Found. Comput. Math. 59–85 (2005)

DOI: 10.1007/s10208-004-0134-1

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

Model Selection for Regularized Least-Squares Algorithm in Learning Theory

E. De Vito,[1] A. Caponnetto,[2] and L. Rosasco[2]

[1]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

[2]D.I.S.I.
Università di Genova
Via Dodecaneso 35
16146 Genova, Italy
and
I.N.F.M.
Sezione di Genova
Via Dodecaneso 33
16146 Genova, Italy
{caponnetto,rosasco}@disi.unige.it

Abstract. We investigate the problem of model selection for learning algorithms depending on a continuous parameter. We propose a model selection procedure based on a worst-case analysis and on a data-independent choice of the parameter. For the regularized least-squares algorithm we bound the generalization error of the solution

Date received: March 22, 2004. Final version received: July 14, 2004. Date accepted: July 20, 2004. Communicated by Steve Smale. Online publication: December 3, 2004.
AMS classification: Primary 68T05, 68P30.
Key words and phrases: Model selection, Optimal choice of parameters, Regularized least-squares algorithm.

by a quantity depending on a few known constants and we show that the corresponding model selection procedure reduces to solving a bias-variance problem. Under suitable smoothness conditions on the regression function, we estimate the optimal parameter as a function of the number of data and we prove that this choice ensures consistency of the algorithm.

1. Introduction

One of the main goals of Learning Theory is the definition of an algorithm that, given a set of examples $(x_i, y_i)_{i=1}^{\ell}$, returns a function f such that $f(x)$ is an effective estimate of the output y when a new input x is given. The map f is chosen from a suitable space of functions, called *hypothesis space*, encoding some a priori knowledge on the relation between x and y.

A learning algorithm is an *inference* process from a finite set of data based on a model represented by the hypothesis space. If the inference process is correct and the model realistic, as the number of available data increases, we expect the solution to approximate the best possible solution. This property is usually called *consistency* [**7**], [**8**], [**10**], [**12**], [**21**].

A central problem of Learning Theory is a quantitative assessment of the inference property of a learning algorithm. A number of seminal works, see, for instance, [**1**], [**7**], [**8**], and [**21**], show that the essential feature of an algorithm should be the capability of controlling the *complexity* of the solution. Roughly speaking, if the model is too complex the algorithm solution overfits the data. In order to overcome overfitting, different *complexity measures* are introduced, such as VC-dimension [**21**], V_γ-dimension [**1**], and covering numbers [**7**], [**24**]. Interestingly, the good behavior of a large class of algorithms has also been recently explained in terms of *stability* with respect to variations of the given training set [**4**], [**17**].

For both approaches it is natural to introduce a parametric family of learning algorithms in which the parameters control the generalization properties. Typical instances of such algorithms are regularized (*à la* Tikhonov) algorithms, see, for instance, [**6**], [**10**], [**11**], [**16**], and [**20**]. In this context a central problem is the *optimal* choice of the parameter as a function of the number of examples.

In this paper we address this issue for the learning algorithms arising in the minimization of the *regularized empirical error*,

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

where the minimization takes place on a Hilbert space of continuous functions. The corresponding algorithm is usually called the *regularized least-squares* algorithm or regularization networks [**6**], [**10**], [**11**], [**16**].

In the above functional the first term measures the error of f on the given examples, while the second term is a penalty term that controls the "smoothness"

of f in order to avoid overfitting. The parameter λ controls the trade-off between these two terms, that is, the balance between the fitting capability of the solution and its complexity.

Our results are in the spirit of [**6**] and [**15**]. In particular, our aim is to provide a selection rule for the parameter λ which is optimal for any number ℓ of examples and provides the desired asymptotic behavior when ℓ goes to infinity. As usual, see [**7**], [**10**], and [**21**], we describe the relation between x and y by means of an unknown probability distribution $\rho(x, y)$. Given a solution f, the expected risk

$$\int (y - f(x))^2 \, d\rho(x, y)$$

measures how well the probabilistic relation between x and y is described by the deterministic rule $y = f(x)$. Following [**6**], the optimal parameter is the one that provides the solution with minimal expectation risk. Since the expected risk is unknown, we need a probabilistic bound on it to have a feasible selection rule. The bound we propose relies on the stability properties of the regularized least-squares algorithm [**4**], and does not depend on any complexity measure on $\mathcal{H}$, which is a central tool in [**6**] and [**15**]. A different point of view is given in [**20**].

The paper is organized as follows. In Section 2 we recall some basic concepts of Learning Theory and we discuss the problem of model selection for algorithms depending on a parameter. In Section 3, we specialize this problem to the regularized least-squares algorithm and find a probabilistic bound for the expected risk of the solution, which is the main result of the paper. In Section 4 we estimate the optimal parameter and prove the consistency of the regularized least-squares algorithm.

2. Learning Theory and Optimal Choice

This section is devoted to the following issues. First, we briefly recall the basic concepts of Learning Theory. Second, we discuss the problem of the choice of the parameter for families of learning algorithms labeled by one real-valued parameter. In particular, we give some insights into the relation between the Bayesian approach (average case) and the learning approach (worst case), and we propose a general framework for parameter selection in a *worst-case scenario* approach. Finally, we discuss the problem of a *data-dependent* choice of the parameter. In the following we assume the reader to have a basic knowledge of Learning Theory (for reviews, see [**5**], [**7**], [**10**], [**12**], and [**21**]).

2.1. *Learning Theory*

In Learning Theory, examples are drawn from two sets: the input space X and the output space Y. The relation between the variable $x \in X$ and the variable $y \in Y$ is

not deterministic and is described by a probability distribution ρ, which is known only by means of ℓ examples $D = ((x_1, y_1), \ldots, (x_\ell, y_\ell))$, drawn identically and independently from $X \times Y$ according to ρ. The set of examples D is called a *training set*. For regression problems, which we deal with in this paper, the output space Y is a subset of real numbers.

The aim of Learning Theory is to *learn* from a training set a function $f : X \to Y$, called an estimator, such that $f(x)$ is an effective estimate of y when x is given. The inference property of f is measured by its *expected risk*,

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

Since we are dealing with regression problems, the choice of the quadratic loss function is natural. However, the discussion of this section holds for a wide class of loss functions (for a discussion of the properties of arbitrary loss functions, see [**10**], [**18**], and [**20**]).

In Learning Theory one approximates the probabilistic relation between x and y by means of functions $y = f(x)$ defined on the input space X and belonging to some a priori given set of model functions, called a *hypothesis space*. For regression problems the model functions f are real-valued.

A *learning algorithm* is a map that, given a training set D, outputs an estimator f_D chosen in the hypothesis space. A good algorithm is such that the expected risk $I[f_D]$ is as small as possible, at least for *generic* training sets.

A well-known example of a learning algorithm is the *empirical risk minimization* algorithm, see, for instance, [**7**], [**10**], and [**21**]. For a training set D, the estimator f_D is defined as the one that minimizes the empirical risk

$$I^D_{\text{emp}}[f_D] = \frac{1}{\ell} \sum_{i=1}^{\ell} (f_D(x_i) - y_i)^2$$

over the hypothesis space. Different choices of the hypothesis space give rise to different algorithms, so one usually introduces a sequence of nested hypothesis spaces,

$$\mathcal{H}_{\lambda_1} \subset \mathcal{H}_{\lambda_n} \subset \cdots \subset \mathcal{H},$$

where $\lambda_1 > \lambda_2 > \cdots > \lambda_n$ and $\mathcal{H}_{\lambda_k}$ is the subset of functions in the model space $\mathcal{H}$ that have *complexity* less than $1/\lambda_k$, according to some suitable measure on complexity (e.g., the inverse of the norm of f [**7**], or its VC-dimension [**10**], [**21**]). The regularized least-squares algorithm discussed in the Introduction is another example of a learning algorithm: for a training set D, the estimator f_D is defined as the minimizer of

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

where the minimum is on a Hilbert space $\mathcal{H}$ of functions defined on X (other loss functions and penalty terms can be considered instead of $\|f\|^2$, see [**10**], [**20**], and [**21**] for a discussion).

In both examples there is a family of algorithms labeled by a parameter controlling the complexity of the estimator. The problem of model selection corresponds to characterizing a rule for the choice of the parameter in such a way that some criterion of optimality with respect to the inference and consistency properties is satisfied. The following section discusses this question.

2.2. *Optimal Choice and Model Selection Rule*

In the following we consider a family of learning algorithms that are labeled by a positive parameter λ; we assume that the complexity of the solution decreases with λ. Given a parameter λ and a training set D, the algorithm provides an estimator $f_D^\lambda \in \mathcal{H}$, where $\mathcal{H}$ is some given (vector) space of functions. In this paper the space $\mathcal{H}$ is given, however, one can easily extend the present discussion to the case of $\mathcal{H}$ being labeled by some parameters, as kernel parameters when $\mathcal{H}$ is a reproducing kernel Hilbert space [**2**].

In this framework the problem of model selection is the problem of the choice of the *optimal parameter* λ. In applications the parameter λ is usually chosen through an a posteriori procedure like cross-validation or using a validation set; see, for instance, [**22**].

To give an a priori definition of the optimal regularization parameter we recall that a good estimator f_D^λ should have a small expected risk $I[f_D^\lambda]$. So a natural definition for the *optimal* parameter λ_{opt}, is the value of λ minimizing $I[f_D^\lambda]$. However, this statement needs some careful examination since $I[f_D^\lambda]$ is a random variable on the sample space of all training sets. This observation introduces some degree of freedom in the definition of λ_{opt}. Following a Bayesian approach, a possible definition is the following:

$$\lambda_{\text{opt}} := \operatorname*{argmin}_{\lambda>0} E_D(I[f_D^\lambda]), \tag{1}$$

where E_D denotes the expectation with respect to the training sets [**23**]. The above definition can be refined by also considering the variance σ^2 of the random variable, for example, considering

$$\lambda_{\text{opt}} := \operatorname*{argmin}_{\lambda>0}\{E_D(I[f_D^\lambda]) + \sigma^2(I[f_D^\lambda])\}. \tag{2}$$

A third possibility is the *worst-case* analysis. Given a confidence level $\eta \in (0, 1)$, we define the quantity

$$E_{\text{opt}}(\lambda, \eta) := \inf_{t\in[0;+\infty)}\{t \mid \text{Prob}\left\{D \in Z^\ell | I[f_D^\lambda] > t\right\} \leq \eta\},$$

and we let the optimal parameter be

$$\lambda_{\text{opt}}(\eta) := \operatorname*{argmin}_{\lambda>0} E_{\text{opt}}(\lambda, \eta). \tag{3}$$

We notice that the first two definitions require the knowledge of the first- (and second-)order momentum of the random variable $I[f_D^\lambda]$. This is a satisfactory characterization if we assume we are dealing with normal distributions. On the other hand, it is easy to see that the third definition amounts to a complete knowledge of the random variable $I[f_D^\lambda]$. Indeed, given λ, $E_{\text{opt}}(\lambda, \eta)$, viewed as function of $1 - \eta$, is the inverse of the distribution function of $I[f_D^\lambda]$.

However, in Learning Theory, the above definitions have only a theoretical meaning since the distribution ρ and, hence, the random variable $I[f_D^\lambda]$, are unknown. To overcome this problem, one studies the random variable $I[f_D^\lambda]$ through a known probabilistic bound $E(\lambda, \ell, \eta)$ of the form

$$\text{Prob}\{D \in Z^\ell \mid |I[f_D^\lambda] > E(\lambda, \ell, \eta)\} \leq \eta. \tag{4}$$

For the worst-case setting the above expression leads to the following model selection rule:

$$\lambda_0(\ell, \eta) := \operatorname*{argmin}_{\lambda > 0} E(\lambda, \ell, \eta). \tag{5}$$

In order to make the above definition rigorous we assume that E extends to a continuous function of λ on $[0, +\infty]$ into itself, and replace (5) by

$$\lambda_0(\ell, \eta) = \max \operatorname*{argmin}_{\lambda \in [0, +\infty]} E(\lambda, \ell, \eta). \tag{6}$$

The continuity of E ensures that the definition is well stated, even if λ_0 could occasionally be zero or infinite. We select the maximum among the minimizers of E to enforce the uniqueness of λ_0: this choice appears quite natural since it corresponds to the most regular estimator fitting the constraint of minimizing the bound.

Some remarks are in order. First of all, different bounds give rise to different selection criteria. Moreover, to have a meaningful selection rule the bound E has to be a function only of known quantities. In this paper we exhibit a bound that gives rise to an optimal parameter defined through a simple algebraic equation. Second, the random variable $I[f_D^\lambda]$ depends on the number ℓ of examples in the training set and, as a consequence, the optimal parameter λ_0 is a function of ℓ. So it is natural to study the asymptotic properties of our selection rule when ℓ goes to infinity. In particular, a basic requirement is *consistency*, that is, the fact that $I[f_D^{\lambda_0(\ell)}]$ approaches the smallest attainable expected risk as the number of data goes to infinity. The concept of (weak universal) consistency is formally expressed by the following definition [**8**].

Definition 2.1. *The one-parameter family of estimators f_D^λ provided with a model selection rule $\lambda_0(\ell)$ is said to be consistent if, for every $\varepsilon > 0$, it holds that*

$$\lim_{\ell \to \infty} \sup_{\rho} \text{Prob}\left\{ D \in Z^\ell \middle| I[f_D^{\lambda_0(\ell)}] > \inf_{f \in \mathcal{H}} I[f] + \varepsilon \right\} = 0,$$

where the sup *is over the set of all probability measures on $X \times Y$.*

In the above definition, the number $\inf_{f\in\mathcal{H}} I[f]$ represents a sort of bias error [**12**], associated with the choice of $\mathcal{H}$ and, hence, it cannot be controlled by the parameter λ. In particular, if there exists $f_{\mathcal{H}} \in \mathcal{H}$ such that $I[f_{\mathcal{H}}] = \inf_{f\in\mathcal{H}} I[f]$, the estimator $f_{\mathcal{H}}$ is the *best* possible deterministic description we can give of the relation between x and y, for a given $\mathcal{H}$. For the sake of clarity, we notice that, for the empirical risk minimization algorithm, the bias error is usually called the *approximation error* and it is controlled by the choice of the hypothesis space [**7**], [**15**].

2.3. *Data Dependency*

The choices of the parameter λ discussed in the above section are a priori since they do not depend on the training set D. One could also consider a posteriori choices where λ_0 depends on the training set D. In a worst-case analysis, this corresponds to considering a bound depending explicitly on D, that is,

$$\text{Prob}\left\{D \in Z^{\ell} \middle| I[f_D^{\lambda}] > E(\lambda, \ell, \eta, D)\right\} \leq \eta. \tag{7}$$

A well-known example of the above bound is the principle of structural risk minimization [**10**], [**21**]. For the Empirical Risk Minimization algorithm in nested hypothesis spaces

$$\mathcal{H}_{\lambda_1} \subset \mathcal{H}_{\lambda_n} \subset \cdots \subset \mathcal{H},$$

the parameter λ is chosen in order to minimize the bound

$$E(\lambda, \ell, \eta, D) = I_{\text{emp}}^{D}[f_D^{\lambda}] + \mathcal{G}(\lambda, \ell, \eta, D), \tag{8}$$

where $\mathcal{G}(\lambda, \ell, \eta, D)$ is a term that controls the complexity of the solution. Usually, $\mathcal{G}$ does not depend on the training set since the measure of complexity is uniform on the hypothesis space $\mathcal{H}_{\lambda}$. However, we get a dependence of the bound on D because of the presence of the empirical term $I_{\text{emp}}^{D}[f_D^{\lambda}]$ in (8).

Now, mimicking the idea of the previous discussion, we could define the optimal parameter as

$$\lambda_0(\ell, \eta, D) := \operatorname*{argmin}_{\lambda>0} E(\lambda, \ell, \eta, D).$$

Clearly, we get a dependence of the optimal parameter on the training set D. This dependence can be, in principle, problematic, due to the probabilistic nature of (4). Indeed, for every λ, we can define the collection of *good* training sets for which the bound is tight, i.e.,

$$\mathcal{A}(\lambda) = \{D \in Z^{\ell} \mid I[f_D^{\lambda}] \leq E(\lambda, \ell, \eta, D)\}.$$

By definition of the bound E, the probability of drawing a good training set $D \in \mathcal{A}(\lambda)$ is greater than $1-\eta$. However, the previous confidence level cannot be applied

to $I[f_D^{\lambda_0(D)}]$. Indeed, it can happen that the probability of drawing a training set D in the set

$$\{D \in Z^\ell \mid I[f_D^{\lambda_0(D)}] \leq E(\lambda_0(D), \ell, \eta, D)\} = \{D \in Z^\ell \mid D \in \mathcal{A}(\lambda_0(D))\}$$

could be much smaller than $1 - \eta$, depending on the structure of the sets $\mathcal{A}(\lambda)$ in the sample space Z^ℓ. Simple toy examples of this pathology can be built.

A possible solution to this kind of problem requires further analyses, see, for instance, [**3**], [**8**], and [**21**]. In this paper we avoid the problem by considering data-independent bounds and hence a priori model selection rules.

3. A Probabilistic Bound for the Regularized Least-Squares Algorithm

We consider throughout the problem of model selection for the regularized least-squares algorithm in the regression setting.

In the present section we first show that the expected risk of the estimator f_D^λ concentrates around the expected risk of f^λ, where f^λ is the minimizer of the regularized expected risk $I[f] + \lambda \|f\|_{\mathcal{H}}^2$.

Moreover, we give a probabilistic bound of the difference between $I[f_D^\lambda]$ and $I[f^\lambda]$ in terms of a function $S(\lambda, \ell, \eta)$ depending on the parameter λ, the number of examples ℓ, and the confidence level $1 - \eta$. Our results are based on the stability properties of the regularized least-squares algorithm [**4**], and the McDiarmid concentration inequality [**13**]. In particular, we do not make use of any complexity measure on the hypothesis space, like the VC-dimension [**21**], or the covering number [**7**], [**15**]. We stress that the bound S depends on $\mathcal{H}$ only through two simple constants related to the topological properties of X and Y.

Compared to previous results (see, for instance, [**3**], [**4**], [**7**], [**14**], [**15**], [**21**]) we are not interested in the deviation of the empirical risk from the expected risk and we bound directly the expected risk of the estimator f_D^λ. Moreover, our result concentrates $I[f_D^\lambda]$ around $I[f^\lambda]$ both from above and below, so that $I[f^\lambda]$ will play the role of approximation error and $S(\lambda, \ell, \eta)$ the role of sample error (our terminology is close to the definition of [**6**], which is somewhat different from the notation of [**7**] and [**15**]).

Finally, in order to obtain algorithmically computable results, we make some smoothness assumption on the probability distribution ρ. By means of standard results in approximation theory [**19**], we find a bound, depending only on known quantities.

Before stating the main theorem of this section we set the notations.

3.1. *Notations*

We assume that the input space X is a compact subset of $\mathbb{R}^d$ and that the output space Y is a compact subset of $\mathbb{R}$. The assumption of compactness is for technical reasons and simplifies the proofs.

We let ρ be the unknown probability measure on $Z = X \times Y$ describing the relation between $x \in X$ and $y \in Y$, and ν the marginal distribution of ρ on X. Moreover, for ν-almost all $x \in X$, let ρ_x be the conditional distribution of y with respect to x.

If ξ is a random variable on Z, we denote its mean value by $E_Z(\xi)$,

$$E_Z(\xi) = \int_{X\times Y} \xi(x, y)\, d\rho(x, y).$$

As usual, $L^2(X, \nu)$ is the Hilbert space of square-integrable functions on X and $\|\cdot\|_\nu$, $\langle\cdot,\cdot\rangle_\nu$ are the corresponding norm and scalar product.

We let $\mathcal{C}(X)$ be the space of (real) continuous functions on X equipped with the uniform norm, $\|f\|_\infty = \sup_{x\in X} |f(x)|$.

We denote by f_0 and σ_0 the *regression* and *noise* functions defined, respectively, as

$$f_0(x) = \int_Y y\, d\rho_x(y), \tag{9}$$

$$\sigma_0^2(x) = \left(\int_Y y^2\, d\rho_x(y)\right) - (f_0(x))^2, \tag{10}$$

that belong to $L^2(X, \nu)$ due to the compactness of X and Y.

Given $\ell \in \mathbb{N}$, let Z^ℓ be the set of all training sets with ℓ examples. We regard Z^ℓ as a probability space with respect to the product measure $\rho^\ell = \rho \otimes \cdots \otimes \rho$. If ξ is a random variable on Z^ℓ we denote its mean value by $E_D(\xi)$,

$$E_D(\xi) = \int_{Z^\ell} \xi(D)\, d\rho^\ell(D).$$

Given $D \in Z^\ell$, let ρ_D be the empirical measure on $X \times Y$ defined by D, that is,

$$\rho_D = \frac{1}{\ell}\sum_{i=1}^{\ell} \delta_{x_i}\delta_{y_i},$$

where δ_x and δ_y are the Dirac measures at x and y, respectively.

We assume the hypothesis space $\mathcal{H}$ to be a reproducing kernel Hilbert space with a continuous kernel $K\colon X \times X \to \mathbb{R}$. The assumption on the kernel ensures $\mathcal{H}$ being a Hilbert space of continuous functions. We let $\|\cdot\|_\mathcal{H}$ and $\langle\cdot,\cdot\rangle_\mathcal{H}$ be the corresponding norm and scalar product. We define

$$K_x(s) = K(s, x), \qquad x \in X,$$

$$\kappa = \sup\{\sqrt{K(x, x)} \mid x \in X\}, \tag{11}$$

$$\delta = \sup\{|y| \mid y \in Y\}. \tag{12}$$

It is well known [**2**], [**7**] that

$$\begin{aligned} \mathcal{H} &\subset \mathcal{C}(X) \subset L^2(X, \nu), \\ f(x) &= \langle f, K_x \rangle_{\mathcal{H}}, \qquad x \in X, \\ \|\cdot\|_\nu &\leq \|\cdot\|_\infty \leq \kappa \|\cdot\|_{\mathcal{H}}. \end{aligned} \tag{13}$$

We denote by L_ν the integral operator on $L^2(X, \nu)$ with kernel K, that is,

$$(L_\nu f)(s) = \int_X K(s, x) f(x)\, d\nu(x), \qquad s \in X, \tag{14}$$

and by P the projection onto the closure of the range of L_ν. In particular, one has that the closure of $\mathcal{H}$ with respect to the norm of $L^2(X, \nu)$ is $PL^2(X, \nu)$.

We recall that, given $f \in L^2(X, \nu)$, the *expected risk* of f is

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

We let $I_{\mathcal{H}}$ be the *bias error*,

$$I_{\mathcal{H}} = \inf_{f \in \mathcal{H}} I[f].$$

For any $\lambda > 0$ we denote by f^λ the solution of

$$\min_{f \in \mathcal{H}} \{I[f] + \lambda \|f\|_{\mathcal{H}}^2\}, \tag{15}$$

which exists and is unique, see, for instance, [**6**].

Finally, given $D \in Z^\ell$, the *empirical risk* of $f \in \mathcal{C}(X)$ is given by

$$I_{\text{emp}}^D[f] = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - f(x_i))^2.$$

For all $\lambda > 0$, the estimator f_D^λ is defined as the unique solution of

$$\min_{f \in \mathcal{H}} \{I_{\text{emp}}^D[f] + \lambda \|f\|_{\mathcal{H}}^2\}, \tag{16}$$

which exists and is unique [**6**].

In the following we will always consider the square root of the unbiased risk that we indicate with $\mathcal{R}[f]$ to simplify the notation, that is,

$$\mathcal{R}[f] = \sqrt{I[f] - I_{\mathcal{H}}}.$$

Indeed, (21) below will show that this quantity can conveniently be interpreted as a distance in $L^2(X, \nu)$.

3.2. *Main Results*

The following theorem shows that, given the parameter λ, the expected risk of the estimator f_D^λ provided by the regularized least-squares algorithm concentrates around the value $I[f^\lambda]$. Moreover, the deviation can be bounded by a simple function S depending only on the confidence level, the number of examples, and two constants, κ and δ, encoding some topological properties of X, Y, and the kernel (see (11) and (12)).

Theorem 3.1. *Given* $0 < \eta < 1$, $\ell \in \mathbb{N}$, *and* $\lambda > 0$, *with probability at least* $1 - \eta$,

$$|\mathcal{R}[f_D^\lambda] - \mathcal{R}[f^\lambda]| \leq S(\lambda, \ell, \eta),$$

where

$$S(\lambda, \ell, \eta) = \frac{\delta \kappa^2}{\lambda \sqrt{\ell}} \left(1 + \frac{\kappa}{\sqrt{\lambda}}\right) \left(1 + \sqrt{2 \log \frac{2}{\eta}}\right). \tag{17}$$

The proof of Theorem 3.1 is postponed to Section 3.3 after a brief discussion and some remarks on the result.

Let us interpret the quantities occurring in our inequality. The data-independent term $\mathcal{R}[f^\lambda]$ can be interpreted as the price paid by replacing the regression function f_0 with the regularized solution f^λ, in short, as the *approximation error*, compare with [**6**] and [**15**].

The term $S(\lambda, \ell, \eta)$ is a bound on $\left|\mathcal{R}[f_D^\lambda] - \mathcal{R}[f^\lambda]\right|$, that is, on the *sample error* made by approximating f^λ through a finite training set D, compare with [**6**], [**15**] and [**21**].

Since Theorem 3.1 bounds $\mathcal{R}[f_D^\lambda]$ both from above and below and $S(\lambda, \ell, \eta)$ goes to zero for ℓ going to $+\infty$, the expected risk of f_D^λ concentrates around the expected risk of f^λ. Then the splitting of $\mathcal{R}[f_D^\lambda]$ into the approximation error and the sample error appears quite natural and intrinsic to the problem.

3.3. *Proofs*

In the following, before dealing with the main result, we briefly sketch the scheme of the proof and we show some useful lemmas.

The proof of Theorem 3.1 is essentially based on two steps:

- We show that the regularized least-squares algorithm satisfies a kind of stability property with respect to variation of the training set, compare with [**4**].
- We give an estimate of the mean value of $\mathcal{R}[f_D^\lambda]$ (and hence of the mean value of the expected risk).

More precisely, given λ, we regard $\mathcal{R}[f_D^\lambda]$ as a real random variable on Z^ℓ and we

apply the McDiarmid inequality [**13**]. This inequality tells us that

$$\text{Prob}\{D \in Z^\ell \mid |\mathcal{R}[f_D^\lambda] - E_D\big(\mathcal{R}[f_D^\lambda]\big)| \geq \varepsilon\} \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^\ell c_i^2} \tag{18}$$

provided that

$$\sup_{D\in Z^\ell} \sup_{(x',y')\in Z} |\mathcal{R}[f_D^\lambda] - \mathcal{R}[f_{D^i}^\lambda]| \leq c_i, \tag{19}$$

where D^i is the training set with the ith example being replaced by (x', y').

To work out the proof, we recall some preliminary facts. Since we are considering the quadratic loss the expected risk of $f \in \mathcal{H}$ can be written in the following way:

$$\begin{aligned} I[f] &= \|f - f_0\|_\nu^2 + \|\sigma_0\|_\nu^2 \\ &= \|P(f - f_0)\|_\nu^2 + \|(I - P)(f - f_0)\|_\nu^2 + \|\sigma_0\|_\nu^2 \\ &= \|f - Pf_0\|_\nu^2 + \|(I - P)(f - f_0)\|_\nu^2 + \|\sigma_0\|_\nu^2 \\ &= \|f - Pf_0\|_\nu^2 + \|(I - P)f_0\|_\nu^2 + \|\sigma_0\|_\nu^2, \end{aligned} \tag{20}$$

where f_0 and σ_0 are given by (9) and (10), and $(I - P)f = 0$ since $f \in \mathcal{H} \subset PL^2(X, \nu)$.

It follows that $I_\mathcal{H} = \inf_{f\in\mathcal{H}} I[f] = \|(I - P)f_0\|_\nu^2 + \|\sigma_0\|_\nu^2$ and

$$\mathcal{R}[f] = \sqrt{I[f] - I_\mathcal{H}} = \|f - Pf_0\|_\nu. \tag{21}$$

We now recall the explicit form of the minimizers of (15) and (16). One has that

$$f^\lambda = (T + \lambda)^{-1} g_\rho, \tag{22}$$

$$f_D^\lambda = (T_\mathbf{x} + \lambda)^{-1} g_D, \tag{23}$$

where T and $T_\mathbf{x}$ are positive operators on $\mathcal{H}$ given by

$$T = \int_X \langle \cdot, K_x \rangle_\mathcal{H} K_x \, d\nu(x), \tag{24}$$

$$T_\mathbf{x} = \frac{1}{\ell} \sum_{i=1}^\ell \langle \cdot, K_{x_i} \rangle_\mathcal{H} K_{x_i}, \tag{25}$$

and g_ρ and g_D are functions in $\mathcal{H}$ defined by

$$g_\rho = \int_X K_x f_0(x) \, d\nu(x), \tag{26}$$

$$g_D = \frac{1}{\ell} \sum_{i=1}^\ell y_i K_{x_i}, \tag{27}$$

(in (24) the integral is taken with respect to the trace operator norm, while in (26) the integral is with respect to the norm of $\mathcal{H}$. Both integrals are finite since the integrands are continuous and X is compact).

In order to present the proof of Theorem 3.1 we need some preliminary lemmas. The first one provides an upper bound on $\|f_D^\lambda\|_\mathcal{H}$. The proof is standard and we report it for completeness; see, for instance, [**4**].

Lemma 3.1. *For all* $\lambda > 0$,

$$\|f_D^\lambda\|_\mathcal{H} \leq \frac{\delta}{\sqrt{\lambda}}.$$

Proof. Since by definition, see (16),

$$f_D^\lambda = \operatorname*{argmin}_{f\in\mathcal{H}} \{I_{\text{emp}}^D[f] + \lambda\|f\|_\mathcal{H}^2\},$$

with the choice $f = 0$, it follows that

$$I_{\text{emp}}^D[f_D^\lambda] + \lambda\|f_D^\lambda\|_\mathcal{H}^2 \leq I_{\text{emp}}^D[0] + \lambda\|0\|_\mathcal{H}^2 = \frac{1}{\ell}\sum_{i=1}^{\ell} y_i^2 \leq \delta^2,$$

where in the last inequality we recalled the definition of δ, see (12). The thesis follows from the fact that $I_{\text{emp}}^D[f_D^\lambda] \geq 0$. □

The next step is the estimate of the expectation value of $\|T_\mathbf{x} - T\|$ (here $\|\cdot\|$ denotes the operator norm in $\mathcal{H}$) and that of $\|g_D - g_\rho\|_\mathcal{H}$. To this end, we recall the following property regarding vector-valued random variables on Z. Let $\mathcal{K}$ be a Hilbert space and ξ a $\mathcal{K}$-valued random variable on Z, then

$$E_D\left(\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi(x_i, y_i) - E_Z(\xi)\right\|_\mathcal{K}\right)^2 \leq E_D\left(\left\|\frac{1}{\ell}\sum_{i=1}^{\ell}\xi(x_i, y_i) - E_Z(\xi)\right\|_\mathcal{K}^2\right)$$
$$= \frac{1}{\ell}\big(E_Z(\|\xi\|_\mathcal{K}^2) - \|E_Z(\xi)\|_\mathcal{K}^2\big). \tag{28}$$

The first inequality is a consequence of the Schwarz inequality and the equality is a straightforward extension of the well-known property of real random variables [**9**].

Lemma 3.2. *Let* $a_1 = \int_{X\times X}(K(x,x)^2 - K(x,x')^2)\,d\nu(x)\,d\nu(x')$, *then*

$$E_D(\|T_\mathbf{x} - T\|) \leq \sqrt{\frac{a_1}{\ell}} \leq \frac{\kappa^2}{\sqrt{\ell}}.$$

Proof. Let $B_2(\mathcal{H})$ be the Hilbert space of self-adjoint Hilbert–Schmidt operators on $\mathcal{H}$ with scalar product $\langle A, B\rangle_{B_2(\mathcal{H})} = \mathrm{Tr}(AB)$. Notice that, for all $x \in X$, the rank-one operator $\langle \cdot, K_x\rangle_{\mathcal{H}} K_x$ is in $B_2(\mathcal{H})$, so we can define the $B_2(\mathcal{H})$-valued random variable ξ as

$$\xi(x) = \langle \cdot, K_x\rangle_{\mathcal{H}} K_x.$$

Since $\mathrm{Tr}(\langle \cdot, K_x\rangle_{\mathcal{H}} K_{x'}) = \langle K_{x'}, K_x\rangle_{\mathcal{H}}$, then

$$\begin{aligned} \|\xi(x) - \xi(x')\|^2_{B_2(\mathcal{H})} &= \mathrm{Tr}((\xi(x) - \xi(x'))^2) \\ &= K^2(x,x) + K^2(x',x') - 2K^2(x',x), \end{aligned}$$

which implies the continuity of ξ. Its mean value is given by

$$\begin{aligned} E_Z(\xi) &= \int_X \langle \cdot, K_x\rangle_{\mathcal{H}} K_x \, d\nu(x), \\ &= T. \end{aligned}$$

Now, observe that

$$\begin{aligned} E_D(\|\xi\|^2_{B_2(\mathcal{H})}) &= E_D(\mathrm{Tr}\,(\langle \cdot, K_x\rangle_{\mathcal{H}} K_x K(x,x))) \\ &= \int_X K(x,x)^2 \, d\nu(x). \end{aligned}$$

Finally,

$$\begin{aligned} \|E_Z(\xi)\|^2_{B_2(\mathcal{H})} &= \mathrm{Tr}(T^2) \\ &= \int_{X\times X} \mathrm{Tr}(\langle \cdot, K_x\rangle_{\mathcal{H}} K_{x'} K(x,x')) \, d\nu(x) \, d\nu(x') \\ &= \int_{X\times X} K(x,x')^2 \, d\nu(x) \, d\nu(x'). \end{aligned}$$

Applying (28) and the definition of a_1, one has that

$$E_D(\|T_{\mathbf{x}} - T\|_{B_2(\mathcal{H})}) \le \sqrt{\frac{a_1}{\ell}}.$$

The thesis follows observing that

$$\|T_{\mathbf{x}} - T\| \le \|T_{\mathbf{x}} - T\|_{B_2(\mathcal{H})}$$

and

$$a_1 \le \int_X K(x,x)^2 \, d\nu(x) \le \kappa^4. \qquad \square$$

Lemma 3.3. *Let* $a_2 = \int_Z y^2 K(x,x)\, d\rho(x,y) - \|g_\rho\|_{\mathcal{H}}^2$, *then*

$$E_D(\|g_D - g_\rho\|_{\mathcal{H}}) \;\leq\; \sqrt{\frac{a_2}{\ell}} \leq \frac{\delta\,\kappa}{\sqrt{\ell}}.$$

Proof. We follow the scheme of the previous proof. Let ξ be the $\mathcal{H}$-valued random variable

$$\xi(x,y) = yK_x.$$

Since

$$\|\xi(x,y) - \xi(x',y')\|_{\mathcal{H}}^2 = y^2K(x,x) + y'^2K(x',x') - 2yy'K(x,x'),$$

ξ is continuous and

$$E_Z(\xi) \;=\; \int_{X\times Y} yK_x\, d\rho(x,y) = g_\rho,$$

by definition of g_ρ.

Moreover,

$$E_Z(\|\xi\|_{\mathcal{H}}^2) \;=\; \int_{X\times Y} y^2K(x,x)\, d\rho(x,y).$$

Applying (28) and the definition of a_2, one has that

$$E_D(\|g_D - g_\rho\|_{\mathcal{H}}) \leq \frac{a_2}{\ell}.$$

The thesis follows observing that

$$0 \leq a_2 \leq \int_{X\times Y} y^2K(x,x)\, d\rho(x,y) \leq \delta^2\kappa^2. \qquad \square$$

The next lemma estimates the expectation value of $\mathcal{R}[f_D^\lambda]$.

Lemma 3.4. *Following the above notations*

$$|E_D(\mathcal{R}[f_D^\lambda]) - \mathcal{R}[f^\lambda]| \;\leq\; \frac{\kappa^2\delta}{\lambda\sqrt{\ell}}\left(1 + \frac{\kappa}{\sqrt{\lambda}}\right).$$

Proof. By means of (23),

$$f_D^\lambda = (T_{\mathbf{x}} + \lambda)^{-1} g_D = (T+\lambda)^{-1}\left(g_D - g_\rho\right) + (T+\lambda)^{-1}(T - T_{\mathbf{x}})(T_{\mathbf{x}} + \lambda)^{-1} g_D + (T+\lambda)^{-1} g_\rho,$$

that is, using (22),

$$f_D^\lambda - f^\lambda = (T+\lambda)^{-1}\left(g_D - g_\rho\right) + (T+\lambda)^{-1}(T - T_{\mathbf{x}}) f_D^\lambda.$$

Using (21) and the triangle inequality we have that

$$\begin{aligned} |\mathcal{R}[f_D^\lambda] - \mathcal{R}[f^\lambda]| &\le \|f_D^\lambda - f^\lambda\|_\nu \\ \text{(Eq.(13))} &\le \kappa \|(T+\lambda)^{-1}(g_D - g_\rho)\|_{\mathcal{H}} + \kappa \|(T+\lambda)^{-1}(T - T_{\mathbf{x}}) f_D^\lambda\|_{\mathcal{H}} \\ &\le \frac{\kappa}{\lambda}\|g_D - g_\rho\|_{\mathcal{H}} + \frac{\kappa\delta}{\lambda^{3/2}}\|T - T_{\mathbf{x}}\|, \end{aligned}$$

where $\|(T+\lambda)^{-1}\| \le 1/\lambda$ and we used Lemma 3.1 to bound $\|f_D^\lambda\|_{\mathcal{H}}$.

Finally we take the mean value on D and use Lemmas 3.2 and 3.3. □

We are now ready to prove the main result of the section.

Proof of Theorem 3.1. The proof uses the McDiarmid inequality (18). Due to (21), the conditions (19) become

$$\sup_{D \in Z^\ell} \sup_{(x', y') \in Z} |\, \|f_D^\lambda - P f_0\|_\nu - \|f_{D^i}^\lambda - P f_0\|_\nu| \le c_i$$

(we recall that D^i is the training set where the ith example is replaced by (x', y')).

In order to compute the constants c_i, we notice that f_D^λ can be decomposed, by means of (23), as

$$f_D^\lambda = (T_{\mathbf{x}} + \lambda)^{-1} g_D = (T_{\mathbf{x}} + \lambda)^{-1}(g_D - g_{D^i}) + (T_{\mathbf{x}} + \lambda)^{-1}(T_{\mathbf{x}^i} - T_{\mathbf{x}})(T_{\mathbf{x}^i} + \lambda)^{-1} g_{D^i} + (T_{\mathbf{x}^i} + \lambda)^{-1} g_{D^i},$$

that is, again using (23) with D^i,

$$f_D^\lambda - f_{D^i}^\lambda = (T_{\mathbf{x}} + \lambda)^{-1}(g_D - g_{D^i}) + (T_{\mathbf{x}} + \lambda)^{-1}(T_{\mathbf{x}^i} - T_{\mathbf{x}}) f_{D^i}^\lambda.$$

By the triangle inequality, we can write

$$\begin{aligned} |\,\|f_D^\lambda - Pf_0\|_\nu - \|f_{D^i}^\lambda - Pf_0\|_\nu| &\le \|f_D^\lambda - f_{D^i}^\lambda\|_\nu \\ \text{(Eq. (13))} &\le \kappa\|(T_{\mathbf{x}}+\lambda)^{-1}(g_D - g_{D^i})\|_{\mathcal{H}} \\ &\quad + \kappa\|(T_{\mathbf{x}}+\lambda)^{-1}(T_{\mathbf{x}} - T_{\mathbf{x}^i})f_{D^i}^\lambda\|_{\mathcal{H}} \\ &\le \frac{2\delta\kappa^2}{\lambda\ell}\left(1+\frac{\kappa}{\sqrt{\lambda}}\right) \\ &:= c_i, \end{aligned}$$

where we used Lemma 3.1 to bound $\|f_{D^i}^\lambda\|_{\mathcal{H}}$ and

$$\begin{aligned} \|(T_{\mathbf{x}}+\lambda)^{-1}\| &\le \frac{1}{\lambda}, \\ \|T_{\mathbf{x}} - T_{\mathbf{x}^i}\| &= \frac{1}{\ell}\|\langle\cdot, K_{x_i}\rangle_{\mathcal{H}}K_{x_i} - \langle\cdot, K_{x_i'}\rangle_{\mathcal{H}}K_{x_i'}\| \le \frac{1}{\ell}2\kappa^2, \\ \|g_D - g_{D^i}\|_{\mathcal{H}} &= \frac{1}{\ell}\|yK_{x_i} - y_i'K_{x_i'}\|_{\mathcal{H}} \le \frac{1}{\ell}2\delta\kappa. \end{aligned}$$

Plugging the constants c_i into the McDiarmid inequality (18), we have that

$$|\mathcal{R}[f_D^\lambda] - E_D(\mathcal{R}[f_D^\lambda])| \le \varepsilon$$

with probability

$$1 - 2e^{-\varepsilon^2/(2(\delta\kappa^2/(\lambda\sqrt{\ell}(1+\kappa/\sqrt{\lambda}))^2} = 1-\eta$$

so, with probability at least $1-\eta$,

$$|\mathcal{R}[f_D^\lambda] - E_D(\mathcal{R}[f_D^\lambda])| \le \frac{\delta\kappa^2}{\lambda\sqrt{\ell}}\left(1+\frac{\kappa}{\sqrt{\lambda}}\right)\sqrt{2\log\frac{2}{\eta}}.$$

The above bound together with Lemma 3.4, once again by the triangle inequality, completes the proof. □

4. Estimate of the Optimal Parameter

We are now in the position to apply the results of the previous section to the actual estimate of the regularization parameter, following the technique presented in Section 2.

From Theorem 3.1 we can easily derive the following bound:

$$\mathcal{R}[f_D^\lambda] \le \mathcal{R}[f^\lambda] + S(\lambda, \ell, \eta), \tag{29}$$

which holds with probability at least $1 - \eta$.

From the explicit form of $S(\lambda, \ell, \eta)$, we have that $S(\lambda, \ell, \eta)$ decreases with λ and goes to $+\infty$ when λ goes to 0. On the other hand, it is easy to check that $\mathcal{R}[f^\lambda]$ is an increasing function of λ and goes to 0 for λ going to 0 [**6**].

The bound (29) is of the form of (4) and can be used in the model selection rule defined by (6). Our definition ensures the existence and uniqueness of the estimate λ_0 of the optimal parameter, however, we still have to prove that λ_0 is finite. We now prove that the bound in (29) provides an estimate λ_0 that is finite for large enough ℓ.

We consider a slightly more general case that will be useful in the following. We let $A(\lambda)$ be an upper bound on the approximation error, that is, an *increasing*, *continuous* function from $[0, +\infty]$ to $[0, +\infty]$ satisfying

$$A(\lambda) \ge \mathcal{R}[f^\lambda],$$

and

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

The following proposition highlights the special role played by the approximation error $\mathcal{R}[f^\lambda]$ with respect to an arbitrary data-independent bound of the form given in (4).

Proposition 4.1. *Let $E(\lambda, \ell, \eta)$ be a bound for $\mathcal{R}[f_D^\lambda]$, that is, with probability at least $1 - \eta$,*

$$\mathcal{R}[f_D^\lambda] \le E(\lambda, \ell, \eta).$$

Assume that $\lim_{\ell \to \infty} E(\lambda, \ell, \eta) = A(\lambda)$ for all η. Then,

$$\mathcal{R}[f^\lambda] \le A(\lambda).$$

Proof. Let $\varepsilon > 0$ and $\eta < \frac{1}{2}$. Since $\lim_{\ell \to \infty} S(\lambda, \ell, \eta) = 0$, there is an ℓ_1 such that

$$S(\lambda, \ell, \eta) \le \varepsilon \quad \text{for all } \ell \ge \ell_1, \tag{30}$$

and, by definition of $A(\lambda)$, there is an ℓ_2 such that

$$|E(\lambda, \ell, \eta) - A(\lambda)| \le \varepsilon \quad \text{for all } \ell \ge \ell_2. \tag{31}$$

Let $\ell_3 = \max\{\ell_1, \ell_2\}$. By Theorem 3.1, with probability at least $1 - \eta$,

$$|\mathcal{R}[f_D^\lambda] - \mathcal{R}[f^\lambda]| \le S(\lambda, \ell_3, \eta) \le \varepsilon, \tag{32}$$

where we used (30). By definition one has that, with probability at least $1-\eta$,

$$\mathcal{R}[f_D^\lambda] \leq E(\lambda, \ell_3, \eta) \leq A(\lambda) + \varepsilon, \tag{33}$$

where we used (31).

It follows from (32) and (33) that, with probability at least $1-2\eta > 0$,

$$A(\lambda) \;\geq\; \mathcal{R}[f_D^\lambda] - \varepsilon \geq \mathcal{R}[f^\lambda] - 2\varepsilon.$$

Since ε is arbitrary, one has the thesis. □

According to the discussion in Section 2, given η and ℓ, we estimate the optimal parameter $\lambda_0(\ell, \eta)$ as the one that minimizes the bound

$$E(\lambda, \ell, \eta) = A(\lambda) + S(\lambda, \ell, \eta).$$

The following proposition shows that λ_0 is finite, at least if ℓ is large enough.

Proposition 4.2. *Assume $Pf_0 \neq 0$ and let $0 < \eta < 1$. There is $\overline{\ell} \in \mathbb{N}$ such that $\lambda_0(\ell, \eta)$ is finite for all $\ell \geq \overline{\ell}$.*

In particular, if $\lim_{\lambda\to\infty} A(\lambda) = +\infty$, then $\lambda_0(\ell, \eta)$ is finite for every ℓ.

Proof. We will prove the finiteness of λ_0 by applying the Weierstrass theorem to the continuous function $E(\lambda, \ell, \eta)$.

Clearly, $\lim_{\lambda\to 0} E(\lambda, \ell, \eta) = +\infty$. Moreover, letting $M = \lim_{\lambda\to\infty} A(\lambda)$, which always exists by the assumed continuity at $\lambda = +\infty$, one has that

$$\lim_{\lambda\to\infty} E(\lambda, \ell, \eta) = M.$$

We now prove that, if ℓ is large enough, there is $\overline{\lambda} > 0$ such that $E(\overline{\lambda}, \ell, \eta) < M$. In fact, since $A(\lambda) \geq \mathcal{R}[f^\lambda]$ and $\lim_{\lambda\to+\infty} \mathcal{R}[f^\lambda] = \|Pf_0\|_\nu > 0$, then $M > 0$. Moreover, since $\lim_{\lambda\to 0} A(\lambda) = 0$, there is $\overline{\lambda} > 0$ such that $A(\overline{\lambda}) < M/2$. Finally, observing that $\lim_{\ell\to\infty} S(\overline{\lambda}, \ell, \eta) = 0$, we conclude that there exists $\overline{\ell} \in \mathbb{N}$ such that

$$S(\overline{\lambda}, \overline{\ell}, \eta) < \frac{M}{2}.$$

It follows that, since S is a decreasing function of ℓ, for all $\ell \geq \overline{\ell}$,

$$E(\overline{\lambda}, \ell, \eta) \leq E(\overline{\lambda}, \overline{\ell}, \eta) < M.$$

Hence, by means of the Weierstrass theorem $E(\lambda, \ell, \eta)$ attains its minimum. Moreover, $\min E \leq E(\overline{\lambda}, \ell, \eta) < M$ so that all the minimizers are finite.

Assume now that $\lim_{\lambda\to\infty} A(\lambda) = +\infty$, then $M = +\infty$ and, clearly, $\min E < +\infty$, so that the minimizers are finite for all ℓ. □

Remark 4.1. The assumption that $Pf_0 \neq 0$ is natural. If $Pf_0 = 0$, the problem of model selection is trivial since $\mathcal{H}$ is too poor to give a reasonable approximation of f_0 still with infinite data.

In order to actually use the estimate λ_0, we have to explicitly compute the minimizer of the bound E. Hence, the function E has to be computable. Though the sample error $S(\lambda, \ell, \eta)$ is a simple function of the parameters λ, ℓ, η, κ, and δ, the *approximation* error $\mathcal{R}[f^\lambda]$ is not directly computable and we need a suitable bound.

We do not discuss the problem of the estimate of the approximation error since there is a large literature on the topic; see [**15**], [**19**] and references therein. We only notice that a common features of these bounds is that one has to make some assumptions on the probability distribution ρ, that is, on the regression function f_0. Clearly, with these hypotheses we lose in generality. However if we want to solve the bias-variance problem in this framework, it seems to be an unavoidable step (compare with [**6**], [**7**], [**15**]).

Using an estimate on the approximation error $A(\lambda)$ given in Theorem 3, Chapter II of [**7**], one easily obtains the following result:[1]

Corollary 4.1. *Let $r \in (0, 1]$ and $C_r > 0$ such that $\|L_\nu^{-r} P f_0\|_\nu \leq C_r$, where f_0 is given by* (9), *L_ν and P by* (14), *then*

$$\mathcal{R}[f_D^\lambda] \leq \lambda^r C_r + S(\lambda, \ell, \eta) =: E_r(\lambda, \ell, \eta),$$

with probability at least $1 - \eta$.

In particular, for all ℓ and η, the bound E_r gives rise to a finite estimate $\lambda_0^r(\ell, \eta)$ of the optimal parameter, which is the unique solution of the following equation:

$$r C_r \lambda^{r+1} = \frac{\delta \kappa^2}{\sqrt{\ell}} \left(1 + \frac{3\kappa}{2\sqrt{\lambda}}\right) \left(1 + \sqrt{2 \log \frac{2}{\eta}}\right).$$

To compare our results with the bounds obtained in the literature we assume that f_0 belongs to the hypothesis space so we can choose $r = \frac{1}{2}$ in the above corollary. Since $I_{\mathcal{H}} = I[f_0]$, $I[f] = \mathcal{R}[f]^2 + I[f_0]$ so that

$$I[f_D^\lambda] = I[f_0] + O(\lambda) + O\left(\frac{1}{\ell \lambda^3}\right)$$

with probability greater than $1 - \eta$. The best rate of convergence is obtained by choosing $\lambda_\ell = 1/\sqrt[4]{\ell}$ so that

$$I[f_D^\lambda] = I[f_0] + O\left(\frac{1}{\sqrt[4]{\ell}}\right).$$

[1] In Appendix A we provide a direct proof of such an estimate.

The rate is comparable with the rate we can deduce from the bounds of [**4**] where, however, the dependence of λ from ℓ is not considered.[2] In [**23**] a bound of the order $O(1/\sqrt{n})$ is obtained using a leave-one-out technique, but with a worse confidence level.[3]

Finally, we notice that our model selection rule, based on an a priori assumption on the target function, is only of theoretical use since the condition that the approximation error is of the order $O(\lambda^r)$ is unstable with respect to the choice of f_0 (if the kernel is infinite dimensional, as the Gaussian kernel) [**19**].

4.1. *Asymptotic Properties and Consistency*

The aim of the present subsection is to state some asymptotic properties, for an increasing number of examples ℓ, of the regularized least-squares algorithm provided with the parameter choice described at the beginning of this section. In particular, we consider properties of the selected parameter $\lambda_0 = \lambda_0(\ell, \eta)$ with respect to the notion of consistency already introduced by Definition 2.1. For clarity we restate here that definition in terms of the modified expected risk $\mathcal{R}[f_D^\lambda]$ defined in Section 3.

Definition 4.1. *The one-parameter family of estimators f_D^λ provided with a model selection rule $\lambda_0(\ell)$ is said to be* (*weakly universally*) *consistent if, for every $\varepsilon > 0$, it holds that*

$$\lim_{\ell\to\infty} \sup_{\rho} \text{Prob}\left\{ D \in Z^\ell \middle| \mathcal{R}[f_D^{\lambda_0(\ell)}] > \varepsilon \right\} = 0,$$

where the sup *is over the set of all probability measures on $X \times Y$.*

From a general point of view consistency can be considered as a property of the algorithm according to which, for a large data amount, the algorithm provides the best possible estimator.

In order to apply this definition to our selection rule we need to specify the explicit dependence of the confidence η on the number of examples ℓ, i.e., to transform the two-parameter family of real positive numbers $\lambda_0(\ell, \eta)$ to the one-parameter family $\lambda_0(\ell) = \lambda_0(\ell, \overline{\eta}(\ell))$ corresponding to a specific choice $\overline{\eta}(\ell)$ of the confidence level. We assume the following power law behavior:

$$\overline{\eta}(\ell) = \ell^{-p} \; p > 0. \tag{34}$$

The main result of this section is contained in the following theorem where we prove that the regularized least-squares algorithm provided by our model selection rule is consistent.

---

[2] See Theorems 12 and 22 of [**4**], in particular, the proof of the latter theorem gives an upper bound of the sample error of the order $O(1/\sqrt{\ell}\lambda^{3/2})$.

[3] See discussions at the end of Sections 4 and 5 of [**23**].

Theorem 4.1. *Given* $\lambda_0(\ell) = \lambda_0(\ell, \overline{\eta}(\ell))$ *where* $\overline{\eta}(\ell)$ *is as in* (34), *then the following three properties hold*:

(1) *if* $\ell' > \ell > 2$, *then* $\lambda_0(\ell') \leq \lambda_0(\ell)$;
(2) $\lim_{\ell \to \infty} \lambda_0(\ell) = 0$;
(3) *the sequence* $(\lambda_0(\ell))_{\ell=1}^{\infty}$ *provides consistency*.

Proof. First of all, let us notice that the dependence of $E(\lambda, \ell, \eta)$ on ℓ and η can be factorized as follows:

$$E(\lambda, \ell, \eta) = A(\lambda) + C(\ell, \eta)s(\lambda), \tag{35}$$

where the sample error term $S(\lambda, \ell, \eta)$ in (17) has been split by means of the functions $C(\ell, \eta)$ and $s(\lambda)$ defined by

$$C(\ell, \eta) = \frac{\delta}{\sqrt{\ell}} \left(1 + \sqrt{2 \log \frac{2}{\eta}}\right), \tag{36}$$

$$s(\lambda) = \frac{\kappa^2}{\lambda} \left(1 + \frac{\kappa}{\sqrt{\lambda}}\right). \tag{37}$$

In order to prove the first part of Theorem 4.1 we show that, if $\ell' > \ell > 2$, then

$$E(\lambda, \overline{\eta}(\ell'), \ell') > E(\lambda_0(\ell), \overline{\eta}(\ell'), \ell') \qquad \text{for every} \quad \lambda > \lambda_0(\ell), \tag{38}$$

implying that the minimizer of $E(\lambda, \overline{\eta}(\ell'), \ell')$, $\lambda_0(\ell')$, is not greater than $\lambda_0(\ell)$, as claimed. Inequality (38) can be proved considering the identity

$$E(\lambda, \overline{\eta}(\ell'), \ell') = E(\lambda, \overline{\eta}(\ell), \ell) - \big(C(\overline{\eta}(\ell), \ell) - C(\overline{\eta}(\ell'), \ell')\big)s(\lambda). \tag{39}$$

First of all, we observe that, due to the power law behavior of $\overline{\eta}(\ell)$, the function $C(\overline{\eta}(\ell), \ell) = (\delta/\sqrt{\ell})(1 + \sqrt{2 \log(2\ell^p)})$ is strictly decreasing for $\ell > 2$, so that under the conditions on ℓ' and ℓ in the text of Theorem 4.1, the difference $C(\overline{\eta}(\ell), \ell) - C(\overline{\eta}(\ell'), \ell')$ is positive. Moreover, since $s(\lambda)$ is a strictly decreasing function, and due to the definition of $\lambda_0(\ell)$ as the maximum of the minimizers of $E(\lambda, \overline{\eta}(\ell), \ell)$ (Eq. (6)), we can bound the two terms in the previous equality as follows:

$$E(\lambda, \overline{\eta}(\ell'), \ell') > E(\lambda_0(\ell), \overline{\eta}(\ell), \ell) - (C(\overline{\eta}(\ell), \ell) - C(\overline{\eta}(\ell'), \ell'))s(\lambda_0(\ell))$$

for every $\lambda > \lambda_0(\ell)$. Using again (39), the previous inequality reduces to (38) as required.

We now prove the remaining two parts of Theorem 4.1. To this end we produce a sequence $(\overline{\lambda}(\ell))_{\ell=1}^{\infty}$ such that

$$\lim_{\ell \to +\infty} E(\overline{\lambda}(\ell), \overline{\eta}(\ell), \ell) = 0. \tag{40}$$

Since, by the definition of $\lambda_0(\ell)$, $E(\lambda_0(\ell), \overline{\eta}(\ell'), \ell')$ is not greater than $E(\overline{\lambda}(\ell), \overline{\eta}(\ell'), \ell')$, a fortiori the following limit holds

$$\lim_{\ell \to +\infty} E(\lambda_0(\ell), \overline{\eta}(\ell), \ell) = 0. \tag{41}$$

Moreover, since $A(\lambda)$ is bounded from above by $E(\lambda, \ell, \eta)$, $A(\lambda)$ also vanishes for increasing ℓ. From the last fact the second part of the theorem follows recalling that $A(\lambda)$ is an increasing function of λ and $A(0) = 0$ (see at the beginning of this section).

Equation (41) also ensures consistency of the sequence $(\lambda_0(\ell))_{\ell=1}^{\infty}$ since, by our definition of the probabilistic bound $E(\lambda, \ell, \eta)$, we can write

$$\text{Prob}\left\{ D \in Z^{\ell} \middle| \mathcal{R}[f_D^{\lambda_0(\ell)}] > E(\lambda_0(\ell), \overline{\eta}(\ell), \ell) \right\} \leq \overline{\eta}(\ell),$$

and, moreover, $\overline{\eta}(\ell)$ goes to zero as ℓ goes to infinity. It remains to show a sequence verifying (40). Let us choose the following $\overline{\lambda}(\ell)$,

$$\overline{\lambda}(\ell) = \ell^{-q} \qquad \text{with} \quad 0 < q < \tfrac{1}{3}.$$

First, since $\overline{\lambda}(\ell)$ vanishes as ℓ increases, then the approximation term $A(\overline{\lambda}(\ell))$ goes to zero. Moreover, recalling the representation in (35), we have to show that

$$\lim_{\ell \to +\infty} C(\overline{\eta}(\ell), \ell) s(\overline{\lambda}(\ell)) = 0,$$

this fact can be directly verified by substitution of the expressions of $\overline{\lambda}$ and $\overline{\eta}$ into the functions s and C. □

The previous result reduces to the approximation error bound considered in Corollary 4.1. In this case the computable bound E_r given in Corollary 4.1 allows us to obtain the explicit asymptotic form for the selected λ_0.

Theorem 4.2. *Given the sequence* $(\lambda_0^r(\ell))_{\ell=1}^{\infty}$ *such that* $\lambda_0^r(\ell)$ *minimizes the bound* $E_r(\lambda, \overline{\eta}(\ell), \ell)$ *of Corollary* 4.1 *then we have that*:

(1) *for all* $\ell > 0$, $\lambda_0^r(\ell)$ *is finite*;

(2) $\lambda_0^r(\ell) = \ell^{-1/(3+2r)} \left(\dfrac{3\kappa^3 \delta}{2r C_r} (1 + \sqrt{2 \log(2\ell^p)}) \right)^{2/(3+2r)} + O\left(\ell^{-3/2(3+2r)}\right)$.

Proof. As in the proof of Theorem 4.1 we introduce the functions C and s, defined in (37)–(37), to factorize in E_r the dependence on ℓ and η, that is, we write

$$E_r(\lambda, \ell, \eta) = \lambda^r C_r + C(\ell, \eta) s(\lambda).$$

Then, the first part of Theorem 4.2 follows immediately from Proposition 4.2 as a consequence of the divergence of E_r for increasing λ.

The second part can be proved exploiting the differentiability of E_r with respect to λ. Since $\lambda_0^r(\ell)$ is a minimizer of E_r, it must be a zero of its derivative. By explicit differentiation we obtain that $\lambda_0^r(\ell)$ is a solution of the following algebraic equation:

$$2rC_r\,\kappa^{-2}\lambda^{(3+2r)/2} = C(\overline{\eta}(\ell),\ell)(2\lambda^{1/2} + 3\,\kappa). \tag{42}$$

It is convenient to reformulate the last equation in terms of the auxiliary variables $x(\ell)$ and $y(\ell)$ defined by

$$\begin{aligned} x(\ell) &= \lambda C(\overline{\eta}(\ell),\ell)^{-2/(3+2r)}, \\ y(\ell) &= C(\overline{\eta}(\ell),\ell)^{1/(3+2r)}. \end{aligned}$$

Using this notation (42) becomes

$$rC_r\,\kappa^{-2}x(\ell)^{1+r} - \tfrac{3}{2}\,\kappa x(\ell)^{-1/2} = y(\ell). \tag{43}$$

The function of the unknown $x(\ell)$ on the right-hand side of (43) has a positive derivative, moreover, it assumes arbitrary real values as its arguments range the positive real numbers. This fact proves that, for a given ℓ, there exists a unique solution $x(\ell)$. This also implies that E_r has a unique finite minimizer.

Since by definition $y(\ell)$ is $O(\ell^{-1/2(3+2r)})$, we deduce that, for increasing ℓ, $x(\ell)$ goes to the zero of the right-hand side of (43), in fact, to the value

$$x_0 = \left(\frac{3\,\kappa^3}{2rC_r}\right)^{2/(3+2r)}.$$

Finally, due to the regular behavior of the function in (43) in x_0, we can write

$$x(\ell) = x_0 + O(\ell^{-1/2(3+2r)}),$$

that provides the claimed result. □

5. Conclusions

In this paper we focus on a functional analytical approach to learning theory, in the same spirit of [**7**]. Unlike other studies we do not examine the deviation of the empirical error from the expected error, but we analyze directly the expected risk of the solution. As a consequence, the splitting of the expected error in an estimation error term and an approximation error follows naturally giving rise to a bias-variance problem.

In this paper we show a possible way to solve this problem by proposing a model selection criterion that relies on the stability properties of the regularized least-squares algorithm and does not make direct use of any complexity measures. In particular, our estimates depend only on a boundedness assumption on the output space and the kernel.

Our analysis uses extensively the special properties of the square loss function, henceforth it would be interesting to extend our approach to other loss functions. We think that our results may be improved by taking into account more information about the structure of the hypothesis space.

Acknowledgments

We thank, for many useful discussions and suggestions, Michele Piana, Alessandro Verri, and the referees. L. Rosasco is supported by an INFM fellowship. A. Caponnetto is supported by a PRIN fellowship within the project "Inverse Problems in Medical Imaging", n. 2002013422. This research has been partially funded by the INFM Project MAIA, the FIRB Project ASTA, and by the EU Project KerMIT.

Appendix A. Bounding the Approximation Error

In this appendix we report an elementary proof of the following result from approximation theory.

Theorem A.1. *Assume that there is $r \in (0, 1]$ such that Pf_0 is in the domain of L_ν^{-r} and let C_r be a constant such that $\|L_\nu^{-r} Pf_0\|_\nu \leq C_r$. Then*

$$\mathcal{R}[f^\lambda] \leq \lambda^r C_r.$$

Proof. Starting from (21) in Section 3.3 and the definition of C_r we just have to show that

$$\|f^\lambda - Pf_0\|_\nu \leq \lambda^r \|L_\nu^{-r} Pf_0\|_\nu.$$

Due to the fact that K is a Mercer kernel, L_ν is a compact positive operator and, by spectral decomposition of L_ν, there is a sequence $(\varphi_n)_{n=1}^N$ in $L^2(X, \nu)$ (possibly $N = +\infty$) such that,

$$\langle \varphi_n, \varphi_m \rangle_\nu = \delta_{nm},$$

$$L_\nu f = \sum_{n=1}^{N} \sigma_n^2 \langle f, \varphi_n \rangle_\nu \varphi_n,$$

where $\sigma_n \geq \sigma_{n+1} > 0$. In particular, $(\varphi_n)_{i=1}^N$ is an orthonormal basis of the range of P.

Moreover, since the function $g(x) = x^r$ is a concave function on $(0, +\infty)$ and $g'(1) = r$, then

$$x^r \leq r(x - 1) + 1 \leq x + 1. \tag{44}$$

Recalling that

$$f^\lambda = (L_\nu + \lambda)^{-1} L_\nu f_0,$$

$$
\begin{aligned}
\|f^\lambda - Pf_0\|_\nu^2 &= \sum_{n=1}^{N} \langle ((L_\nu + \lambda)^{-1} L_\nu - I) Pf_0, \varphi_n \rangle_\nu^2 \\
&= \sum_{n=1}^{N} \left(\frac{\sigma_n^2}{\sigma_n^2 + \lambda} - 1 \right)^2 \langle Pf_0, \varphi_n \rangle_\nu^2 \\
\left(x_n = \frac{\sigma_n^2}{\lambda} \right) &= \sum_{n=1}^{N} \left(\frac{1}{x_n + 1} \right)^2 \langle Pf_0, \varphi_n \rangle_\nu^2 \\
\text{(Eq. (44))} &\leq \sum_{n=1}^{N} \left(\frac{1}{x_n^r} \right)^2 \langle Pf_0, \varphi_n \rangle_\nu^2 \\
&= \lambda^{2r} \sum_{n=1}^{N} \left(\frac{1}{(\sigma_n^2)^r} \right)^2 \langle Pf_0, \varphi_n \rangle_\nu^2 \\
&= \lambda^{2r} \|L_\nu^{-r} Pf_0\|_\nu^2. \qquad \square
\end{aligned}
$$

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *J. ACM* **44** (1997), 615–631.

[2] N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.* **68** (1950), 337–404.

[3] P. Bartlett, S. Boucheron, and G. Lugosi, Model selection and error estimation, *Machine Learning* **48** (2002), 85–113.

[4] O. Bousquet and A. Elisseeff, Stability and generalization, *J. Mach. Learn. Res.* **2** (2002), 499–526.

[5] N. Cristianini and J. Shawe Taylor, *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, 2000.

[6] 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.

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

[8] L. Devroye, L. Györfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer-Verlag, New York, 1996.

[9] R. Dudley, *Real Analysis and Probability*, Cambridge University Press, Cambridge, 2002.

[10] T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. Comput. Math.* **13** (2000), 1–50.

[11] F. Girosi, M. Jones, and T. Poggio, Regularization theory and neural networks architectures, *Neural Comput.* **7** (1995), 219–269.

[12] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, Springer-Verlag, New York, 2001.

[13] C. McDiarmid, On the method of bounded differences, in *Surveys in Combinatorics*, 1989 (Norwich, 1989), Cambridge University Press, Cambridge, 1989, pp. 148–188.

[14] S. Mendelson, A few notes on statistical learning theory, in *Advanced Lectures in Machine Learning* (S. Mendelson and A. Smola, eds.), Springer-Verlag, 2003, pp. 1–40.

[15] P. Niyogi and F. Girosi, Generalization bounds for function approximation from scattered noisy data, *Adv. Comput. Math.* **10** (1999), 51–80.

[16] T. Poggio and F. Girosi, Networks for approximation and learning, in *Proceedings of the IEEE*, Vol. 78, 1990, pp. 1481–1497.

[17] T. Poggio, R. Rifkin, S. Mukherjee, and P. Niyogi, General conditions for predictivity in learning theory, *Nature* **428** (2004), 419–422.
[18] L. Rosasco, E. De Vito, A. Caponnetto, M. Piana, and A. Verri, Are loss functions all the same?, *Neural Comput*. **16** (2004), 1063–1076.
[19] S. Smale and D.-X. Zhou, Estimating the approximation error in learning theory, *Anal*. *Appl*. (*Singap*.) **1** (2003), 17–41.
[20] I. Steinwart, Consistency of support vector machines and other regularized kernel machines, Technical Report 02-03, University of Jena, Department of Mathematics and Computer Science, 2002.
[21] V. Vapnik, *Statistical learning theory*, Wiley, New York, 1998.
[22] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, PA, 1990.
[23] T. Zhang, Leave-one-out bounds for kernel methods, *Neural Comput*. **15** (2003), 1397–1437.
[24] D.-X. Zhou, The covering number in learning theory, *J*. *Complexity* **18** (2002), 739–767.