BULLETIN (New Series) OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 39, Number 1, Pages 1–49
S 0273-0979(01)00923-5
Article electronically published on October 5, 2001

ON THE MATHEMATICAL FOUNDATIONS OF LEARNING

FELIPE CUCKER AND STEVE SMALE

The problem of learning is arguably at the very core of the problem of intelligence, both biological and artificial.

T. Poggio and C.R. Shelton

Introduction

(1) A main theme of this report is the relationship of approximation to learning and the primary role of sampling (inductive inference). We try to emphasize relations of the theory of learning to the mainstream of mathematics. In particular, there are large roles for probability theory, for algorithms such as *least squares*, and for tools and ideas from linear algebra and linear analysis. An advantage of doing this is that communication is facilitated and the power of core mathematics is more easily brought to bear.

We illustrate what we mean by learning theory by giving some instances.

(a) The understanding of language acquisition by children or the emergence of languages in early human cultures.
(b) In Manufacturing Engineering, the design of a new wave of machines is anticipated which uses sensors to sample properties of objects before, during, and after treatment. The information gathered from these samples is to be analyzed by the machine to decide how to better deal with new input objects (see [43]).
(c) Pattern recognition of objects ranging from handwritten letters of the alphabet to pictures of animals, to the human voice.

Understanding the laws of learning plays a large role in disciplines such as (Cognitive) Psychology, Animal Behavior, Economic Decision Making, all branches of Engineering, Computer Science, and especially the study of human thought processes (how the brain works).

Mathematics has already played a big role towards the goal of giving a universal foundation of studies in these disciplines. We mention as examples the theory of Neural Networks going back to McCulloch and Pitts [25] and Minsky and Papert [27], the PAC learning of Valiant [40], Statistical Learning Theory as developed by Vapnik [42], and the use of reproducing kernels as in [17] among many other mathematical developments. We are heavily indebted to these developments. Recent discussions with a number of mathematicians have also been helpful. In

Received by the editors April 2000, and in revised form June 1, 2001.
2000 *Mathematics Subject Classification.* Primary 68T05, 68P30.
This work has been substantially funded by CERG grant No. 9040457 and City University grant No. 8780043.

particular this includes Gregorio Malajovich, Massimiliano Pontil, Yuan Yao, and especially Ding-Xuan Zhou.

(2) We now describe some cases of learning where we have simplified to the extreme.

Case 1. A classical example of learning is that of learning a physical law by curve fitting to data. Assume that the law at hand, an unknown function $f:\mathbb{R}\to\mathbb{R}$, has a specific form and that the space of all functions having this form can be parameterized by N real numbers. For instance, if f is assumed to be a polynomial of degree d, then $N=d+1$ and the parameters are the unknown coefficients $w_0,\ldots,w_d$ of f. In this case, finding the *best fit* by the *least squares method* estimates the unknown f from a set of pairs $(x_1,y_1),\ldots,(x_m,y_m)$. If the measurements generating this set were exact, then $f(x_i)$ would be equal to y_i. But in general one expects the values y_i to be affected by noise. One computes the vector of coefficients w such that the value

$$\sum_{i=1}^{m}(f_w(x_i)-y_i)^2, \qquad \text{with} \qquad f_w(x)=\sum_{j=0}^{d} w_j x^j$$

is minimized where, typically, $m>N$. In general, the value above is not minimized at 0. The least squares technique, going back to Gauss and Legendre, which is computationally efficient and relies on numerical linear algebra, solves this minimization problem.

In some contexts the x_i, rather than being chosen, are also generated by a probability measure. Thus, one might take as a starting point, instead of the unknown f, a probability measure on $\mathbb{R}$ varying with $x\in\mathbb{R}$. Then y_i is a sample for a given x_i. The starting point could be even a single measure on $\mathbb{R}\times\mathbb{R}$ from which the pairs (x_i,y_i) are randomly drawn. The latter is the point of view taken here.

A more general form of the functions in our approximating class could be given by

$$f_w(x)=\sum_{i=1}^{N} w_i\phi_i(x)$$

where the ϕ_i are part of a "preconditioning step". This is reminiscent of neural nets where the w_i are the weights to be adjusted by "training".

Case 2. A standard example of pattern recognition is that of recognizing handwritten characters. Consider the problem of classifying handwritten letters of the English alphabet. Here, elements in our space X could be matrices with entries in the interval $[0,1]$ —each entry representing a pixel in a certain grey scale of a photo of the handwritten letter or some features extracted from the letters. We may take Y to be

$$Y=\left\{ y\in\mathbb{R}^{26} \mid y=\sum_{i=1}^{26}\lambda_i e_i \quad \text{s.t.} \quad \sum_{i=1}^{26}\lambda_i=1 \right\}.$$

Here e_i is the ith coordinate vector in $\mathbb{R}^{26}$ (each coordinate corresponding to a letter). If $\Delta\subset Y$ is the set of points y as above such that $0\le\lambda_i\le 1$, for $i=1,\ldots,26$, one can interpret a point in Δ as a probability measure on the set $\{\mathtt{A},\mathtt{B},\mathtt{C},\ldots,\mathtt{X},\mathtt{Y},\mathtt{Z}\}$. The problem is to learn the ideal function $f:X\to Y$ which associates, to a given handwritten letter x, the point $\{\mathrm{Prob}\{x=\mathtt{A}\},\mathrm{Prob}\{x=\mathtt{B}\},\ldots,$ $\mathrm{Prob}\{x=\mathtt{Z}\}\}$. Non-ambiguous letters are mapped into a coordinate vector, and in the (pure) classification problem f takes values on these e_i. "Learning f" means to find a sufficiently good approximation of f within a given prescribed class.

The approximation of f is constructed from a set of samples of handwritten letters, each of them with a label in Y. The set $\{(x_1, y_1), \ldots, (x_m, y_m)\}$ of these m samples is randomly drawn from $X \times Y$ according to a measure ρ on $X \times Y$, and the function f to be learned is the regression function f_ρ of ρ. That is, $f_\rho(x)$ is the average of the y values of $\{x\} \times Y$ (we will be more precise about ρ and the regression function in Section 1 in the next chapter).

Case 3 (Monte Carlo integration). An early instance of randomization used in algorithms is for computing integrals. Let $f : [0,1]^n \to \mathbb{R}$. A way of approximating the integral $\int_{x \in [0,1]^n} f(x)dx$ consists of randomly drawing points $x_1, \ldots, x_m \in [0,1]^n$ and computing

$$I_m(f) = \frac{1}{m} \sum_{i=1}^{m} f(x_i).$$

Under mild conditions on f, $I_m(f) \to \int f$ with probability 1; i.e., for all $\varepsilon > 0$,

$$\lim_{m \to \infty} \operatorname*{Prob}_{x_1, \ldots, x_m} \left\{ \left| I_m(f) - \int f \right| > \varepsilon \right\} \to 0.$$

We find again the theme of learning an object (here a single real number, although defined in a non-trivial way through f) from a sample. In this case the measure governing the sample is known (the measure in $[0,1]^n$ inherited from the standard Lebesgue measure on $\mathbb{R}^n$), but the same idea can be used for an unknown measure. If ρ_X is a probability measure on $X \subset \mathbb{R}^n$, a domain or manifold, $I_m(f)$ will approximate $\int_{x \in X} f(x) d\rho_X$, for large m with high probability, as long as the points $x_1, \ldots, x_m$ are drawn from X according to the measure ρ_X.

Case 4. The approximation of characteristic (or indicator) functions of sets is known as PAC learning (from Probably Approximately Correct). Let T (the *target concept*) be a subset of $\mathbb{R}^n$ and ρ_X be a probability measure on $\mathbb{R}^n$ which we assume is not known in advance. Intuitively, a set $S \subset \mathbb{R}^n$ approximates T when the symmetric difference $S \Delta T = (S - T) \cup (T - S)$ is small, i.e. has a small measure. Note that if f_S and f_T denote the characteristic functions of S and T respectively, this measure, called the *error of* S, is $\int_{\mathbb{R}^n} (f_S - f_T)^2 d\rho_X$.

Let $\mathcal{C}$ be a class of subsets of $\mathbb{R}^n$ and assume that $T \in \mathcal{C}$. A strategy to construct an approximation of T is the following. First, draw points $x_1, \ldots, x_m \in \mathbb{R}^n$ according to ρ_X and label each of them with 1 or 0 according to whether or not they belong to T. Secondly, compute any function $f_S : \mathbb{R}^n \to \{0,1\}$, $f_S \in \mathcal{C}$, which coincides with the labeling above over $\{x_1, \ldots, x_m\}$. Such a function will provide a good approximation S of T as long as m is large enough and $\mathcal{C}$ is not too wild. Thus the measure ρ_X is used in both capacities, governing the sample drawing and measuring the error set $S \Delta T$.

A major goal in PAC learning is to estimate as a function of ε and δ how large m needs to be to obtain an ε approximation of T with probability at least $1 - \delta$.

A common characteristic of the cases above is the existence of both an "unknown" function $f : X \to Y$ and a probability measure allowing one to randomly draw points in $X \times Y$. That measure can be on X (Cases 3 and 4), on Y varying with $x \in X$ (Case 1), or on the product $X \times Y$ (Case 2). It can be known (Case 3) or unknown. The only requirement it satisfies is that, if for $x \in X$ a point $y \in Y$ can be randomly drawn, then the expected value of y is $f(x)$.

The development in this paper, for reasons of unity and generality, will be based upon a single measure on $X \times Y$. Yet, one should keep in mind the distinction between "inputs" $x \in X$ and "outputs" $y \in Y$.

In the sequel, we will try to give a rigorous development of what we have found to be the central ideas of learning theory. However, learning theory in its various forms is vast, and we don't even touch on important parts such as "unsupervised learning", relations with dynamics, with neural nets, and so on. "Classification" is not covered directly. However, this report could be of use in further foundational studies in these areas.

Since the readers will have diverse mathematical backgrounds, we sketch the proofs of some standard theorems, with references to the literature for fuller accounts. When the result is new, we are more complete.

Practical results are not the goal of this paper. Understanding is. We try to write in the spirit of H. Weyl and J. von Neumann's contributions to the foundations of quantum mechanics.

Chapter I: Sample Error

1. A formal setting: The probability measure on the product space and the error

Since we want to study learning from random sampling, the primary object in our development is a probability measure ρ governing the sampling and which is not known in advance (however, the goal is not to reveal ρ).

Let X be a compact domain or a manifold in Euclidean space and $Y = \mathbb{R}^k$. For convenience we will take $k = 1$ for the time being. Let ρ be a Borel probability measure on $Z = X \times Y$ whose regularity properties will be assumed as needed. In the following we try to utilize concepts formed naturally and solely from X, Y and ρ.

Throughout this paper, if ξ is a random variable, i.e. a real valued function on a probability space Z, we will use $\mathbf{E}(\xi)$ to denote the expected value (or average, or mean) of ξ and $\sigma^2(\xi)$ to denote its variance. Thus

$$\mathbf{E}(\xi) = \int_Z \xi \, d\rho \qquad \text{and} \qquad \sigma^2(\xi) = \mathbf{E}((\xi - \mathbf{E}(\xi))^2) = \mathbf{E}(\xi^2) - (\mathbf{E}(\xi))^2.$$

A main concept is the *error* (or *least squares error*) of f defined by

$$\mathcal{E}(f) = \mathcal{E}_\rho(f) = \int_Z (f(x) - y)^2 \quad \text{for } f : X \to Y.$$

For each input $x \in X$ and output $y \in Y$, $(f(x) - y)^2$ is the error suffered from the use of f as a model for the process producing y from x. By integrating over $X \times Y$ (w.r.t. ρ, of course) we average out the error over all pairs (x, y). Hence the word "error" for $\mathcal{E}(f)$.

The problem is posed: *What is the f which minimizes the error $\mathcal{E}(f)$?*

The error $\mathcal{E}(f)$ naturally decomposes as a sum. Let us see how.

For every $x \in X$, let $\rho(y|x)$ be the conditional (w.r.t. x) probability measure on Y and ρ_X be the marginal probability measure on X, i.e. the measure on X defined by $\rho_X(S) = \rho(\pi^{-1}(S))$ where $\pi : X \times Y \to X$ is the projection. Notice that ρ,

$\rho(y|x)$ and ρ_X are related as follows. For every integrable function $\varphi : X \times Y \to \mathbb{R}$ a version of Fubini's Theorem states that

$$\int_{X\times Y} \varphi(x,y)\, d\rho = \int_X \left(\int_Y \varphi(x,y)\, d\rho(y|x) \right) d\rho_X .$$

This "breaking" of ρ into the measures $\rho(y|x)$ and ρ_X corresponds to looking at Z as a product of an input domain X and an output set Y. In what follows, unless otherwise specified, integrals are to be understood over ρ, $\rho(y|x)$ or ρ_X.

Define $f_\rho : X \to Y$ by

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

The function f_ρ is called the *regression function* of ρ. For each $x \in X$, $f_\rho(x)$ is the average of the y coordinate of $\{x\} \times Y$ (in topological terms, the average of y on the fiber of x). Regularity hypotheses on ρ will induce regularity properties on f_ρ.

We will assume throughout this paper that f_ρ is bounded.

Fix $x \in X$ and consider the function from Y to $\mathbb{R}$ mapping y into $(y - f_\rho(x))$. Since the expected value of this function is 0, its variance is

$$\sigma^2(x) = \int_Y (y - f_\rho(x))^2 d\rho(y|x).$$

Averaging over X, define

$$\sigma_\rho^2 = \int_X \sigma^2(x)\, d\rho_X = \mathcal{E}(f_\rho).$$

The number σ_ρ^2 is a measure of how well conditioned ρ is, analogous to the notion of condition number in numerical linear algebra.

Remark 1. (a) It is important to note that, while ρ and f_ρ are mainly "unknown", ρ_X is known in some situations and can even be the Lebesgue measure on X inherited from Euclidean space (as in Case 1 above).

(b) In the rest of this paper, if formulas do not make sense or ∞ appears, then the assertions where these formulas occur should be considered vacuous.

Proposition 1. *For every $f : X \to Y$,*

$$\mathcal{E}(f) = \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.$$

Proposition 1 has the following consequence:

The first term in the right-hand side of Proposition 1 provides an average (over X) of the error suffered from the use of f as a model for f_ρ. In addition, since σ_ρ^2 is independent of f, Proposition 1 implies that f_ρ has the smallest possible error among all functions $f : X \to Y$. Thus σ_ρ^2 represents a lower bound on the error $\mathcal{E}$, and it is due solely to our primary object, the measure ρ.

Thus, Proposition 1 supports:

> *The goal is to "learn" (i.e. to find a good approximation of) f_ρ from random samples on Z.*

Proof of Proposition 1. We have

$$\begin{aligned}
\mathcal{E}(f) &= \int_Z (f(x) - f_\rho(x) + f_\rho(x) - y)^2 \\
&= \int_X (f(x) - f_\rho(x))^2 + \int_X \int_Y (f_\rho(x) - y)^2 \\
&\qquad\qquad + 2\int_X \int_Y (f(x) - f_\rho(x))(f_\rho(x) - y) \\
&= \int_X (f(x) - f_\rho(x))^2 + \sigma_\rho^2.
\end{aligned}$$

□

We now consider the sampling. Let

$$\mathbf{z} \in Z^m, \qquad \mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$$

be a *sample* in Z^m, i.e. m examples independently drawn according to ρ. Here Z^m denotes the m-fold Cartesian product of Z. We define the *empirical error* of f (w.r.t. $\mathbf{z}$) to be

$$\mathcal{E}_\mathbf{z}(f) = \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

If ξ is a random variable on Z, we denote the *empirical mean* of ξ (w.r.t. $\mathbf{z}$) by $\mathbf{E}_\mathbf{z}(\xi)$. Thus,

$$\mathbf{E}_\mathbf{z}(\xi) = \frac{1}{m} \sum_{i=1}^{m} \xi(z_i).$$

For any function $f : X \to Y$ we denote by f_Y the function

$$\begin{aligned}
f_Y : X \times Y &\to Y \\
(x, y) &\mapsto f(x) - y.
\end{aligned}$$

With these notations we may write $\mathcal{E}(f) = \mathbf{E}(f_Y^2)$ and $\mathcal{E}_\mathbf{z}(f) = \mathbf{E}_\mathbf{z}(f_Y^2)$. We already remarked that the expected value of $f_{\rho Y}$ is 0; we now remark that its variance is σ_ρ^2.

Remark 2. Consider the setting of PAC learning discussed in Case 4 where $X = \mathbb{R}^n$. The measure ρ_X described there can be extended to a measure ρ on Z by defining, for $A \subset Z$,

$$\rho(A) = \rho_X(\{x \in X \mid (x, f_T(x)) \in A\}).$$

The marginal measure on X of ρ is our original ρ_X. In addition, $\sigma_\rho^2 = 0$, the error above specializes to the error mentioned in that discussion, and the regression function f_ρ of ρ coincides with f_T except for a set of measure zero in X.

2. Convergence in probability

Toward the proof of our main Theorems B and C we recall some basic inequalities in probability theory. The first one, Chebyshev's inequality, is classical. For a proof of the second one, which is an exponential extension of Chebyshev's inequality for bounded random variables, see [32].

Proposition 2. *Let ξ be a random variable on a probability space Z with mean $\mathbf{E}(\xi) = \mu$ and variance $\sigma^2(\xi) = \sigma^2$.*

[Chebyshev] *For all $\varepsilon > 0$*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \varepsilon\right\} \leq \frac{\sigma^2}{m\varepsilon^2}.$$

[Bernstein] *If $|\xi(z) - \mathbf{E}(\xi)| \leq M$ for almost all $z \in Z$, then, for all $\varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \varepsilon\right\} \leq 2e^{-\frac{m\varepsilon^2}{2\left(\sigma^2+\frac{1}{3}M\varepsilon\right)}}.$$

□

Remark 3. (i) The inequalities in Proposition 2 can be seen as quantitative versions of the law of large numbers.

(ii) Bernstein's inequality without the absolute value provides a bound without the first 2, i.e. $e^{-\frac{m\varepsilon^2}{2\left(\sigma^2+\frac{1}{3}M\varepsilon\right)}}$ (see [32]).

(iii) Another exponential version of Chebyshev's inequality, due to Hoeffding, is often used in the learning literature. With the notations used in the statement of Proposition 2, Hoeffding's inequality reads

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\left|\frac{1}{m}\sum_{i=1}^{m}\xi(z_i) - \mu\right| \geq \varepsilon\right\} \leq 2e^{-\frac{m\varepsilon^2}{2M^2}}.$$

Notice that when we replace σ^2 by its obvious bound M^2, the exponent in Bernstein's inequality becomes

$$-\frac{m\varepsilon^2}{2M^2 + \frac{2}{3}M\varepsilon}$$

which is slightly worse than Hoeffding's. Since we may assume $\varepsilon \leq M$ (otherwise the probability in the statement is zero) we have $2M^2 + \frac{2}{3}M\varepsilon \leq 2M^2(1 + 1/3)$. It follows that this exponent is multiplied by a factor of at most $3/4$. However, in the other extreme, when $\sigma^2 = 0$, the exponent in Bernstein's inequality becomes

$$-\frac{3m\varepsilon}{2M}$$

which is much better than the exponent in Hoeffding's inequality.

We also note that Chebyshev's inequality yields a better bound than both Bernstein's and Hoeffding's for small m.

Let $f : X \to Y$. The *defect function* of f is

$$L_{\mathbf{z}}(f) = L_{\rho,\mathbf{z}}(f) = \mathcal{E}(f) - \mathcal{E}_{\mathbf{z}}(f).$$

Notice that the theoretical error $\mathcal{E}(f)$ cannot be measured directly while $\mathcal{E}_{\mathbf{z}}(f)$ can. A bound on $L_{\mathbf{z}}(f)$ becomes useful since it allows one to bound the actual error from an observed quantity.

Our first main result, Theorem A, states bounds for $\operatorname{Prob}\{|L_{\mathbf{z}}(f)| \leq \varepsilon\}$ for a single function $f : X \to Y$. This bound follows from Proposition 2 by taking $\xi = f_Y^2$.

Theorem A. *Let $M > 0$ and $f : X \to Y$ be such that $|f(x) - y| \le M$ almost everywhere. Then, for all $\varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|L_{\mathbf{z}}(f)| \le \varepsilon\} \ge 1 - 2e^{-\frac{m\varepsilon^2}{2\left(\sigma^2+\frac{1}{3}M^2\varepsilon\right)}}$$

where σ^2 is the variance of f_Y^2. □

Remark 4. (1) Note that the *confidence* (i.e. the right hand side in the inequality above) is positive when m is larger than $\frac{2(\sigma^2+\frac{1}{3}M^2\varepsilon)}{\varepsilon^2}$ and approaches 1 exponentially fast with m.

(2) A case implying the condition $|f(x) - y| \le M$ a.e. is the following. Define

$$M_\rho = \inf\left\{\overline{M} \ge 0 \mid \{(x,y) \in Z \mid |y - f_\rho(x)| \ge \overline{M}\} \text{ has measure zero}\right\}.$$

Then take $M = P + M_\rho$ where $P \ge \|f - f_\rho\|_\infty = \sup_{x\in X} |f(x) - f_\rho(x)|$.

3. Hypothesis spaces and target functions

Learning processes do not take place in a vacuum. Some structure needs to be present at the beginning of the process. The nature of this structure in the instance of language acquisition mentioned in the introduction is a subject of debate among linguists. In our formal development, we will assume that this structure takes the form of a class of functions. The goal of the learning process will thus be to find the best approximation of f_ρ within this class. Therefore, we now move the focus from a function $f : X \to Y$ to a family $\mathcal{H}$ of such functions.

Let $\mathcal{C}(X)$ be the Banach space of continuous functions on X with the norm

$$\|f\|_\infty = \sup_{x\in X} |f(x)|.$$

We consider a compact subset $\mathcal{H}$ of $\mathcal{C}(X)$ —in the sequel called *hypothesis space*— where algorithms will work to find, as well as possible, the best approximation for f_ρ. A main choice in our paper is a compact, infinite dimensional, subset of $\mathcal{C}(X)$, but we will also consider closed balls in finite dimensional subspaces of $\mathcal{C}(X)$. It is important for us to choose $\mathcal{H}$ in this way so that the existence of $f_{\mathcal{H}}$ and $f_{\mathbf{z}}$ (see below) is guaranteed, Proposition 3 below can be proved, and covering numbers are finite (see Section 4).

If $f_\rho \in \mathcal{H}$, simplifications will occur. But in general, we will not even assume that $f_\rho \in \mathcal{C}(X)$, and we will have to consider a *target function* $f_{\mathcal{H}}$ in $\mathcal{H}$.

Let $f_{\mathcal{H}}$ be a function minimizing the error $\mathcal{E}(f)$ over $f \in \mathcal{H}$, i.e. an optimizer of

$$\min_{f\in\mathcal{H}} \int_Z (f(x) - y)^2.$$

Notice that, since $\mathcal{E}(f) = \int_X (f - f_\rho)^2 + \sigma_\rho^2$, $f_{\mathcal{H}}$ is also an optimizer of

$$\min_{f\in\mathcal{H}} \int_X (f - f_\rho)^2.$$

The existence of $f_{\mathcal{H}}$ follows from the compactness of $\mathcal{H}$ and the continuity of $\mathcal{E} : \mathcal{C}(X) \to \mathbb{R}$ (see Remark 7 below). It is not necessarily unique. However, we will see a uniqueness result in Section 7 when $\mathcal{H}$ is convex.

Let $\mathbf{z} \in Z^m$ be a sample. We define the *empirical target function* $f_{\mathcal{H},\mathbf{z}} = f_{\mathbf{z}}$ to be a function minimizing the empirical error $\mathcal{E}_{\mathbf{z}}(f)$ over $f \in \mathcal{H}$, i.e. an optimizer of

$$\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2.$$

Note that while $f_{\mathbf{z}}$ is not produced by an algorithm, it is close to algorithmic. It is "empirical" from its dependence on the sample $\mathbf{z}$. The existence of $f_{\mathbf{z}}$ follows from the compactness of $\mathcal{H}$ and the continuity of $\mathcal{E}_{\mathbf{z}}$ where the use of $\| \, \|_\infty$ is now crucial (again, see Remark 7 below). Observe that $f_{\mathbf{z}}$ does not depend on ρ. Note also that $\mathcal{E}(f_{\mathbf{z}})$ and $\mathcal{E}_{\mathbf{z}}(f)$ are different objects, as are $\mathcal{E}(f_{\mathcal{H}})$ and $\mathcal{E}_{\mathcal{H}}(f)$ below.

For a given hypothesis space $\mathcal{H}$, the *error in* $\mathcal{H}$ of a function $f \in \mathcal{H}$ is the normalized error

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

Note that $\mathcal{E}_{\mathcal{H}}(f) \geq 0$ for all $f \in \mathcal{H}$ and that $\mathcal{E}_{\mathcal{H}}(f_{\mathcal{H}}) = 0$.

Continuing the discussion after Proposition 1, note that it follows from our definitions and that proposition that

$$\mathcal{E}(f_z) = \mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) + \mathcal{E}(f_{\mathcal{H}}) = \int_X (f_{\mathbf{z}} - f_\rho)^2 + \sigma_\rho^2. \tag{1}$$

Consider the sum $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) + \mathcal{E}(f_{\mathcal{H}})$. The second term in this sum depends on the choice of $\mathcal{H}$ but is independent of sampling. We will call it the *approximation error*. The first term, $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})$, is called the *sample error*.[1]

Equation (1) thus breaks our goal —to estimate $\int_X (f_{\mathbf{z}} - f_\rho)^2$ or, equivalently, $\mathcal{E}(f_{\mathbf{z}})$— into two different problems corresponding to finding estimates for the sample and approximation errors. Note that the first problem is posed on the space $\mathcal{H}$ and the second is independent of the sample $\mathbf{z}$. For fixed $\mathcal{H}$ the sample error decreases when the number m of examples increases (as we will see in Theorem C). Fix m instead. Then, typically, the approximation error will decrease when enlarging $\mathcal{H}$, but the sample error will increase. This latter feature is sometimes called the *bias-variance trade-off* (see e.g. [6] and page 41 in [28]). The "bias" is the approximation error and the "variance" is the sample error. This suggests the problem of how to choose $\dim \mathcal{H}$ (or another measure of the size of $\mathcal{H}$) when m is fixed. We will examine this problem in the next chapter. The focus of this chapter is on estimating the sample error. We want to estimate how close one may expect $f_{\mathbf{z}}$ and $f_{\mathcal{H}}$ to be, depending on the size of the sample and with a given confidence. Or, equivalently,

> *How many examples do we need to draw to assert, with a confidence greater than $1 - \delta$, that $\int_X (f_{\mathbf{z}} - f_{\mathcal{H}})^2$ is not more than ε?*

There have been many results in recent years doing this (cf. [18], [42]). Our main results in this chapter, Theorems C and C* below, give such estimates in a general and sharp setting.

We now describe some examples of hypothesis spaces. Our development in this and the next chapter will be accompanied by the development of these examples.

Example 1 (Homogeneous polynomials)**.** Let $\mathcal{H}_d = \mathcal{H}_d(\mathbb{R}^{n+1})$ be the linear space of homogeneous polynomials of degree d in $x_0, x_1, \ldots, x_n$. Let $X = S(\mathbb{R}^{n+1})$, the

[1] The sample error is often called *estimation error* in the literature.

n-dimensional unit sphere. An element in $\mathcal{H}_d$ defines a function from X to $\mathbb{R}$ and can be written as

$$f = \sum_{|\alpha|=d} w_\alpha x^\alpha.$$

Here, $\alpha = (\alpha_0, \ldots, \alpha_n) \in \mathbb{N}^n$ is a "multi-index", $|\alpha| = \alpha_0 + \cdots + \alpha_n$, and $x^\alpha = x_0^{\alpha_0} \cdots x_n^{\alpha_n}$. Thus, $\mathcal{H}_d$ is a vector space of dimension

$$N = \binom{n+d}{n}.$$

We may consider $\mathcal{H} = \{f \in \mathcal{H}_d \mid \|f\|_\infty \leq 1\}$ as a hypothesis space. Because of the scaling $f(\lambda x) = \lambda^d f(x)$, taking the bound $\|f\|_\infty \leq 1$ causes no loss. The number N is exponential in n and d. We notice however that in some situations one may consider a linear space of polynomials with a given monomial structure, i.e. in which only a prespecified set of monomials may appear.

Example 2 (Finite dimensional function spaces)**.** This generalizes the previous example. Let $\phi_1, \ldots, \phi_N \in \mathcal{C}(X)$ and $\mathbb{E}$ be the linear subspace of $\mathcal{C}(X)$ spanned by $\{\phi_1, \ldots, \phi_N\}$. Here we may take $\mathcal{H} = \{f \in \mathbb{E} \mid \|f\|_\infty \leq R\}$ for some $R > 0$.

The next two examples deal with infinite dimensional linear spaces. In both of them, the space $\mathcal{L}^2_\nu(X)$ of square integrable functions is central.

Let ν be a Borel measure on X and L be the linear space of functions $f : X \to Y$ such that the integral

$$\int_X f^2(x)\, d\nu$$

exists. The space $\mathcal{L}^2_\nu(X)$ is defined to be the quotient of L under the equivalence relation $\equiv$ given by

$$f \equiv g \iff \int_X (f(x) - g(x))^2\, d\nu = 0.$$

This is a Hilbert space with the scalar product

$$\langle f, g\rangle_\nu = \int_X f(x)g(x)\, d\nu.$$

We will denote by $\| \;\|_\nu$ the norm induced by this inner product. In case $\nu = \rho_X$ we will write $\| \;\|_\rho$ instead of the more cumbersome $\| \;\|_{\rho_X}$.

A linear map $J : \mathbb{E} \to \mathbb{F}$ between the Banach spaces $\mathbb{E}$ and $\mathbb{F}$ is called *compact* if the closure $\overline{J(B)}$ of $J(B)$ is compact for any bounded set $B \subset \mathbb{E}$.

Example 3 (Sobolev spaces)**.** Let X be a compact domain in $\mathbb{R}^n$ with smooth boundary. Then, the space $\mathcal{C}^\infty(X)$ of infinitely differentiable functions on X is well-defined. For every $s \in \mathbb{N}$ we can define an inner product in $\mathcal{C}^\infty(X)$ by

$$\langle f, g\rangle_s = \int_X \sum_{|\alpha| \leq s} D^\alpha f D^\alpha g.$$

Here, $\alpha \in \mathbb{N}^n$, $D^\alpha f$ is the partial derivative $\frac{\partial^\alpha f}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}}$, and we are integrating with respect to the Lebesgue measure μ on X inherited from Euclidean space. We will denote by $\| \;\|_s$ the norm induced by $\langle\;,\;\rangle_s$. Notice that when $s = 0$, the inner product above coincides with that of $\mathcal{L}^2_\mu(X)$. In particular, $\| \;\|_0 = \| \;\|_\mu$. We define the Sobolev space $H^s(X)$ to be the completion of $\mathcal{C}^\infty(X)$ with respect to the norm $\| \|_s$. The Sobolev Embedding Theorem asserts that, for $s > n/2$, the inclusion

$$J_s : H^s(X) \hookrightarrow \mathcal{C}(X)$$

is well-defined and bounded. From Rellich's Theorem it follows that this embedding is actually compact. The definition of $H^s(X)$ can be extended to $s \in \mathbb{R}$, $s \geq 0$, by using a Fourier transform argument (see also [38]). A reference for the above is [39].

Thus, if B_R denotes the closed ball of radius R in $H^s(X)$, we may take $\mathcal{H}_{R,s} = \mathcal{H} = \overline{J_s(B_R)}$.

Example 4 (Spaces associated to a kernel)**.** Let $K : X \times X \to \mathbb{R}$ be continuous and symmetric. Assume that, in addition, K is *positive definite*, i.e. that for all finite sets $\{x_1, \dots, x_k\} \subset X$ the $k \times k$ matrix $K[\mathbf{x}]$ whose (i,j) entry is $K(x_i, x_j)$ is positive definite. We will call such function a *Mercer kernel.* Let ν be any Borel measure on X. Let $L_K : \mathcal{L}^2_\nu(X) \to \mathcal{C}(X)$ be the linear operator given by

$$(L_K f)(x) = \int K(x,t) f(t) dt.$$

Then L_K is well-defined, positive, and compact (cf. Section 1 of Chapter III). In Section 3 of Chapter III it is proved that there exists a Hilbert space $\mathcal{H}_K$ of continuous functions on X (called *reproducing kernel Hilbert space*, RKHS for short) associated to K and X and independent of ν such that the linear map $L_K^{1/2}$ is a Hilbert isomorphism between $\mathcal{L}^2_\nu(X)$ and $\mathcal{H}_K$. Here $L_K^{1/2}$ denotes the square root of L_K, i.e. the only linear operator satisfying $L_K^{1/2} \circ L_K^{1/2} = L_K$. Thus, we have the following diagram:

$$\begin{tikzcd}
\mathcal{L}^2_\mu(X) \arrow[r, "L^{1/2}_{K,\mathcal{C}}"] \arrow[dr, "L_K^{1/2}"', "\approx"] & \mathcal{C}(X) \\
& \mathcal{H}_K \arrow[u, "I_K"']
\end{tikzcd}$$

where we write $L_{K,\mathcal{C}}$ to emphasize that the target is $\mathcal{C}(X)$ and I_K denotes the inclusion. In Section 5 of Chapter III we will prove that if K is $\mathcal{C}^\infty$, then I_K is compact. For a $\mathcal{C}^\infty$ Mercer kernel K we may thus consider $\overline{I_K(B_R)}$ as a hypothesis space. This choice will occupy us in Chapter III, where, in particular, Mercer kernels are shown to exist.

Remark 5. The examples above fit into a *general setting* which we will refer to in the sequel. Let $\mathbb{E}$ be a Banach space of functions on X and $J_\mathbb{E} : \mathbb{E} \to \mathcal{C}(X)$ a compact embedding. We then define, for $R > 0$,

$$\mathcal{H} = \mathcal{H}_R = \mathcal{H}_{\mathbb{E},R} = \overline{J_\mathbb{E}(B_R)}$$

where B_R denotes the closed ball of radius R in $\mathbb{E}$. Of course our definition of hypothesis space includes some which do not fit into the general setting.

4. Uniform estimates on the defect

Our second main result, Theorem B, extends Theorem A to families of functions. While Theorem A is an immediate application of Bernstein's inequality, Theorem B is a version of the main uniformity estimate in Statistical Learning Theory as developed by Vapnik (see e.g. [18], [42]). The topology on the family of functions $\mathcal{H}$, in particular via supposing that $\mathcal{H} \subset \mathcal{C}(X)$ and that $\mathcal{H}$ is compact as in Section 3, enables our statement and proof of the uniformity estimates to become quite economical.

Let S be a metric space and $s > 0$. We define the *covering number* $\mathcal{N}(S, s)$ to be the minimal $\ell \in \mathbb{N}$ such that there exist ℓ disks in S with radius s covering S. When S is compact, as in our case, this number is finite.

Theorem B. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \leq M$ almost everywhere. Then, for all $\varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \sup_{f \in \mathcal{H}} |L_{\mathbf{z}}(f)| \leq \varepsilon \right\} \geq 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{8M}\right) 2e^{-\frac{m\varepsilon^2}{4\left(2\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}}.$$

Here $\sigma^2 = \sigma^2(\mathcal{H}) = \sup_{f \in \mathcal{H}} \sigma^2(f_Y^2)$.

Notice the resemblance to Theorem A. The only essential difference is the inclusion of the covering number, which takes into account the extension from a single f to the family $\mathcal{H}$. This has the effect of requiring the sample size m to increase accordingly to achieve the confidence level of Theorem A.

Let $f_1, f_2 \in \mathcal{C}(X)$. We first estimate the quantity

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)|$$

linearly by $\|f_1 - f_2\|_\infty$ for almost all $\mathbf{z} \in Z^m$ (a Lipshitz estimate).

Proposition 3. *If $|f_j(x) - y| \leq M$ on a set $U \subset Z$ of full measure for $j = 1, 2$, then for $\mathbf{z} \in U^m$*

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| \leq 4M\|f_1 - f_2\|_\infty.$$

Proof. First note that since

$$(f_1(x) - y)^2 - (f_2(x) - y)^2 = (f_1(x) - f_2(x))(f_1(x) + f_2(x) - 2y)$$

we have

$$\begin{aligned} |\mathcal{E}(f_1) - \mathcal{E}(f_2)| &= \left| \int (f_1(x) - f_2(x))(f_1(x) + f_2(x) - 2y) \right| \\ &\leq \|f_1 - f_2\|_\infty \int |(f_1(x) - y) + (f_2(x) - y)| \\ &\leq \|f_1 - f_2\|_\infty 2M. \end{aligned}$$

Also, for $\mathbf{z} \in U^m$, we have

$$\begin{aligned} |\mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}_{\mathbf{z}}(f_2)| &= \frac{1}{m} \left| \sum_{i=1}^m (f_1(x_i) - f_2(x_i))(f_1(x_i) + f_2(x_i) - 2y_i) \right| \\ &\leq \|f_1 - f_2\|_\infty \frac{1}{m} \sum_{i=1}^m |(f_1(x_i) - y) + (f_2(x_i) - y_i)| \\ &\leq \|f_1 - f_2\|_\infty 2M. \end{aligned}$$

Thus

$$|L_{\mathbf{z}}(f_1) - L_{\mathbf{z}}(f_2)| = |\mathcal{E}(f_1) - \mathcal{E}_{\mathbf{z}}(f_1) - \mathcal{E}(f_2) + \mathcal{E}_{\mathbf{z}}(f_2)| \leq \|f_1 - f_2\|_\infty 4M.$$

□

Remark 6. Notice that for bounding $|\mathcal{E}_{\mathbf{z}}(f_1)-\mathcal{E}_{\mathbf{z}}(f_2)|$ in the proof above —in contrast with the bound for $|\mathcal{E}(f_1)-\mathcal{E}(f_2)|$— one crucially needs the use of the $\|\ \|_\infty$ norm. Nothing less would do.

Remark 7. Let $\mathcal{H}\subseteq\mathcal{C}(X)$ such that, for all $f\in\mathcal{H}$, $|f(x)-y|\le M$ almost everywhere. Then the bounds $|\mathcal{E}(f_1)-\mathcal{E}(f_2)|\le 2M\|f_1-f_2\|_\infty$ and $|\mathcal{E}_{\mathbf{z}}(f_1)-\mathcal{E}_{\mathbf{z}}(f_2)|\le 2M\|f_1-f_2\|_\infty$ imply that $\mathcal{E},\mathcal{E}_{\mathbf{z}}:\mathcal{H}\to\mathbb{R}$ are continuous.

Lemma 1. *Let $\mathcal{H}=S_1\cup\ldots\cup S_\ell$ and $\varepsilon>0$. Then*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)|\ge\varepsilon\right\}\le\sum_{j=1}^{\ell}\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in S_j}|L_{\mathbf{z}}(f)|\ge\varepsilon\right\}.$$

Proof. It follows from the equivalence

$$\sup_{f\in\mathcal{H}}|L_{\mathbf{z}}(f)|\ge\varepsilon\quad\Longleftrightarrow\quad\exists j\le\ell \text{ s.t. } \sup_{f\in S_j}|L_{\mathbf{z}}(f)|\ge\varepsilon$$

and the fact that the probability of a union of events is bounded by the sum of the probabilities of these events. $\square$

Proof of Theorem B. Let $\ell=\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{4M}\right)$ and consider $f_1,\ldots,f_\ell$ such that the disks D_j centered at f_j and with radius $\frac{\varepsilon}{4M}$ cover $\mathcal{H}$. Let U be a full measure set on which $|f(x)-y|\le M$. By Proposition 3, for all $\mathbf{z}\in U^m$ and all $f\in D_j$,

$$|L_{\mathbf{z}}(f)-L_{\mathbf{z}}(f_j)|\le 4M\|f-f_j\|_\infty\le 4M\frac{\varepsilon}{4M}=\varepsilon.$$

Since this holds for all $\mathbf{z}\in U^m$ and all $f\in D_j$ we get

$$\sup_{f\in D_j}|L_{\mathbf{z}}(f)|\ge 2\varepsilon\Rightarrow|L_{\mathbf{z}}(f_j)|\ge\varepsilon.$$

We conclude that, for $j=1,\ldots,\ell$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in D_j}|L_{\mathbf{z}}(f)|\ge 2\varepsilon\right\}\le\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{|L_{\mathbf{z}}(f_j)|\ge\varepsilon\}\le 2e^{-\frac{m\varepsilon^2}{2\left(\sigma^2(f_{jY}^2)+\frac{1}{3}M^2\varepsilon\right)}}$$

with the last estimate using Theorem A. The statement now follows from Lemma 1 by replacing ε by $\varepsilon/2$. $\square$

Remark 8. We noted in Remark 3 that Bernstein's inequality can be seen as a quantitative instance of the law of large numbers. An "abstract" uniform version of this law can be extracted from the proof of Theorem B.

Proposition 4. *Let $\mathcal{F}$ be a family of functions from a probability space Z to $\mathbb{R}$ and d a distance on $\mathcal{F}$. Let $U\subset Z$ be of full measure such that*

(a) *$|\xi(z)|\le B$ for all $\xi\in\mathcal{F}$ and all $z\in U$, and*
(b) *$|L_{\mathbf{z}}(\xi_1)-L_{\mathbf{z}}(\xi_2)|\le\mathbf{L}\,d(\xi_1,\xi_2)$, for all $\xi_1,\xi_2\in\mathcal{F}$ and all $\mathbf{z}\in U^m$*

where $L_{\mathbf{z}}(f)=\displaystyle\int_Z\xi(f,z)-\frac{1}{m}\sum_{i=1}^m\xi(f,z_i)$. *Then, for all $\varepsilon>0$,*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{\xi\in\mathcal{F}}|L_{\mathbf{z}}(\xi)|\le\varepsilon\right\}\ge 1-\mathcal{N}\left(\mathcal{F},\frac{\varepsilon}{2\mathbf{L}}\right)2e^{-\frac{m\varepsilon^2}{4\left(2\sigma^2+\frac{1}{3}B\varepsilon\right)}}.$$

Here $\sigma^2=\sigma^2(\mathcal{F})=\displaystyle\sup_{\xi\in\mathcal{F}}\sigma^2(\xi)$. $\square$

5. Estimating the sample error

How good can we expect $f_{\mathbf{z}}$ to be as an approximation of $f_{\mathcal{H}}$? Or, in other words, how small can we expect the sample error $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})$ to be? The third main result in this chapter, Theorem C below, gives an answer.

Lemma 2. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Let $\varepsilon > 0$ and $0 < \delta < 1$ such that*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}} |L_{\mathbf{z}}(f)| \le \varepsilon\right\} \ge 1-\delta.$$

Then

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le 2\varepsilon\} \ge 1-\delta.$$

Proof. By hypothesis we have, with probability at least $1-\delta$,

$$\mathcal{E}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \varepsilon$$

and

$$\mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) \le \mathcal{E}(f_{\mathcal{H}}) + \varepsilon.$$

Moreover, since $f_{\mathbf{z}}$ minimizes $\mathcal{E}_{\mathbf{z}}$ on $\mathcal{H}$ we have

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}).$$

Therefore, with probability at least $1-\delta$,

$$\mathcal{E}(f_{\mathbf{z}}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}}) + \varepsilon \le \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) + \varepsilon \le \mathcal{E}(f_{\mathcal{H}}) + 2\varepsilon$$

and thus, $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le 2\varepsilon$. $\square$

Replacing ε by $\varepsilon/2$ in Lemma 2 and using Theorem B, one obtains the following.

Theorem C. *Let $\mathcal{H}$ be a compact subset of $\mathcal{C}(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x)-y| \le M$ almost everywhere. Let*

$$\sigma^2 = \sigma^2(\mathcal{H}) = \sup_{f\in\mathcal{H}} \sigma^2(f_Y^2)$$

where $\sigma^2(f_Y^2)$ is the variance of f_Y^2. Then, for all $\varepsilon > 0$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{16M}\right) 2e^{-\frac{m\varepsilon^2}{8\left(4\sigma^2+\frac{1}{3}M^2\varepsilon\right)}}.$$

$\square$

In case $\mathcal{H}$ is convex Theorem C* in Section 7 improves the dependence on ε. Its Corollary 5 estimates directly $\|f_{\mathbf{z}} - f_{\mathcal{H}}\|_\rho$ as well.

Remark 9. Theorem C helps to deal with the question posed in Section 3. Given $\varepsilon, \delta > 0$, to ensure that

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1-\delta$$

it is sufficient that the number m of examples satisfies

$$m \ge \frac{8\left(4\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}{\varepsilon^2}\left[\ln\left(2\mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{16M}\right)\right) + \ln\left(\frac{1}{\delta}\right)\right]. \tag{2}$$

To prove this, take $\delta = \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{16M}\right) 2e^{-\frac{m\varepsilon^2}{8\left(4\sigma^2+\frac{1}{3}M^2\varepsilon\right)}}$ and solve for m. But note further that (2) gives a relation between the three basic variables ε, δ and m.

6. Estimation of covering numbers

As we have seen, the estimates in Theorems B and C have as a factor the covering numbers $\mathcal{N}(\mathcal{H}, \eta)$. Here we give estimates for this factor in our series of examples.

Our first result estimates the covering number of balls in finite dimensional Banach spaces. Let $\mathbb{E}$ be such a space and denote by B_R the closed ball of radius R centered at the origin, i.e.,

$$B_R = \{x \in \mathbb{E} \mid \|x\| \le R\}.$$

Proposition 5. *Let $N = \dim \mathbb{E}$. Then* $\ln \mathcal{N}(B_R, \eta) \le N \ln \left(\frac{4R}{\eta}\right)$.

Proposition 5 allows one to bound the covering numbers appearing in Example 2. The proof we next give is essentially taken from [9]. We first introduce some numbers occurring in functional analysis.

Let S be a metric space. For $k \ge 1$ define

$$\varepsilon_k(S) = \inf\{\varepsilon > 0 \mid \exists \text{ closed balls } D_1, \dots, D_k \text{ with radius } \varepsilon \text{ covering } S\}.$$

Note that

$$\varepsilon_k(S) \le \eta \iff \mathcal{N}(S, \eta) \le k \tag{3}$$

since both inequalities are equivalent to the existence of a covering of S by k balls of radius η. Also, note that ε_k scales well in the sense that, for all $R > 0$, $\varepsilon_k(RS) = R\varepsilon_k(S)$. Here $RS = \{Rx \mid x \in S\}$.

Also, for $k \ge 1$, define

$$\varphi_k(S) = \sup\{\delta > 0 \mid \exists x_1, \dots, x_{k+1} \in S \text{ s.t. for } i \ne j,\ d(x_i, x_j) > 2\delta\}.$$

Lemma 3. (i) *For all $k \ge 1$, $\varphi_k(S) \le \varepsilon_k(S) \le 2\varphi_k(S)$.*

(ii) *Let $\mathbb{E}$ be a Banach space of dimension N and B_1 the unit ball in $\mathbb{E}$. For all $k \ge 1$, $k^{-\frac{1}{N}} \le \varepsilon_k(B_1) \le 4(k+1)^{-\frac{1}{N}}$.*

Proof. Part (i) is easy to prove. For part (ii), first note that $\varphi_k(B_1) \le 1$ for all $k \in \mathbb{N}$. Let $\rho < \varphi_k(B_1)$. Then there exist $x_1, \dots, x_{k+1}$ such that $d(x_i, x_j) > 2\rho$ for $1 \le i \ne j \le k+1$. Let $D_j = x_j + \rho B_1$, $j = 1, \dots, k+1$. Clearly, $D_i \cap D_j = \emptyset$ if $i \ne j$. In addition, for all $x \in D_j$, $\|x\| \le \|x - x_j\| + \|x_j\| \le \rho + 1 < 2$. Therefore, $D_j \subseteq B_2$.

As a vector space, $\mathbb{E}$ is isomorphic to $\mathbb{R}^N$. Any such isomorphism induces on $\mathbb{E}$ a measure ν which is invariant under translations and is homogeneous of degree N with respect to homotheties (i.e. $\nu(\lambda B) = \lambda^N \nu(B)$ for every measurable set B). Using this measure we get

$$\begin{aligned} \sum_{i=1}^{k+1} \nu(D_i) \le \nu(B_2) &\Rightarrow \sum_{i=1}^{k+1} \rho^N \nu(B_1) \le 2^N \nu(B_1) \\ \Rightarrow \quad (k+1)\rho^N &\le 2^N \Rightarrow \rho \le 2(k+1)^{-\frac{1}{N}}. \end{aligned}$$

From here it follows that $\varepsilon_k(B_1) \le 4(k+1)^{-\frac{1}{N}}$.

For the other inequality in (ii) consider any $\varepsilon > \varepsilon_k(B_1)$. Then there exist closed balls $D_1, \dots, D_k$ of radius ε covering B_1, and consequently $\nu(B_1) \le k\varepsilon^N \nu(B_1)$ which implies $k^{-\frac{1}{N}} \le \varepsilon$. $\square$

Let $x \in \mathbb{R}$. We denote by $\lceil x \rceil$ the largest integer smaller than or equal to x.

Proof of Proposition 5. Let $k = \left\lceil \left(\frac{4R}{\eta}\right)^N - 1 \right\rceil$. Then $k+1 \geq \left(\frac{4R}{\eta}\right)^N$ and

$$4(k+1)^{-\frac{1}{N}} \leq \frac{\eta}{R} \Rightarrow \varepsilon_k(B_1) \leq \frac{\eta}{R} \iff \varepsilon_k(B_R) \leq \eta \iff \mathcal{N}(B_R, \eta) \leq k.$$

From here the statement follows since $k \leq (\frac{4R}{\eta})^N$. □

To deal with Examples 3 and 4 we introduce a logarithmic version of $\varepsilon_k(S)$. For $k \geq 1$ define the kth *entropy number* of a metric space S to be[2]

$$e_k(S) = \inf\{\varepsilon > 0 \mid \exists \text{ closed balls } D_1, \ldots, D_{2^k-1} \text{ with radius } \varepsilon \text{ covering } S\}.$$

If $\mathbb{E}$ and $\mathbb{F}$ are Banach spaces and $T : \mathbb{E} \to \mathbb{F}$ is a linear map, then we define

$$e_k(T) = e_k(T(B_1)).$$

Lemma 4. (a) $e_k(T) \leq \eta \iff \mathcal{N}(T(B_1), \eta) \leq 2^k - 1$, *and*
(b) $e_k(T(B_R)) = Re_k(T)$.

Proof. For (a) note that, using (3),

$$e_k(T) \leq \eta \iff \varepsilon_{2^k-1}(T(B_1)) \leq \eta \iff \mathcal{N}(T(B_1), \eta) \leq 2^k - 1.$$

Part (b) is clear. □

Example 3 (continued). Recall that $H^s(X)$ is a Sobolev space and we are assuming that $s > n/2$ from which it follows that the inclusion

$$J_s : H^s(X) \hookrightarrow \mathcal{C}(X)$$

is a compact embedding. Let B_R be the closed ball of radius R centered at the origin in $H^s(X)$ and $\mathcal{H} = \overline{J_s(B_R)}$ be its image in $\mathcal{C}(X)$.

A main result —of a kind going back to the work of Birman and Solomyak [5]— concerning entropy numbers of Sobolev spaces states that, if $X \subset \mathbb{R}^n$ is a compact domain with smooth ($\mathcal{C}^\infty$) boundary and $s > n/2$, then, for all $k \geq 1$,

$$e_k(J_s) \leq C\left(\frac{1}{k}\right)^{s/n}. \tag{4}$$

For a proof, take $s_1 = s, s_2 = 0, p_1 = 2, p_2 = \infty$ in a very general theorem of Edmunds and Triebel ([16], page 105). Here C is a "constant" independent of k (which depends though on X and s). It would be useful to see this constant bounded explicitly.

Remark 10. In general in this paper, we have tried to estimate the value of the constants occurring in our bounds. In some cases, however, as with the constant C above, we have lost control.

Proposition 6. *Let B_R be the closed ball of radius R centered at the origin in $H^s(X)$ and $\mathcal{H} = \overline{J_s(B_R)}$ be its image in $\mathcal{C}(X)$. Then, for all $\varepsilon > 0$,*

$$\ln \mathcal{N}(\mathcal{H}, \varepsilon) \leq \left(\frac{RC}{\varepsilon}\right)^{n/s} + 1.$$

[2] Sometimes in the literature (e.g. [9]) $\varepsilon_k(S)$ and $\varphi_k(S)$ are called inner and outer entropy numbers respectively. Following [16] we reserve the expression *entropy number* for $e_k(S)$.

Proof. Let $\eta = R\varepsilon$ and $k = \left\lceil \left(\frac{C}{\eta}\right)^{n/s} \right\rceil$. Then $\eta \geq C\left(\frac{1}{k}\right)^{s/n}$. By inequality (4) we thus have $e_k(J_s) \leq \eta$ and therefore, $\mathcal{N}(J_s(B_1), \eta) \leq 2^k - 1$. Hence,

$$\ln \mathcal{N}(J_s(B_R), R\eta) = \ln \mathcal{N}(J_s(B_1), \eta) < k < \left(\frac{RC}{\varepsilon}\right)^{n/s} + 1.$$

□

In the use of Proposition 6 we may and will delete the constant 1 by supposing C is slightly enlarged. Proposition 6 can be generalized to other function spaces via the mentioned result in [16].

Example 4 (continued). Recall that $K : X \times X \to \mathbb{R}$ is a $\mathcal{C}^\infty$ Mercer kernel and

$$I_K : \mathcal{H}_K \to \mathcal{C}(X)$$

is the compact embedding defined by K. The following result will be proved in Section 5 of Chapter III. Let B_R be the ball of radius R in $\mathcal{H}_K$. Then, for all $h > n$, $\eta > 0$, and $R > 0$,

$$\ln \mathcal{N}\left(\overline{I_K(B_R)}, \eta\right) \leq \left(\frac{RC_h}{\eta}\right)^{\frac{2n}{h}}$$

where C_h is a constant independent of η and R.

As a consequence the sample error satisfies that given $\varepsilon, \delta > 0$, in order to have

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \leq \varepsilon\} \geq 1 - \delta$$

it is enough that the number m of examples satisfies

$$m \geq \frac{8\left(4\sigma^2 + \frac{1}{3}M^2\varepsilon\right)}{\varepsilon^2} \left[\left(\frac{16MRC_h}{\varepsilon}\right)^{\frac{2n}{h}} + 1 + \ln\left(\frac{1}{\delta}\right)\right].$$

Remark 11. In the examples above, seen as particular cases of the general setting, with $J_{\mathbb{E}} : \mathbb{E} \to \mathcal{C}(X)$, we obtain estimates of the entropy numbers for $J_{\mathbb{E}}$ of the form $e_k(J_{\mathbb{E}}) \leq C_{\mathbb{E}}\left(\frac{1}{k}\right)^{\ell_{\mathbb{E}}}$ for some positive constants $C_{\mathbb{E}}$ and $\ell_{\mathbb{E}}$. Actually this estimate is always true if we allow $\ell_{\mathbb{E}}$ to be zero, so, in what follows, we will assume the estimate as a part of the general setting.

Note we thus have, for $\mathcal{H} = \mathcal{H}_{\mathbb{E},R}$, that $\ln \mathcal{N}(\mathcal{H}, \varepsilon) \leq \left(\frac{RC_{\mathbb{E}}}{\varepsilon}\right)^{1/\ell_E}$.

We close this section by noting that the use of entropy numbers in learning theory has been discussed in [46]. On the other hand, entropy numbers have a strong history in related contexts (see [21], [44], [24], [41]). See also [45] for contributions to these matters coming from statistics.

7. Convex hypothesis spaces

A simple computation shows that in the noise-free case, i.e. when $\sigma_\rho^2 = 0$, one has that, for all $f \in \mathcal{L}_\rho^2(X)$, $\sigma^2(f_Y^2) = 0$. It follows that $\sigma_{\mathcal{H}}^2 = 0$ and the exponent in the bound in Theorem C becomes $\frac{3m\varepsilon}{8M^2}$. Thus the dependency on ε of this exponent passes from quadratic to linear. In several situations, notably in those covered in the general setting described in Remark 5, the hypothesis space $\mathcal{H}$ is convex. In this case, in Theorem C* below, at the cost of worsening the constant 3/8 above, we are able to obtain such a linear dependency on ε without assuming $\sigma_\rho^2 = 0$. In a related context, [3], [22] have shown a similar passage from ε^2 to ε.

Theorem C*. *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(X)$. Assume that, for all $f \in \mathcal{H}$, $|f(x) - y| \le M$ almost everywhere. Then, for all $\varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m} \{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{24M}\right) e^{-\frac{m\varepsilon}{288M^2}}.$$

Theorem C* applies to Examples 1 to 4. Before proceeding with the proof of Theorem C* we revisit these examples.

Example 2 (continued). Let $\phi_1, \ldots, \phi_N \in \mathcal{C}(X)$, $\mathbb{E}$ be the subspace of $\mathcal{C}(X)$ spanned by $\{\phi_1, \ldots, \phi_N\}$ and $\mathcal{H} = \{f \in \mathbb{E} \mid \|f\|_\infty \le R\}$ for some $R > 0$. As in Remark 9, given $\varepsilon, \delta > 0$, to have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m} \{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \delta,$$

it is sufficient that the number m of examples satisfies

$$m \ge \frac{288M^2}{\varepsilon}\left[N \ln\left(\frac{96RM}{\varepsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right].$$

This follows from Theorem C* together with Proposition 5.

Example 3 (continued). Recall that $H^s(X)$ is a Sobolev space and that we are assuming that $s > n/2$, from which it follows that the inclusion

$$J_s : H^s(X) \hookrightarrow \mathcal{C}(X)$$

is a compact embedding. Let B_R be the closed ball of radius R centered at the origin in $H^s(X)$ and $\mathcal{H} = \overline{J_s(B_R)}$ be its image in $\mathcal{C}(X)$.

As above, using Proposition 6, given $\varepsilon, \delta > 0$, to have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m} \{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \delta$$

it is sufficient that the number m of examples satisfies

$$m \ge \frac{288M^2}{\varepsilon}\left[\left(\frac{24CRM}{\varepsilon}\right)^{n/s} + \ln\left(\frac{1}{\delta}\right)\right]. \tag{5}$$

Here C is the constant of (4).

Example 4 (continued). Recall that $I_K : \mathcal{H}_K \to \mathcal{C}(X)$ is a compact embedding defined by a $\mathcal{C}^\infty$ Mercer kernel $K : X \times X \to \mathbb{R}$, B_R is the ball of radius R in $\mathcal{H}_K$ and $\mathcal{H} = \overline{I_K(B_R)}$. As above, given $\varepsilon, \delta > 0$, to have

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m} \{\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) \le \varepsilon\} \ge 1 - \delta$$

it is enough that the number m of examples satisfies

$$m \ge \frac{288M^2}{\varepsilon}\left[\left(\frac{24MRC_h}{\varepsilon}\right)^{\frac{2n}{h}} + \ln\left(\frac{1}{\delta}\right)\right].$$

Here $h > n$ and C_h are as in Section 6.

Remark 12. Note that in the bounds in Examples 3 and 4 there is no dependency on the dimension of $\mathcal{H}$ (which is now infinite), in contrast with the bound shown in Example 2. These results may be said to be "dimension-free". The parameter R in Examples 3 and 4 determines the size of the hypothesis space and is our replacement for the VC dimension (which is infinite in these examples).

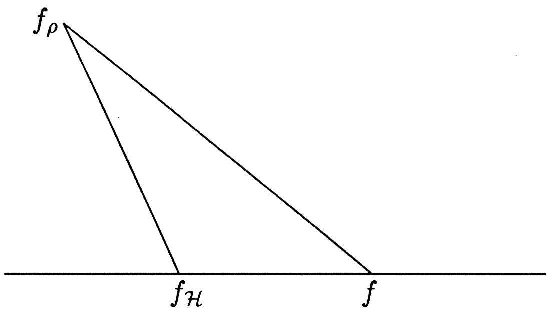
Toward the proof of Theorem C* we show an additional property of convex hypothesis spaces.

From the discussion in Section 3 it follows that $f_{\mathcal{H}}$ is a function in $\mathcal{H}$ whose distance in $\mathcal{L}^2_\rho(X)$ to f_ρ is minimal. We next prove that, if $\mathcal{H}$ is convex, it is unique.

Lemma 5. *Let $\mathcal{H}$ be a convex subset of $\mathcal{C}(X)$ such that $f_{\mathcal{H}}$ exists. Then $f_{\mathcal{H}}$ is unique as an element in $\mathcal{L}^2_\rho(X)$ and, for all $f \in \mathcal{H}$,*

$$\int_X (f_{\mathcal{H}} - f)^2 \le \mathcal{E}_{\mathcal{H}}(f).$$

Proof. Let $s = \overline{f_{\mathcal{H}} f}$ be the line segment with extremities $f_{\mathcal{H}}$ and f.



Since $\mathcal{H}$ is convex, $s \subset \mathcal{H}$. And, since $f_{\mathcal{H}}$ minimizes the distance in $\mathcal{L}^2_\rho(X)$ to f_ρ over $\mathcal{H}$, we have that, for all $g \in s$, $\|f_{\mathcal{H}} - f_\rho)\|_\rho \le \|g - f_\rho\|_\rho$. This implies that the angle $\widehat{f_\rho f_{\mathcal{H}} f}$ is obtuse, and that implies (note that the squares are crucial)

$$\|f_{\mathcal{H}} - f\|^2_\rho \le \|f - f_\rho\|^2_\rho - \|f_{\mathcal{H}} - f_\rho\|^2_\rho,$$

i.e.

$$\int_X (f_{\mathcal{H}} - f)^2 \le \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}).$$

This proves the desired inequality. The uniqueness of $f_{\mathcal{H}}$ follows by considering the line segment joining two minimizers $f'_{\mathcal{H}}$ and $f''_{\mathcal{H}}$. Reasoning as above, one shows that both angles $\widehat{f_\rho f'_{\mathcal{H}} f''_{\mathcal{H}}}$ and $\widehat{f_\rho f''_{\mathcal{H}} f'_{\mathcal{H}}}$ are obtuse. This is only possible if $f''_{\mathcal{H}} = f'_{\mathcal{H}}$. □

Corollary 1. *With the hypotheses of Theorem C*, for all $\varepsilon > 0$,*

$$\operatorname*{Prob}_{\mathbf{z} \in Z^m} \left\{ \int (f_{\mathbf{z}} - f_{\mathcal{H}})^2 \le \varepsilon \right\} \ge 1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{24M}\right) e^{-\frac{m\varepsilon}{288M^2}}.$$

□

Now, in addition to convexity, assume that $\mathcal{H}$ is a compact subset of $\mathcal{C}(X)$ so that the covering numbers $\mathcal{N}(\mathcal{H}, \eta)$ make sense and are finite. Also, assume that there exists $M > 0$ such that, for all $f \in \mathcal{H}$, $|f(x) - y| \le M$ a.e. The following analogue of Theorem B is the main steppingstone towards the proof of Theorem C*.[3]

For a sample $\mathbf{z} \in Z^m$, the *empirical error in* $\mathcal{H}$ of $f \in \mathcal{H}$ is $\mathcal{E}_{\mathcal{H},\mathbf{z}}(f) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})$. Note that $\mathcal{E}_{\mathcal{H},\mathbf{z}}(f_{\mathbf{z}}) \le 0$.

[3]The writing of the rest of this section benefitted greatly from discussions with Partha Niyogi and a remark by Peter Bartlett.

Proposition 7. *For all* $\varepsilon > 0$ *and* $0 < \alpha < 1$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in\mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f)+\varepsilon} \geq 3\alpha\right\} \leq \mathcal{N}\left(\mathcal{H}, \frac{\alpha\varepsilon}{4M}\right) e^{-\frac{\alpha^2 m\varepsilon}{8M^2}}.$$

□

Before proving Proposition 7 we show how Theorem C* follows from it.

Proof of Theorem C.* Put $\alpha = 1/6$ in Proposition 7. By this proposition, with probability at least

$$1 - \mathcal{N}\left(\mathcal{H}, \frac{\varepsilon}{24M}\right) e^{-\frac{m\varepsilon}{288M^2}}$$

we have

$$\sup_{f\in\mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f)+\varepsilon} < \frac{1}{2},$$

and therefore, for all $f \in \mathcal{H}$, $\frac{1}{2}\mathcal{E}_{\mathcal{H}}(f) < \mathcal{E}_{\mathcal{H},\mathbf{z}}(f) + \frac{1}{2}\varepsilon$. Take $f = f_{\mathbf{z}}$. Then, multiplying by 2,

$$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) < 2\mathcal{E}_{\mathcal{H},\mathbf{z}}(f_{\mathbf{z}}) + \varepsilon,$$

but $\mathcal{E}_{\mathcal{H},\mathbf{z}}(f_{\mathbf{z}}) \leq 0$ by definition of $f_{\mathbf{z}}$ from which $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) < \varepsilon$ and the theorem follows. □

We now proceed with the proof of Proposition 7. Let $\ell(f) : Z \to Y$ be defined by $f_Y^2 - f_{\mathcal{H},Y}^2$. Thus, $\mathbf{E}\ell(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{E}_{\mathcal{H}}(f)$ and, for $\mathbf{z} \in Z^m$, $\mathbf{E}_{\mathbf{z}}\ell(f) = \mathcal{E}_{\mathbf{z}}(f) - \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}}) = \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)$. In addition, we note that for all $f \in \mathcal{H}$, $|\ell(f)(x,y)| \leq M^2$ a.e.

Convexity plays a major role in the following result. Let $\sigma^2 = \sigma^2(\ell(f))$ denote the variance of $\ell(f)$.

Lemma 6. *For all* $f \in \mathcal{H}$, $\sigma^2 \leq 4M^2\mathcal{E}_{\mathcal{H}}(f)$.

Proof. Because

$$\sigma^2 \leq \mathbf{E}\ell(f)^2 = \mathbf{E}[(f_{\mathcal{H}} - f)^2(y - f + y - f_{\mathcal{H}})^2] \leq 4M^2\mathbf{E}[(f_{\mathcal{H}} - f)^2],$$

it is enough to prove that $\mathbf{E}[(f_{\mathcal{H}} - f)^2] \leq \mathcal{E}_{\mathcal{H}}(f)$. This is exactly Lemma 5. □

Our next result is a form of Theorem A for the random variable $\ell(f)$.

Lemma 7. *Let* $f \in \mathcal{H}$. *For all* $\varepsilon, \alpha > 0$, $\alpha \leq 1$,

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f)+\varepsilon} \geq \alpha\right\} \leq e^{-\frac{\alpha^2 m\varepsilon}{8M^2}}.$$

Proof. Let $\mu = \mathcal{E}_{\mathcal{H}}(f)$. Using the one-sided Bernstein's inequality (see Remark 3) applied to $\ell(f)$ and the fact that $|\ell(f)(z)| \leq M^2$ a.e., we get

$$\operatorname*{Prob}_{\mathbf{z}\in Z^m}\left\{\frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mu+\varepsilon} \geq \alpha\right\} \leq e^{-\frac{(\alpha(\mu+\varepsilon))^2 m}{2\left(\sigma^2 + \frac{1}{3}M^2\alpha(\mu+\varepsilon)\right)}}.$$

We only need to show that

$$\begin{aligned}
& \frac{\varepsilon}{8M^2} \leq \frac{(\mu+\varepsilon)^2}{2\left(\sigma^2 + \frac{1}{3}M^2\alpha(\mu+\varepsilon)\right)} \\
\iff \quad & \frac{\varepsilon}{4M^2}\left(\sigma^2 + \frac{1}{3}M^2\alpha(\mu+\varepsilon)\right) \leq (\mu+\varepsilon)^2 \\
\iff \quad & \frac{\varepsilon\sigma^2}{4M^2} + \frac{\varepsilon\alpha\mu}{12} + \frac{\varepsilon^2\alpha}{12} \leq (\mu+\varepsilon)^2.
\end{aligned}$$

The second and third terms on the left are respectively bounded by $\mu\varepsilon$ and ε^2 since $\alpha \leq 1$. The first one is smaller than $\varepsilon\mu$ since, by Lemma 6, σ^2 is bounded by $4M^2\mu$. The result follows since $2\mu\varepsilon + \varepsilon^2 \leq (\mu + \varepsilon)^2$. □

Lemma 8. *Let $0 < \alpha < 1$, $\varepsilon > 0$, and $f \in \mathcal{H}$ such that*

$$\frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(f) + \varepsilon} < \alpha.$$

For all $g \in \mathcal{H}$ such that $\|f - g\|_\infty \leq \frac{\alpha\varepsilon}{4M}$ we have

$$\frac{\mathcal{E}_{\mathcal{H}}(g) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} < 3\alpha.$$

Proof.

$$\begin{aligned}
\frac{\mathcal{E}_{\mathcal{H}}(g) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(g)}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} &= \frac{\mathcal{E}(g) - \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}_{\mathbf{z}}(g) + \mathcal{E}_{\mathbf{z}}(f_{\mathcal{H}})}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} = \frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f_{\mathcal{H}})}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} \\
&= \frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f) + L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_{\mathcal{H}})}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} \\
&= \frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} + \frac{L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_{\mathcal{H}})}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon}.
\end{aligned}$$

If the first term above is negative, then it is certainly smaller than α. Otherwise we have

$$\frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} \leq \frac{L_{\mathbf{z}}(g) - L_{\mathbf{z}}(f)}{\varepsilon} \leq \frac{4M\alpha\varepsilon}{4M\varepsilon} = \alpha,$$

where the last inequality follows from using $\|f - g\|_\infty \leq \frac{\alpha\varepsilon}{4M}$ in Proposition 3. For the second term, note that, using the first part in the proof of Proposition 3,

$$\mathcal{E}(f) - \mathcal{E}(g) \leq 2M\|f - g\|_\infty \leq 2M\frac{\alpha\varepsilon}{4M} < \varepsilon$$

since $\alpha < 1$. This implies that

$$\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H}}(g) = \mathcal{E}(f) - \mathcal{E}(g) \leq \varepsilon \leq \mathcal{E}_{\mathcal{H}}(g) + \varepsilon$$

or, equivalently, that $\frac{\mathcal{E}_{\mathcal{H}}(f)+\varepsilon}{\mathcal{E}_{\mathcal{H}}(g)+\varepsilon} \leq 2$. But then

$$\frac{L_{\mathbf{z}}(f) - L_{\mathbf{z}}(f_{\mathcal{H}})}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} = \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},\mathbf{z}}(f)}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} \leq \alpha\frac{\mathcal{E}_{\mathcal{H}}(f) + \varepsilon}{\mathcal{E}_{\mathcal{H}}(g) + \varepsilon} \leq 2\alpha.$$

□

Proposition 7 follows from Lemma 8 by applying the same argument used to prove Theorem B from Proposition 3.

Remark 13. Note that, to obtain Theorem C*, we only used convexity to prove Lemma 5. But the inequality proved in this lemma may hold true in other situations as well. A case which stands out is when $f_\rho \in \mathcal{H}$. In this case $f_{\mathcal{H}} = f_\rho$ and the inequality in Lemma 5 is trivial.

8. Final remarks

Remark 14. In this chapter we have assumed that $Y = \mathbb{R}$. They can, however, be extended to Y, a finite dimensional inner product space.

Remark 15. The least squares error function $\mathcal{E}(f)$ above is only one of the many used in the learning theory literature. Our view is that it is the central notion because of mathematical tradition and algorithmic simplicity. However, the least squares error has its limitations and problems. It would be interesting to analyze some other error functions in the framework of our paper. See e.g. [11].

Remark 16. Let us compare what we have done with the more traditional approach in learning theory, especially inspired by Vapnik, with the use of VC (Vapnik-Chervonenkis) dimension and its variants (see e.g. [18], [42]). As we have remarked, the hypothesis space $\mathcal{H}$ plays a central role in the learning process. The earlier choice of hypothesis space is a space of functions on X which carries no topology. The development proceeds with a more combinatorial flavor to achieve results which cannot be compared directly with our Theorems B, C, and C*. In that setting, covering numbers usually depend on the sample, and the sample error estimate will depend on the VC dimension.

Our approach, with its function space $\mathcal{H} \subset \mathcal{C}(X)$, leads quickly to classical functional analysis. The VC dimension is replaced by the radius R of a ball which defines the hypothesis space in a Sobolev space or in a reproducing kernel Hilbert space.

Moreover we emphasize the continuous (regression) perspective and are led to the approximation questions of the next chapter.

Chapter II: Approximation Error

For a given hypothesis space $\mathcal{H}$, the error $\mathcal{E}(f_{\mathbf{z}})$ of the empirical target $f_{\mathbf{z}}$ decomposes as

$$\mathcal{E}(f_{\mathbf{z}}) = \mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) + \mathcal{E}(f_{\mathcal{H}}).$$

The first term in this sum, the sample error, has been the focus of Chapter I. The second term, the approximation error, will be the focus of this chapter. The approximation error depends only on $\mathcal{H}$ and ρ and, by Proposition 1, is equal to $\int_X (f_{\mathcal{H}} - f_\rho)^2 + \sigma_\rho^2$. Note that σ_ρ^2 does not depend on the choice of $\mathcal{H}$. Therefore, when studying the approximation error we will examine the integral $\int_X (f_{\mathcal{H}} - f_\rho)^2$. Since f_ρ is not known and we have made no assumptions on it besides being bounded, there are limits on how much one can say about the approximation error. We note that if $f_\rho \in \mathcal{H}$, then $f_{\mathcal{H}} = f_\rho$ and the integral above is zero. This chapter is devoted to estimates of the integral for various $\mathcal{H}$ and to the implications for the bias-variance problem.

1. Fourier series and the approximation error

In this section we give an example of a finite dimensional hypothesis space (Example 5 below) and an estimate for the corresponding approximation error. To get this estimate, we will need to estimate the growth of the eigenvalues of a given operator. Growth of eigenvalues, or the highly related growth of entropy numbers, is a recurring theme in our report.

On one hand, Fourier series give a link from our problem in learning theory to the mathematical analysis known to many scientists. On the other hand, the interested

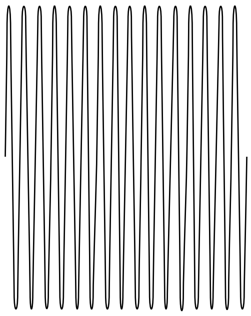


FIGURE 1. Shape of ϕ_α for α large, $n = 1$.

reader will be able to discover the relations (via Greens' functions) to our integral operators and to entropy numbers (see Appendix A of Chapter III) as well as our use of Sobolev spaces, which were originally developed to better understand elliptic operators.

Let S^1 be the circle, say, described by a real number t mod 2π, and $X = (S^1)^n$ the n-dimensional torus. For each $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{Z}^n$ consider the complex valued function on X, ϕ_α, given by $\phi_\alpha(x) = (2\pi)^{-n/2}e^{i(\alpha \cdot x)}$. Here $i = \sqrt{-1}$. By taking the real part from de Moivre's formula one can obtain a real valued function on X. Thus we may deal with complex valued functions on X.

Let $\mathcal{L}^2_\mu(X)$ be the space of square integrable functions on X with respect to the Lebesgue measure induced on X as a quotient of $\mathbb{R}^n$. Recall that a sequence $\{\phi_k\}$ in a Hilbert space H is said to be a *complete orthonormal system* (or a *Hilbert basis*) if the following conditions hold:

1. for all $k, q \geq 1$, $\langle \phi_k, \phi_q \rangle = 0$;
2. for all $k \geq 1$, $\|\phi_k\| = 1$; and
3. for all $f \in H$, $f = \sum_{k=1}^{\infty} \langle f, \phi_k \rangle \phi_k$.

The set $\{\phi_\alpha\}_{\alpha \in \mathbb{Z}^n}$ forms a Hilbert basis of $\mathcal{L}^2_\mu(X)$ with respect to the inner product $\langle f, g \rangle = \int f\overline{g}$, $\overline{g}$ the complex conjugate of g. Thus, every function $f \in \mathcal{L}^2_\mu(X)$ can be written as

$$f = \sum_{\alpha \in \mathbb{Z}^n} c_\alpha \phi_\alpha.$$

But if $\|\alpha\|$ is large, the function ϕ_a oscillates with high frequency, and thus each of these terms gives a fine structure, beyond sensitivity of measurement devices. See Figure 1.

This heuristic indicates how, for purposes of the hypothesis space of Section 3 in Chapter I, it makes sense to consider the subspace $\mathcal{H}_N \subset \mathcal{L}^2_\mu(X)$ spanned by the set $\{\phi_\alpha\}_{\|\alpha\|^2 \leq B}$ for some B with the induced structure of Hilbert space. The dimension $N = N(B)$ of this space is the number of integer lattice points in the ball of radius B of $\mathbb{R}^n$. Thus, a crude bound is $N(B) \leq (2B)^{n/2}$. The ball $\mathcal{H}_{N,R}$ of radius R with respect to the norm $\| \|_\infty$ in $\mathcal{H}_N$ is a candidate for the $\mathcal{H}$ of Chapter I.

Remark 1. Let $\Delta : \mathcal{C}^\infty(X) \to \mathcal{C}^\infty(X)$ be the Laplace operator,

$$\Delta(f) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

It is immediate to check that, for all $\alpha \in \mathbb{Z}^n$, $\Delta(\phi_\alpha) = -\|\alpha\|^2 \phi_\alpha$. Therefore, ϕ_α is an eigenvector of $-\Delta$ with eigenvalue $\|\alpha\|^2$.

Since the n-dimensional torus is not a very suitable space for most examples of learning theory, we extend the setting as suggested by Remark 1.

Example 5. Consider now a bounded domain X in $\mathbb{R}^n$ with smooth boundary ∂X, and a Hilbert basis $\{\phi_k\}_{k\geq 1}$ of $\mathcal{C}^\infty$ functions in $\mathcal{L}^2_\mu(X)$ satisfying

$$\begin{cases} -\Delta\phi_k = \zeta_k \phi_k & \text{in } X, \text{ for all } k \geq 1 \\ \phi_k = 0 & \text{on } \partial X, \text{ for all } k \geq 1 \end{cases}$$

with

$$0 < \zeta_1 \leq \zeta_2 \leq \zeta_3 \ldots \uparrow \infty.$$

Here μ is the Lebesgue measure on X inherited from $\mathbb{R}^n$. The existence of $\{\phi_k\}_{k\geq 1}, \{\zeta_k\}_{k\geq 1}$ as above uses a main theorem in the theory of elliptic differential equations.

For $N \in \mathbb{N}$ consider $\mathcal{H}_N$, the subspace of $\mathcal{L}^2_\mu(X)$ generated by $\{\phi_1, \ldots, \phi_N\}$. The higher frequency justification for the cutoff in the case of Fourier series still applies. This comes from the Courant Nodal Theorem or the many variables Morse Index Theorem (see [36] for a formal account). Also, as above, let $\mathcal{H} = \mathcal{H}_{N,R}$ be the ball of radius R with respect to the norm $\| \; \|_\infty$ in $\mathcal{H}_N$ and let $f_{\mathcal{H}}$ be the corresponding target function.

Recall we have assumed that f_ρ is bounded on X. Then, $f_\rho \in \mathcal{L}^2_\rho(X)$ and $f_\rho \in \mathcal{L}^2_\mu(X)$. Suppose in addition that $R \geq \|f_\rho\|_\infty$. Then $R \geq \|f_\rho\|_\rho$ and $f_{\mathcal{H}}$ is the orthogonal projection of f_ρ on $\mathcal{H}_N$ w.r.t. the inner product in $\mathcal{L}^2_\rho(X)$. The main result of this section bounds the approximation error $\mathcal{E}(f_{\mathcal{H}})$.

Let $\mathcal{D}_{\mu\rho}$ denote the operator norm $\|J\|$ where J is the identity function

$$\mathcal{L}^2_\mu(X) \xrightarrow{J} \mathcal{L}^2_\rho(X).$$

We will call $\mathcal{D}_{\mu\rho}$ the *distortion* of ρ (with respect to μ). It measures how much ρ distorts the ambient measure μ. It is often reasonable to suppose that the distortion $\mathcal{D}_{\mu\rho}$ is finite.

Since ρ is not known, then $\mathcal{D}_{\mu\rho}$ is not known in general as well. But our estimate in Theorem 1 below gives a relation between the approximation error and $\mathcal{D}_{\mu\rho}$. Moreover, the context could lead to some information about $\mathcal{D}_{\mu\rho}$. An important case is the one in which, in spite of ρ not being known, we do know ρ_X. In this case $\mathcal{D}_{\mu\rho}$ may be derived.

For $f = \sum_{k=1}^\infty c_k \phi_k$, let $\|f\|_K$ denote

$$\left(\sum_{k=1}^{\infty} c_k^2 \zeta_k \right)^{1/2}.$$

The set of f such that this series is convergent is a linear subspace of $\mathcal{L}^2_\mu(X)$ on which $\| \; \|_K$ is a norm. Motivation for this norm is given in the next section, in which a similar construction is described for an integral operator given by a Mercer kernel K (hence the notation).

Theorem 1. *Let $\mathcal{H}$ and $f_{\mathcal{H}}$ be as above. The approximation error satisfies*

$$\mathcal{E}(f_{\mathcal{H}}) \le \mathcal{D}_{\mu\rho}^2 \left(\frac{\mathrm{Vol}(X)}{N+1}\right)^{2/n} \|f_\rho\|_K^2 + \sigma_\rho^2.$$

Towards the proof of Theorem 1 first note that

$$\|f_\rho - f_{\mathcal{H}}\|_\rho = d_\rho(f_\rho, \mathcal{H}_N) \le \|J\| d_\mu(f_\rho, \mathcal{H}_N).$$

Recall that ℓ^2 is the linear space of all square summable sequences $(a_k)_{k\ge1}$. It is a Hilbert space with the inner product

$$\langle (a_k), (b_k)\rangle = \sum_{k\ge1} a_k b_k.$$

Since $f_\rho \in \mathcal{L}_\mu^2(X)$, there exists a sequence $\{a_k\}_{k\ge1} \in \ell^2$ such that $f_\rho = \sum a_k\phi_k$. Then

$$\begin{aligned} d_\mu(f_\rho, \mathcal{H}_N)^2 &= \left\| \sum_{k=N+1}^{\infty} a_k\phi_k \right\|_\mu^2 \\ &= \sum_{k=N+1}^{\infty} a_k^2 = \sum_{k=N+1}^{\infty} a_k^2 \zeta_k \frac{1}{\zeta_k} \\ &\le \frac{1}{\zeta_{N+1}} \|f_\rho\|_K^2. \end{aligned}$$

The next lemma deals with the growth of the eigenvalues ζ_k.

Lemma 1. *For $k \ge 1$, $\zeta_k \ge \left(\frac{k}{\mathrm{Vol}(X)}\right)^{2/n}$.*

Proof. Under the hypothesis described at the beginning of this example, a version of a result of H. Weyl by Li and Yau [23] (pointed out to us by Roderick Wong) states that, for all $k \ge 1$,

$$\zeta_k \ge \frac{n}{n+2} 4\pi^2 \left(\frac{k}{B_n \,\mathrm{Vol}(X)}\right)^{2/n} \tag{6}$$

where B_n is the volume of the unit ball in $\mathbb{R}^n$ and $\mathrm{Vol}(X)$ the volume of X.

Stirling's inequality, $\sqrt{2\pi} u^{u-\frac12} e^{-u} \le \Gamma(u)$ (see [1], Chapter 5, Section 2.5, Exercise 2), implies that

$$\Gamma(n/2) \ge \sqrt{2\pi}\left(\frac{n}{2}\right)^{\frac{n-1}{2}} e^{-\frac{n}{2}}$$

and, consequently, since $B_n = \frac{1}{n}\frac{(2\pi)^{n/2}}{\Gamma(n/2)}$, that

$$B_n \le \frac{1}{n} \frac{(2\pi)^{\frac{n}{2}} e^{\frac{n}{2}}}{\sqrt{2\pi}\left(\frac{n}{2}\right)^{\frac{n-1}{2}}} = \frac{(4\pi)^{\frac{n-1}{2}} e^{\frac{n}{2}}}{n^{\frac{n+1}{2}}}.$$

Placing this bound in inequality (6) we obtain, for all $k \in \mathbb{N}$,

$$\begin{aligned}
\zeta_k &\geq \frac{n}{n+2} 4\pi^2 \left(\frac{k n^{\frac{n+1}{2}}}{\mathrm{Vol}(X)(4\pi)^{\frac{n-1}{2}} e^{\frac{n}{2}}} \right)^{2/n} \\
&= \frac{n}{n+2} \pi \left(\frac{k}{\mathrm{Vol}(X)} \right)^{2/n} \frac{(4\pi)^{\frac{1}{n}} n^{\frac{n+1}{n}}}{e} \\
&= \frac{n^{2+\frac{1}{n}}}{n+2} \frac{\pi}{e} \left(\frac{k}{\mathrm{Vol}(X)} \right)^{2/n} (4\pi)^{\frac{1}{n}} \geq \left(\frac{k}{\mathrm{Vol}(X)} \right)^{2/n}
\end{aligned}$$

since $\frac{n^{2+\frac{1}{n}}}{n+2} \frac{\pi}{e} (4\pi)^{\frac{1}{n}} \geq 1$ for all $n \in \mathbb{N}$. □

Proof of Theorem 1. Using Lemma 1 we obtain

$$\begin{aligned}
d_\rho(f_\rho, \mathcal{H}_N)^2 &\leq \mathcal{D}_{\mu\rho}^2 d_\mu(f_\rho, \mathcal{H}_N)^2 \\
&\leq \mathcal{D}_{\mu\rho}^2 \frac{1}{\zeta_{N+1}} \|f_\rho\|_K^2 \\
&\leq \mathcal{D}_{\mu\rho}^2 \left(\frac{\mathrm{Vol}(X)}{N+1} \right)^{2/n} \|f_\rho\|_K^2.
\end{aligned}$$

□

We already remarked that our goal is to minimize $\mathcal{E}(f_{\mathbf{z}})$ which equals the sum

$$\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}}) + \int (f_{\mathcal{H}} - f_\rho)^2 + \sigma_\rho^2.$$

A form of the bias-variance problem is to minimize this sum over $N \in \mathbb{N}$ assuming R, m and δ are fixed. So, fix m, $R = \|f_\rho\|_\infty$, and $\delta > 0$. From Section 7 in Chapter I it follows that, if

$$m \geq \frac{288M^2}{\varepsilon} \left[N \ln \left(\frac{96RM}{\varepsilon} \right) + 1 + \ln \left(\frac{1}{\delta} \right) \right],$$

then, with probability at least $1-\delta$, the sample error is bounded by ε. From this equation it follows that, for given m, δ, R, M and N, with probability at least $1-\delta$, the sample error is bounded by any quantity ε satisfying

$$\varepsilon - \frac{288M^2}{m} \left(N \ln \left(\frac{96RM}{\varepsilon} \right) + 1 + \ln \left(\frac{1}{\delta} \right) \right) \geq 0.$$

The equation obtained by taking the equality in this inequality has exactly one positive solution. This is due to the form $f(t) = 0$ for this equation with $f(t) = t + c\ln(t) - d$, $c, d > 0$. Thus $f(0) = -\infty$, $f(+\infty) = +\infty$, and $f'(t) = 1 + \frac{c}{t}$, which is always positive, showing that f monotonically increases in $(0, +\infty)$. We denote this solution by $\varepsilon(N)$, thus emphasizing the functional dependency of $\varepsilon(N)$ with respect to N.

From Theorem 1, we know that the approximation error is bounded by

$$\mathcal{A}(N) = \mathcal{D}_{\mu\rho}^2 \left(\frac{\mathrm{Vol}(X)}{N+1} \right)^{2/n} \|f_\rho\|_K^2 + \sigma_\rho^2.$$

The integer N minimizing $\mathcal{A}(N) + \varepsilon(N)$ will thus be a solution of the bias-variance problem above. While we have no explicit form for the solution of this minimization problem, it is easy to numerically deal with it. One may also derive some qualitative information about N.

This development is valid for the case of any compact submanifold X of Euclidean space. A general reference for the material in this section is [33].

2. Abstract approximation error

A linear operator $L : H \to H$ on a Hilbert space H is said to be *self-adjoint* if, for all $f, g \in H$, $\langle Lf, g\rangle = \langle f, Lg\rangle$. It is said to be *positive* (resp. *strictly positive*) if it is self-adjoint and, for all non-trivial $f \in H$, $\langle Lf, f\rangle \geq 0$ (resp. $\langle Lf, f\rangle > 0$).

The next result, the Spectral Theorem for compact operators (see Section 4.10 of [12] for a proof), will be useful in this and the next chapter.

Theorem 2. *Let L be a compact linear operator on an infinite dimensional Hilbert space H. Then there exists in H a complete orthonormal system $\{\phi_1, \phi_2, \dots\}$ consisting of the eigenvectors of L. If λ_k is the eigenvalue corresponding to ϕ_k, then the set $\{\lambda_k\}$ is either finite or $\lambda_k \to 0$ when $k \to \infty$. In addition, $\max_{k\geq 1} |\lambda_k| = \|L\|$. The eigenvalues are real if L is self-adjoint. If, in addition, L is positive, then $\lambda_k \geq 0$ for all $k \geq 1$, and if L is strictly positive, then $\lambda_k > 0$ for all $k \geq 1$.* □

If L is a strictly positive operator, then L^τ is defined, for all $\tau \geq 0$, by

$$L^\tau\left(\sum a_k\phi_k\right) = \sum \lambda_k^\tau a_k\phi_k.$$

If $\tau < 0$, L^τ is defined by the same formula on the subspace

$$S_\tau = \left\{\sum a_k\phi_k \mid \sum (a_k\lambda_k^\tau)^2 \text{ is convergent}\right\}.$$

For $\tau < 0$, the expression $\|L^\tau a\|$ must be understood as ∞ if $a \notin S_\tau$.

Theorems 3 and 5 in this and the next section are taken from [38], where one can find a more substantial development of the approximation error.

Theorem 3. *Let H be a Hilbert space and A a self-adjoint, strictly positive compact operator on H. Let $s, r \in \mathbb{R}$ such that $s > r > 0$.*

(1) *Let $\gamma > 0$. Then, for all $a \in H$*

$$\min_{b\in H}\left(\|b-a\|^2 + \gamma\|A^{-s}b\|^2\right) \leq \gamma^r\|A^{-sr}a\|^2.$$

(2) *Let $R > 0$. Then, for all $a \in H$*

$$\min_{b \text{ s.t. } \|A^{-s}b\|\leq R} \|b-a\| \leq \left(\frac{1}{R}\right)^{\frac{r}{s-r}} \|A^{-r}a\|^{\frac{s}{s-r}}.$$

In both cases the minimizer $\hat{b}$ exists and is unique. In addition, in part (1), $\hat{b} = (\mathrm{Id} + \gamma A^{-2s})^{-1}a$.

Proof. First note that by replacing A by A^s we can reduce the problem in both parts (1) and (2) to the case $s = 1$.

Now, for part (1), consider

$$\varphi(b) = \|b-a\|^2 + \gamma\|A^{-1}b\|^2.$$

If a point $\hat{b}$ minimizes φ, then it must be a zero of the derivative $D\varphi$. That is, $\hat{b}$ satisfies $(\mathrm{Id} + \gamma A^{-2})\hat{b} = a$, which implies $\hat{b} = (\mathrm{Id} + \gamma A^{-2})^{-1}a$. Note that the operator $\mathrm{Id} + \gamma A^{-2}$ is invertible since it is the sum of the identity and a positive operator.

If $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ denotes the eigenvalues of A,

$$\begin{aligned}
\varphi(\hat{b}) &= \|((\mathrm{Id} + \gamma A^{-2})^{-1} - \mathrm{Id})a\|^2 + \gamma\|A^{-1}(\mathrm{Id} + \gamma A^{-2})^{-1}a\|^2 \\
&= \sum_{k=1}^{\infty}\left\{\left(\frac{1}{1+\gamma\lambda_k^{-2}} - 1\right)^2 + \gamma\sum_{k=1}^{\infty}\left(\frac{1}{\lambda_k(1+\gamma\lambda_k^{-2})}\right)^2\right\} a_k^2 \\
&= \sum_{k=1}^{\infty}\left(\frac{\gamma^2\lambda_k^{-4} + \gamma\lambda_k^{-2}}{(\lambda_k^{-2}(\lambda_k^2+\gamma))^2}\right) a_k^2 = \gamma\sum_{k=1}^{\infty}\left(\frac{1}{\lambda_k^2+\gamma}\right) a_k^2 \\
&= \gamma\sum_{k=1}^{\infty}\left(\frac{\lambda_k^{2r}}{\lambda_k^2+\gamma}\right)\lambda_k^{-2r}a_k^2 \leq \gamma\left(\sup_{t\in\mathbb{R}^+}\frac{t^r}{t+\gamma}\right)\|A^{-r}a\|^2.
\end{aligned}$$

Let $\psi(t) = \frac{t^r}{t+\gamma}$. Then

$$\psi'(t) = \frac{rt^{r-1}}{t+\gamma} - \frac{t^r}{(t+\gamma)^2} = 0 \qquad \text{iff} \qquad t = \hat{t} = \sqrt{\frac{\gamma r}{1-r}}.$$

Thus

$$\psi(\hat{t}) = \gamma^{r-1}r^r(1-r)^{1-r} \leq \gamma^{r-1}.$$

We conclude that

$$\varphi(\hat{b}) = \min_{b\in H}\|b-a\|^2 + \gamma\|A^{-1}b\|^2 \leq \gamma^r\|A^{-r}a\|^2$$

and hence (1).

For part (2) first note that if $\|A^{-1}a\| \leq R$, then the minimum in the statement is zero and the theorem is obviously true. Assume from now on that this is not the case. Then we notice that the point $\hat{b}$ minimizing $\|a-b\|$ in the subset of H given by $\|A^{-1}b\| \leq R$ is in the boundary of this subset, i.e. $\|A^{-1}\hat{b}\| = R$.

Now, a well known result in constrained optimization states that there exists $\gamma \geq 0$ (the Lagrange multiplier) such that the point $\hat{b}$ is a zero of the Lagrangian

$$D(\|b-a\|^2) + \gamma D(\|A^{-1}b\|^2).$$

But this Lagrangian coincides with $D\varphi$ of part (1), and we proved in this part that $\varphi(\hat{b}) \leq \gamma^r\|A^{-r}a\|^2$. From this inequality we deduce firstly that

$$\gamma R^2 \leq \gamma^r\|A^{-r}a\|^2$$

and secondly, since $\gamma \geq 0$, that

$$\|\hat{b}-a\|^2 \leq \gamma^r\|A^{-r}a\|^2.$$

From the first of these two inequalities it follows that

$$\gamma \leq \left(\frac{1}{R}\right)^{\frac{2}{1-r}}\|A^{-r}a\|^{\frac{2}{1-r}}.$$

Replacing this bound for γ in the second inequality, one gets the statement in part (2). ☐

Remark 2. In Example 3, $A = (-\Delta + \mathrm{Id})^{-1/2}$ and $s = \tau$ as in the proof of Theorem 5 below. In Example 4, $A = L_K^{1/2}$, $s = 1$.

Remark 3. The quantity

$$K(a,\gamma) = \min_{b\in H} \left(\|b-a\|^2 + \gamma\|A^{-s}b\|^2\right)$$

is a modification of the K-functional of interpolation theory [4]. Moreover,

$$I(a,R) = \min_{b \text{ s.t. } \|A^{-s}b\|\le R} \|b-a\|$$

is an object of study also in [4]). The proof of Theorem 3 shows that $K(a,\gamma) = \gamma\|(A^{2s}+\gamma\,\mathrm{Id})^{-1/2}a\|^2$, and, for $\gamma > 0$ and $R = R(\gamma) = \|A^{-s}\hat{b}_\gamma\|$, $I(a,R) = K(a,\gamma) - \gamma R^2$.

Remark 4. We now introduce a *general setting in Hilbert space.* Let ν be a Borel measure on X and $A : \mathcal{L}^2_\nu(X) \to \mathcal{L}^2_\nu(X)$ a compact strictly positive operator. Fix $s > 0$ and define $\mathbb{E} = \{g \in \mathcal{L}^2_\nu(X) \mid \|A^{-s}g\|_\nu < \infty\}$. We can make $\mathbb{E}$ a Hilbert space with the inner product

$$\langle g,h\rangle_{\mathbb{E}} = \langle A^{-s}g, A^{-s}h\rangle_\nu .$$

Thus, $A^{-s} : \mathcal{L}^2_\nu(X) \to \mathbb{E}$ is a Hilbert isomorphism. The general setting in Hilbert space is the setting above together with the assumption that the inclusion $\mathbb{E} \hookrightarrow \mathcal{L}^2_\nu(X)$ factors

$$\begin{tikzcd} \mathbb{E} \arrow[r] \arrow[dr, "J_{\mathbb{E}}"'] & \mathcal{L}^2_\nu(X) \\ & \mathcal{C}(X) \arrow[u] \end{tikzcd}$$

with $J_{\mathbb{E}}$ compact. Therefore the hypothesis space $\mathcal{H} = \mathcal{H}_{\mathbb{E},R}$ is $\overline{J_{\mathbb{E}}(B_R)}$ where B_R is the ball of radius R in $\mathbb{E}$. Note that the target $f_{\mathcal{H}}$ is the $\hat{b}$ of Theorem 3 (2), for $H = \mathcal{L}^2_\nu(X)$, and we may consider the corresponding approximation error.

As in Section 1 we consider $\mathcal{D}_{\nu\rho}$, the distortion of ρ with respect to ν, i.e. the operator norm of

$$\mathcal{L}^2_\nu(X) \xrightarrow{J} \mathcal{L}^2_\rho(X).$$

Theorem 4. *In the general setting in Hilbert space, for $0 < r < s$, the approximation error satisfies*

$$\mathcal{E}(f_{\mathcal{H}}) = \|f_{\mathcal{H}} - f_\rho\|^2_\rho + \sigma^2_\rho \le \mathcal{D}^2_{\nu\rho}\left(\frac{1}{R}\right)^{\frac{2r}{s-r}} \|A^{-r}f_\rho\|_\nu^{\frac{2s}{s-r}} + \sigma^2_\rho.$$

Proof.

$$\|f_\rho - f_{\mathcal{H}}\|_\rho = \min_{g\in B_R}\|f_\rho - g\|_\rho \le \mathcal{D}_{\nu\rho}\min_{g\in B_R}\|f_\rho - g\|_\mu \le \mathcal{D}_{\nu\rho}\left(\frac{1}{R}\right)^{\frac{r}{s-r}} \|A^{-r}f_\rho\|_\nu^{\frac{s}{s-r}}$$

with the last inequality from Theorem 3 (2) with $H = \mathcal{L}^2_\nu(X)$ and $a = f_\rho$. ∎

While in Example 3 we always take $\nu = \mu$, the Lebesgue measure, in our most interesting example (Example 4) we will usually suppose $\nu = \rho$ so $\mathcal{D}_{\nu\rho} = 1$.

3. Approximation error in Sobolev spaces and RKHS

We continue our discussion of Example 3 in the context of the approximation error. In this section $X \subset \mathbb{R}^n$ is a compact domain with smooth boundary.

Theorem 5. *Let $s > n/2$ and r such that $0 < r < s$. Consider $R > 0$, B_R the ball of radius R in $H^s(X)$ and $\mathcal{H} = \overline{J_s(B_R)}$. Then the approximation error satisfies*

$$\mathcal{E}(f_{\mathcal{H}}) \leq \mathcal{D}^2_{\mu\rho} C \left(\frac{1}{R}\right)^{\frac{2r}{s-r}} (\|f_\rho\|_r)^{\frac{2s}{s-r}} + \sigma_\rho^2$$

where C is a constant which depends only on s, r and X.

Proof. Let $\Delta : H^2(X) \to \mathcal{L}^2_\mu(X)$ denote the Laplacian and $\mathcal{A} = (-\Delta + \mathrm{Id})^{-1/2}$. For all $\tau \geq 0$, $\mathcal{A}^\tau : \mathcal{L}^2_\mu(X) \to H^\tau(X)$ is a compact linear map with bounded inverse. There exist $C_0, C_1 > 0$ such that, for all $g \in H^\tau(X)$,

$$C_0\|g\|_\tau \leq \|\mathcal{A}^{-\tau} g\|_\mu \leq C_1 \|g\|_\tau. \tag{7}$$

By composing with the inclusion $H^1(X) \hookrightarrow \mathcal{L}^2_\mu(X)$ and slightly abusing notation we may assume $\mathcal{A} : \mathcal{L}^2_\mu(X) \to \mathcal{L}^2_\mu(X)$ and consider the general setting in Hilbert space.

Let $\mathbb{E}$ be the space defined in this setting with $A = \mathcal{A}$ and $s = \tau$. Then the ball $B_{RC_0}(\mathbb{E})$ of radius RC_0 in $\mathbb{E}$ is included in the ball $B_R(H^s(X))$ in $H^s(X)$ and consequently

$$\mathcal{E}(f_{\mathcal{H}}) = \min_{g \in B_R(H^s(X))} \|f_\rho - g\|^2_\rho + \sigma^2_\rho \leq \min_{g \in B_{RC_0}(\mathbb{E})} \|f_\rho - g\|^2_\rho + \sigma^2_\rho.$$

Now, apply Theorem 4 to obtain

$$\min_{g \in B_{RC_0}(\mathbb{E})} \|f_\rho - g\|^2_\rho + \sigma^2_\rho \leq \mathcal{D}^2_{\nu\rho} \left(\frac{1}{RC_0}\right)^{\frac{2r}{s-r}} \|\mathcal{A}^{-r} f_\rho\|_\mu^{\frac{2s}{s-r}} + \sigma^2_\rho.$$

Apply finally (7) with $\tau = r$ to get

$$\|\mathcal{A}^{-r} f_\rho\|_\mu \leq C_1 \|f_\rho\|_r.$$

The result follows by taking $C = C_0^{\frac{-2r}{s-r}} C_1^{\frac{2s}{s-r}}$.

For the facts about Sobolev spaces mentioned in this proof see [39]. ∎

Remark 5. (i) In Theorem 5 we have some freedom to choose r. For example if f_ρ is a characteristic function and $0 < r < 1/2$, then $\|f_\rho\|_r < \infty$ (see [38]) and we obtain information in the classification problem of learning theory.
(ii) The essence of Theorem 5, for the case $n = 1$, appears in [14].

It is also possible to use Theorem 4 to derive bounds for the approximation error in Example 4.

Theorem 6. *Let K be a Mercer kernel, ν a Borel measure on X, $R > 0$, and $\mathcal{H} = \overline{I_K(B_R)}$. The approximation error satisfies, for $0 < r < 1$,*

$$\mathcal{E}(f_{\mathcal{H}}) \leq \mathcal{D}^2_{\nu\rho} \left(\frac{1}{R}\right)^{\frac{2r}{1-r}} \|L_K^{-r/2} f_\rho\|_\nu^{\frac{2}{1-r}} + \sigma^2_\rho.$$

Proof. Take $A = L_K^{1/2}$ and $s = 1$ in Theorem 4. Then, we will see in Section 3 in Chapter III that for all $f \in \mathcal{L}^2_\nu(X)$, $\|f\|_K = \|A^{-1} f\|_\nu$, which implies that $\mathbb{E}$ is the reproducing kernel Hilbert space of Example 4. Now apply Theorem 4. ∎

4. The bias-variance problem

Consider the general setting in Hilbert space described in Remark 4. Fix a sample size m and a confidence $1-\delta$ with $0<\delta<1$. For each $R>0$ a hypothesis space $\mathcal{H}=\mathcal{H}_{\mathbb{E},R}$ is determined, and we can consider $f_{\mathcal{H}}$ and, for $\mathbf{z}\in Z^m$, $f_{\mathbf{z}}$. The *bias-variance problem in the general setting* consists of finding the value of R which minimizes a natural bound for the error $\mathcal{E}(f_{\mathbf{z}})$ (with confidence $1-\delta$). This value of R determines a particular hypothesis space in the family of such spaces parametrized by R, or, using a terminology common in the learning literature, it selects a *model.*

Theorem 7. *For all $m\in\mathbb{N}$ and $\delta\in\mathbb{R}$, $0<\delta<1$, and all r with $0<r<s$, there exists a unique solution R^* of the bias-variance problem in the general setting.*

Proof. We first describe the natural bound we are going to minimize. Recall that $\mathcal{E}(f_{\mathbf{z}})$ equals the sum $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})+\mathcal{E}(f_{\mathcal{H}})$ of the sample and approximation error. Theorem 4 bounds the approximation error, for $0<r<s$, by an expression

$$\alpha(R)=\mathcal{D}^2_{\nu\rho}\left(\frac{1}{R}\right)^{\frac{2r}{s-r}}\|A^{-r}f_\rho\|_\nu^{\frac{2s}{s-r}}+\sigma_\rho^2.$$

We now want to bound the sample error. To do so let

$$M=M(R)=\|J_{\mathbb{E}}\|R+M_\rho+\|f_\rho\|_\infty.$$

Then, almost everywhere, $|f(x)-y|\le M$ since

$$|f(x)-y|\le|f(x)|+|y|\le|f(x)|+|y-f_\rho(x)|+|f_\rho(x)|\le\|J_{\mathbb{E}}\|R+M_\rho+\|f_\rho\|_\infty.$$

By Theorem C*, the sample error ε with confidence $1-\delta$ satisfies

$$\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{24M}\right)e^{-\frac{m\varepsilon}{288M^2}}\ge\delta;$$

i.e.

$$\frac{m\varepsilon}{288M^2}-\ln\left(\frac{1}{\delta}\right)-\ln\left(\mathcal{N}\left(\mathcal{H},\frac{\varepsilon}{24M}\right)\right)\le 0.$$

Then, as in Remark 11 of Chapter I,

$$\frac{m\varepsilon}{288M^2}-\ln\left(\frac{1}{\delta}\right)-\left(\frac{24M^2C_{\mathbb{E}}}{\|J_{\mathbb{E}}\|\varepsilon}\right)^{1/\ell_E}\le 0$$

where we have also used that $R\|J_{\mathbb{E}}\|\le M$. Write $v=\frac{\varepsilon}{M^2}$. Then the inequality above takes the form

$$c_0v-c_1-c_2v^{-d}\le 0 \tag{8}$$

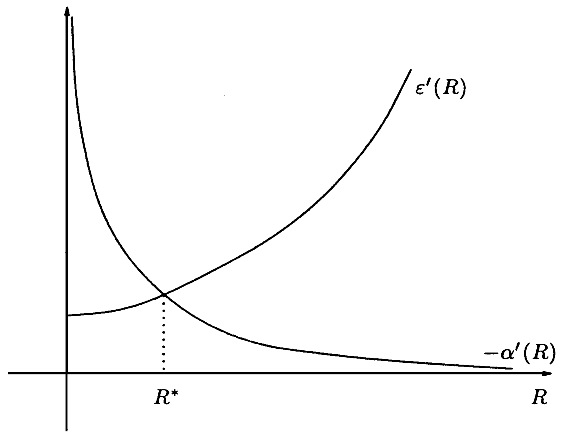
where $c_0=\frac{m}{288}$, $c_1=\ln\left(\frac{1}{\delta}\right)$, $c_2=\left(\frac{24C_{\mathbb{E}}}{\|J_{\mathbb{E}}\|}\right)^{1/\ell_E}$, and $d=1/\ell_{\mathbb{E}}$.

If we take the equality in (8) we obtain an equation which, it is easy to see, has exactly one positive solution for v. Let $v^*(m,\delta)$ be this solution. Then, $\varepsilon(R)=M^2v^*(m,\delta)$ is the best bound we can obtain from Theorem C* for the sample error.

We will therefore minimize $\alpha(R)+\varepsilon(R)$.

For a point $R>0$ to be a minimum of $\alpha(R)+\varepsilon(R)$ it is necessary that $\varepsilon'(R)=-\alpha'(R)$. Taking derivatives, we get

$$\alpha'(R)=-C_A\frac{-2r}{s-r}R^{\frac{-(s+r)}{s-r}}\qquad\text{and}\qquad\alpha''(R)=C_A\frac{2r(s+r)}{(s-r)^2}R^{\frac{-2s}{s-r}}$$



where $C_A = \mathcal{D}^2_{\nu\rho}\|A^{-r}f_\rho\|_\nu^{\frac{2s}{s-r}}$, and

$$\varepsilon'(R) = 2Mv^*(m,\delta) \qquad \text{and} \qquad \varepsilon''(R) = 2v^*(m,\delta).$$

Since $C_A \geq 0$ we deduce that $-\alpha'(R)$ is a positive function monotonically decreasing on $(0,+\infty)$. On the other hand, since $v^*(m,\delta) > 0$, it follows that $\varepsilon'(R)$ is a positive function strictly increasing on $(0,+\infty)$. Since $\varepsilon'(+\infty) = +\infty$, $-\alpha'(+\infty) = 0$, $\varepsilon'(0) < +\infty$ and $-\alpha'(0) = +\infty$, we deduce the existence of a unique R^* such that $\varepsilon'(R^*) = -\alpha'(R^*)$. □

For different instances of the general setting the value of R^* may be numerically computed.

Remark 6. In this section we considered a form of the bias-variance problem which optimized the parameter R fixing all the others. One may consider other forms of the bias-variance problem by optimizing other parameters. For instance, in Example 4, one may consider the degree of smoothness of the kernel K. The smoother K is, the smaller $\mathcal{H}_K$ is. Therefore, the sample error decreases and the approximation error increases with a parameter reflecting this smoothness.

Chapter III: Algorithms

1. Operators defined by a kernel

Recall that X is a compact domain or manifold in Euclidean space with $\dim X = n$. However, for much of this chapter it is sufficient to take X to be a compact metric space. Let ν be a Borel measure on X and $\mathcal{L}^2_\nu(X)$ be the Hilbert space of square integrable functions on X. Note that ν can be any Borel measure. Significant particular cases are Lebesgue measure or the marginal measure ρ_X of Chapter I.

Let $K : X \times X \to \mathbb{R}$ be a continuous function. Then the linear map

$$L_K : \mathcal{L}^2_\nu(X) \to \mathcal{C}(X)$$

given by the following integral transform

$$(L_K f)(x) = \int K(x,t)f(t)d\nu(t)$$

is well-defined. Composition with the inclusion $\mathcal{C}(X) \hookrightarrow \mathcal{L}^2_\nu(X)$ yields a linear operator $L_K : \mathcal{L}^2_\nu(X) \to \mathcal{L}^2_\nu(X)$ which, abusing notation, we will also denote by L_K.

The function K is said to be the *kernel* of L_K and several properties of L_K follow from properties of K. Let

$$\mathbf{C}_K = \sup_{x,t \in X} |K(x,t)|.$$

Also, for $x \in X$, let $K_x : X \to \mathbb{R}$ be given by $K_x(t) = K(x,t)$.

Proposition 1. *If K is continuous, then L_K is well-defined and compact. In addition, $\|L_K\| \le \sqrt{\nu(X)}\mathbf{C}_K$. Here $\nu(X)$ denotes the measure of X.*

Proof. To see that L_K is well defined we need to show that $L_K f$ is continuous for every $f \in \mathcal{L}^2_\nu(X)$. To do so, consider $f \in \mathcal{L}^2_\nu(X)$ and $x_1, x_2 \in X$. Then

$$\begin{aligned} |(L_K f)(x_1) - (L_K f)(x_2)| &= \left| \int (K(x_1,t) - K(x_2,t)) f(t) \right| \\ &\le \|K_{x_1} - K_{x_2}\| \|f\| \quad \text{by Cauchy-Schwartz} \\ &\le \sqrt{\nu(X)} \max_{t \in X} |K(x_1,t) - K(x_2,t)| \|f\|. \end{aligned}$$

Since K is continuous and X is compact, K is uniformly continuous. This implies the continuity of $L_K f$.

The assertion $\|L_K\| \le \sqrt{\nu(X)}\mathbf{C}_K$ follows from the inequality

$$|(L_K f)(x)| \le \sqrt{\nu(X)} \sup_{t \in X} |K(x,t)| \|f\|$$

which is proved as above.

Finally, to see that L_K is compact, let (f_n) be a bounded sequence in $\mathcal{L}^2_\nu(X)$. Since $\|L_K f\|_\infty \le \mathbf{C}_K \|f\|$ we have that $(L_K f_n)$ is uniformly bounded. And, since $|(L_K f_n)(x_1) - (L_K f_n)(x_2)| \le \sqrt{\nu(X)} \max_{t \in X} |K(x_1,t) - K(x_2,t)| \|f_n\|$ for all $n \ge 1$, we have that the sequence $(L_K f_n)$ is equicontinuous. By Arzela's Theorem (see e.g. §11.4 of [20]), $(L_K f_n)$ contains a uniformly convergent subsequence. □

Two more important properties of L_K follow from properties of K. Recall that we say that K is *positive definite* if for all finite sets $\{x_1, \ldots, x_k\} \subset X$ the $k \times k$ matrix $K[\mathbf{x}]$ whose (i,j) entry is $K(x_i, x_j)$ is positive definite.

Proposition 2. (a) *If K is symmetric, then $L_K : \mathcal{L}^2_\nu(X) \to \mathcal{L}^2_\nu(X)$ is self-adjoint.*

(b) *If, in addition, K is positive definite, then L_K is positive.*

Proof. Part (a) follows easily from Fubini's Theorem and the symmetry of K. For (b), just note that

$$\begin{aligned} \int\int K(x,t) f(x) f(t) &= \lim_{k \to \infty} \frac{\nu(X)}{k^2} \sum_{i,j=1}^{k} K(x_i, x_j) f(x_i) f(x_j) \\ &= \lim_{k \to \infty} \frac{\nu(X)}{k^2} f_{\mathbf{x}}^{\mathrm{T}} K[\mathbf{x}] f_{\mathbf{x}} \end{aligned}$$

where, for all $k \ge 1$, $x_1, \ldots, x_k \in X$ is a set of points conveniently chosen, $f_{\mathbf{x}} = (f(x_1), \ldots, f(x_k))$ and $K[\mathbf{x}]$ is the $k \times k$ matrix whose (i,j) entry is $K(x_i, x_j)$. Since this matrix is positive definite the result follows. □

In the sequel we will consider a Mercer kernel K (i.e. a function $K : X \times X \to \mathbb{R}$ which is continuous, symmetric and positive definite). Then $L_K : \mathcal{L}^2_\nu(X) \to \mathcal{L}^2_\nu(X)$ is a self-adjoint, positive, compact operator and the Spectral Theorem (Theorem 2 of Chapter II) applies. Let λ_k, $k \geq 1$, denote the eigenvalues of L_K and ϕ_k the corresponding eigenfunctions.

Corollary 2. *For $k \geq 1$, if $\lambda_k \neq 0$, then ϕ_k is continuous on X.*

Proof. Use that $\phi_k = \frac{1}{\lambda_k} L_K(\phi_k)$. $\square$

In the sequel we will assume, without loss of generality, that $\lambda_k \geq \lambda_{k+1}$ for all $k \geq 1$.

2. Mercer's Theorem

If $f \in \mathcal{L}^2_\nu(X)$ and $\{\phi_1, \phi_2, \ldots\}$ is a Hilbert basis of $\mathcal{L}^2_\nu(X)$, f can be uniquely written as $f = \sum_{k=1}^\infty a_k \phi_k$ and the partial sums $\sum_{k=1}^N a_k \phi_k$ converge to f in $\mathcal{L}^2_\nu(X)$. If this convergence also holds in $\mathcal{C}(X)$, we say that the series *uniformly converges* to f. Also, we say that a series $\sum a_k$ *converges absolutely* if the series $\sum |a_k|$ is convergent.

Theorem 1. *Let X be a compact domain or a manifold, ν a Borel measure on X, and $K : X \times X \to \mathbb{R}$ a Mercer kernel. Let λ_k be the kth eigenvalue of L_K and $\{\phi_k\}_{k \geq 1}$ the corresponding eigenvectors. For all $x, t \in X$, $K(x,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x) \phi_k(t)$ where the convergence is absolute (for each $x, y \in X \times X$) and uniform (on $X \times X$).*

The proof of Theorem 1 is given in [19] for $X = [0,1]$ and ν the measure inherited by the Lebesgue measure on $\mathbb{R}$, but the proof there is valid in the generality of our statement.

Corollary 3. *The sum $\sum \lambda_k$ is convergent and*

$$\sum_{k=1}^{\infty} \lambda_k = \int_X K(x,x) \leq \nu(X) \mathbf{C}_K.$$

Therefore, for all $k \geq 1$, $\lambda_k \leq \left(\frac{\nu(X)\mathbf{C}_K}{k} \right)$.

Proof. By taking $x = t$ in Theorem 1 we get $K(x,x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)^2$. Integrating on both sides of this equality, we get

$$\sum_{k=1}^{\infty} \lambda_k \int_X \phi_k(x)^2 = \int_X K(x,x) \leq \nu(X) \mathbf{C}_K.$$

But since $\{\phi_1, \phi_2, \ldots\}$ is a Hilbert basis, $\int \phi_k^2 = 1$ for all $k \geq 1$ and the first statement follows. The second statement follows from the assumption $\lambda_k \geq \lambda_j$ for $j > k$. $\square$

3. Reproducing kernel Hilbert spaces

In this section we fix a compact domain or a manifold X, a Borel measure ν on X, and a Mercer kernel $K : X \times X \to \mathbb{R}$. The two main results of this section are the following.

Theorem 2. *There exists a unique Hilbert space $\mathcal{H}_K$ of functions on X satisfying the following conditions:*

(i) *for all $x \in X$, $K_x \in \mathcal{H}_K$;*
(ii) *the span of the set $\{K_x \mid x \in X\}$ is dense in $\mathcal{H}_K$; and*
(iii) *for all $f \in \mathcal{H}_K$, $f(x) = \langle K_x, f\rangle_K$.*

Moreover, $\mathcal{H}_K$ consists of continuous functions, and the inclusion $I_K : \mathcal{H}_K \to \mathcal{C}(X)$ is bounded with $\|I_K\| \leq \mathbf{C}_K^{1/2}$.

Theorem 3. *The map*

$$\begin{array}{rcl} \Phi : X & \to & \ell^2 \\ x & \mapsto & (\sqrt{\lambda_k}\phi_k(x))_{k\in\mathbb{N}} \end{array}$$

is well-defined, continuous, and satisfies

$$K(x,t) = \langle \Phi(x), \Phi(t)\rangle.$$

Corollary 4. *For all $x, t \in X$, $|K(x,t)| \leq K(x,x)^{1/2}K(t,t)^{1/2}$.*

Proof. This is a consequence of the Cauchy-Schwartz inequality and the last statement in Theorem 3. ☐

Remark 1. (i) Note that the space $\mathcal{H}_K$ of Theorem 2 depends only on X and K. It is independent of any measure considered on X.
(ii) In the learning context, the space ℓ^2 in Theorem 3 is often called the *feature space* and the function Φ the *feature map.*
(iii) The Hilbert space $\mathcal{H}_K$ in Theorem 2 is said to be a *reproducing kernel Hilbert space* (or, for short, a RKHS). This terminology is of common use in the learning literature.
(iv) A substantial amount of the theory of reproducing kernel Hilbert spaces was developed by N. Aronszajn [2]. On page 344 of this reference, Theorem 2, in essence, is attributed to E.H. Moore.

Proof of Theorem 2. Let H_0 be the span of the set $\{K_x \mid x \in X\}$. We define an inner product in H_0 as follows. If $f = \sum_{i=1}^{s} \alpha_i K_{x_i}$ and $g = \sum_{j=1}^{r} \beta_i K_{t_j}$, then

$$\langle f, g\rangle = \sum_{\substack{1\leq i\leq s\\ 1\leq j\leq r}} \alpha_i\beta_j K(x_i, t_j).$$

Let $\mathcal{H}_K$ be the completion of H_0 with the associated norm. It is easy to check that $\mathcal{H}_K$ satisfies the three conditions in the statement. We only need to prove that it is unique. So, assume H is another Hilbert space of functions on X satisfying the noted conditions. We want to show that

$$H = \mathcal{H}_K \quad \text{and} \quad \langle\ ,\ \rangle_H = \langle\ ,\ \rangle_{\mathcal{H}_K}. \tag{9}$$

We first observe that $H_0 \subset H$. Also, for any $x, t \in X$, $\langle K_x, K_t\rangle_H = K(x,t) = \langle K_x, K_t\rangle_{\mathcal{H}_K}$. By linearity, for every $f, g \in H_0$, $\langle f, g\rangle_H = \langle f, g\rangle_{\mathcal{H}_K}$. Since both H and $\mathcal{H}_K$ are completions of H_0, (9) follows from the uniqueness of the completion.

To see the remaining assertion consider $f \in \mathcal{H}_K$ and $x \in X$. Then

$$|f(x)| = |\langle K_x, f\rangle| \leq \|f\|\|K_x\| = \|f\|\sqrt{K(x,x)}.$$

This implies $\|f\|_\infty \leq \sqrt{\mathbf{C}_K}\|f\|_{\mathcal{H}_K}$ and thus $\|I_K\| \leq \sqrt{\mathbf{C}_K}$. Therefore, convergence in $\| \, \|_{\mathcal{H}_K}$ implies convergence in $\| \, \|_\infty$, and this shows that f is continuous since f is the limit of elements in H_0 which are continuous. $\square$

Proof of Theorem 3. For every $x \in X$, by Mercer's Theorem, $\sum \lambda_k \phi_k^2(x)$ converges to $K(x,x)$. This shows that $\Phi(x) \in \ell^2$.

Also by Mercer's Theorem, for every $x, t \in X$,

$$K(x,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x)\phi_k(t) = \langle \Phi(x), \Phi(t)\rangle.$$

It only remains to prove that $\Phi : X \to \ell^2$ is continuous. But for any $x, t \in X$,

$$\begin{aligned} \|\Phi(x) - \Phi(t)\| &= \langle \Phi(x), \Phi(x)\rangle + \langle \Phi(t), \Phi(t)\rangle - 2\langle \Phi(x), \Phi(t)\rangle \\ &= K(x,x) + K(t,t) - 2K(x,t) \end{aligned}$$

which by the continuity of K tends to zero when x tends to t. $\square$

We next characterize $\mathcal{H}_K$ through the eigenvalues λ_k of L_K. Theorem 2 of Chapter II guarantees that $\lambda_k \geq 0$ for all $k \geq 1$. In the rest of this section we assume that, in addition, $\lambda_k > 0$ for all $k \geq 1$. There is no loss of generality in doing so (see Remark 3 below).

Let

$$H_K = \left\{ f \in \mathcal{L}^2_\nu(X) \mid f = \sum_{k=1}^{\infty} a_k \phi_k \quad \text{with} \quad \left(\frac{a_k}{\sqrt{\lambda_k}}\right) \in \ell^2 \right\}.$$

We can make H_K a Hilbert space with the inner product

$$\langle f, g\rangle_K = \sum_{k=1}^{\infty} \frac{a_k b_k}{\lambda_k}$$

for $f = \sum a_k \phi_k$ and $g = \sum b_k \phi_k$. Note that the map

$$\begin{aligned} L_K^{1/2} : \mathcal{L}^2_\nu(X) &\to H_K \\ \sum a_k \phi_k &\mapsto \sum a_k \sqrt{\lambda_k} \phi_k \end{aligned}$$

defines an isomorphism of Hilbert spaces. In addition, considered as an operator on $\mathcal{L}^2_\nu(X)$, it is the square root of L_k in the sense that $L_K = L_K^{1/2} \circ L_K^{1/2}$.

Proposition 3. *The elements of H_K are continuous functions on X. In addition, for $f \in H_K$, if $f = \sum a_k \phi_k$, then this series converges absolutely and uniformly to f.*

Proof. Let $g \in H_K$, $g = \sum g_k \phi_k$, and $x \in X$. Then

$$|g(x)| = \left|\sum_{k=1}^{\infty} g_k \phi_k(x)\right| = \left|\sum_{k=1}^{\infty} \frac{g_k}{\sqrt{\lambda_k}} \sqrt{\lambda_k}\phi_k(x)\right| \leq \|g\|_{\mathrm{K}} \|\Phi(x)\| = \|g\|_{\mathrm{K}} K(x,x)^{1/2},$$

the inequality by Cauchy-Schwartz and the last equality by Theorem 1. Thus, $\|g\|_\infty \leq \sqrt{\mathbf{C}_K}\|g\|_{\mathrm{K}}$. Therefore, convergence in $\| \, \|_K$ implies convergence in $\| \, \|_\infty$

which, applied to the series $g_N = f - \sum_{k=1}^{N} a_k\phi_k$, proves the statement about uniform convergence. The continuity of f now follows from that of the ϕ_k (Corollary 2). The absolute convergence follows from the inequality $\sum |g_k\phi_k(x)| \leq \|g\|_K \|\Phi(x)\|$. □

Lemma 1. *Let $x \in X$. The function $\varphi_x : X \to \mathbb{R}$ defined by $\varphi_x(t) = \langle \Phi(x), \Phi(t)\rangle$ belongs to H_K.*

Proof. Use Theorem 3. □

Proposition 4. *For all $f \in H_K$ and all $x \in X$, $f(x) = \langle f, K_x\rangle_K$.*

Proof. For $f \in H_K$, $f = \sum w_k\phi_k$,

$$\begin{aligned}
\langle f, K_x\rangle_{\mathrm{K}} &= \sum_{k=1}^{\infty} w_k\langle \phi_k, K_x\rangle_{\mathrm{K}} = \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k}\langle \phi_k, K_x\rangle \\
&= \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k}\int \phi_k(t)K(x,t) = \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k}(L_K\phi_k)(x) = \sum_{k=1}^{\infty} \frac{w_k}{\lambda_k}\lambda_k\phi_k(x) \\
&= f(x).
\end{aligned}$$

□

Theorem 4. *The Hilbert spaces $\mathcal{H}_K$ and H_K are the same space of functions on X with the same inner product.*

Proof. For any $x \in X$, the function K_x coincides, by Theorem 3, with the function φ_x in the statement of Lemma 1. And this result shows precisely that $\varphi_x \in H_K$. In addition, Proposition 4 shows that for all $f \in H_K$ and all $x \in X$, $f(x) = \langle f, K_x\rangle_{\mathrm{K}}$. We now show that the span of $\{K_x \mid x \in X\}$ is dense in H_K.

To do so, assume that for $f \in H_K$, $\langle f, K_t\rangle_{\mathrm{K}} = 0$ for all $t \in X$. Then, since $\langle f, K_t\rangle_{\mathrm{K}} = f(t)$, we have $f = 0$ on X. This implies the desired density.

The statement now follows from Theorem 2. □

Remark 2. A consequence of Theorem 4 is the fact that the Hilbert space H_K, although being defined through the integral operator L_K and its associated spectra which depend on the measure ν, is actually independent of ν. This follows from Remark 1.

Remark 3. The properties of H_K and Φ have been exposed under the assumption that all eigenvalues of L_K are strictly positive. If the eigenvalues might be zero as well, let $\mathbb{H}$ be the linear subspace of $\mathcal{L}^2_\nu(X)$ spanned by the eigenvectors corresponding to non-zero eigenvalues. If $\mathbb{H}$ is infinite dimensional, all the results in this section remain true if one replaces $\mathcal{L}^2_\nu(X)$ by $\mathbb{H}$. If $\mathbb{H}$ is finite dimensional, this is so if, in addition, we replace ℓ^2 by $\mathbb{R}^N$ where $N = \dim\mathbb{H}$.

4. Mercer kernels exist

Given a kernel K, it is in general straightforward to check its symmetry and continuity. It is more involved to check that it is positive definite. The next result, Proposition 5 below, will be helpful to prove positivity of several kernels. It was originally proved for $\mathbb{R}^n$ by Schoenberg [34] (together with a more difficult converse), but it follows for subsets of $\mathbb{R}^n$ by restricting to such a subset a kernel defined on $\mathbb{R}^n$.

A function $f:(0,\infty)\to\mathbb{R}$ is *completely monotonic* if it is $\mathcal{C}^\infty$ and, for all $r>0$ and $k\geq 0$, $(-1)^k f^{(k)}(r)\geq 0$. Here $f^{(k)}$ denotes the kth derivative of f.

Proposition 5. *Let $X\subset\mathbb{R}^n$, $f:(0,\infty)\to\mathbb{R}$ and $K:X\times X\to\mathbb{R}$ be defined by $K(x,t)=f(\|x-t\|^2)$. If f is completely monotonic, then K is positive definite.* □

Corollary 5. *Let $c\neq 0$. The following kernels, defined on a compact domain $X\subset\mathbb{R}^n$, are Mercer kernels.*

(a) **[Gaussian]** $K(x,t)=e^{-\frac{\|x-t\|^2}{c^2}}$.
(b) $K(x,t)=(c^2+\|x-t\|^2)^{-\alpha}$ *with* $\alpha>0$.

Proof. Clearly, both kernels are continuous and symmetric. In (a) K is positive definite by Proposition 5 with $f(r)=e^{-\frac{r}{c^2}}$. The same for (b) taking $f(r)=(c^2+r)^{-\alpha}$. □

Remark 4. The kernels of (a) and (b) in Corollary 5 satisfy $\mathbf{C}_K=1$ and $\mathbf{C}_K=c^{-2\alpha}$ respectively.

The following is a key example of finite dimensional RKHS induced by a Mercer kernel. In contrast with the Mercer kernels of Corollary 5 we will not use Proposition 5 to show positivity.

Example 1 (continued). Let $\mathcal{H}_d=\mathcal{H}_d(\mathbb{R}^{n+1})$ be the linear space of homogeneous polynomials of degree d in $x_0,x_1,\ldots,x_n$. Thus, we recall, elements $f\in\mathcal{H}_d$ have the form $f=\sum_{|\alpha|=d} w_\alpha x^\alpha$ with $\alpha=(\alpha_0,\alpha_1,\ldots,\alpha_n)\in\mathbb{N}^{n+1}$. It follows that the dimension of $\mathcal{H}_d$ is

$$N=\binom{n+d}{n}.$$

We can make $\mathcal{H}_d$ an inner product space by taking

$$\langle f,g\rangle=\sum_{|\alpha|=d} w_\alpha v_\alpha (C_\alpha^d)^{-1}$$

for $f,g\in\mathcal{H}_d$, $f=\sum w_\alpha x^\alpha$, $g=\sum v_\alpha x^\alpha$. Here

$$C_\alpha^d=\frac{d!}{\alpha_0!\cdots\alpha_n!}$$

is the multinomial coefficient associated to the pair (d,α). This inner product, which we call the *Weyl inner product*, is natural and has important properties such as group invariance. If $\|f\|$ denotes the norm induced by this inner product, then one has

$$|f(x)|\leq\|f\|\|x\|^d$$

where $\|x\|$ is the standard norm of $x\in\mathbb{R}^{n+1}$ (cf. Lemma 7 of Chapter 14 of [8]; this reference gives more background to this discussion).

Let $X=S(\mathbb{R}^{n+1})$ and

$$\begin{aligned} K:X\times X &\to \mathbb{R}\\ (x,t) &\mapsto \langle x,t\rangle^d \end{aligned}$$

where $\langle\ ,\ \rangle$ denotes the Euclidean inner product in $\mathbb{R}^{n+1}$. Let also

$$\begin{aligned} \Phi:X &\to \mathbb{R}^N\\ x &\mapsto \left(x^\alpha (C_\alpha^d)^{1/2}\right). \end{aligned}$$

Then, for $x, t \in X$, we have

$$\langle \Phi(x), \Phi(t) \rangle = \sum_{|\alpha|=d} x^\alpha t^\alpha C^d_\alpha = \langle x, t \rangle^d = K(x, t).$$

This equality enables us to prove that K is positive definite since it implies that, for $t_1, \ldots, t_k \in X$, the entry in row i and column j of $K[\mathbf{t}]$ is $\langle \Phi(t_i), \Phi(t_j) \rangle$. Therefore, if M denotes the matrix whose jth column is $\Phi(t_j)$, we have that $K[\mathbf{t}] = M^{\mathrm{T}} M$ from which the positivity of $K[\mathbf{t}]$ follows. Since K is clearly continuous and symmetric, we conclude that K is a Mercer kernel.

Which is the RKHS associated to K?

Proposition 6. *$\mathcal{H}_d = \mathcal{H}_K$ as function spaces and inner product spaces.*

Proof. We know from the proof of Theorem 2 that $\mathcal{H}_K$ is the completion of H_0, the span of $\{K_x \mid x \in X\}$. Since $H_0 \subseteq \mathcal{H}_d$ and $\mathcal{H}_d$ has finite dimension, the same holds for H_0. But then H_0 is complete and we deduce

$$\mathcal{H}_K = H_0 \subseteq \mathcal{H}_d.$$

The map $\mathcal{V} : \mathbb{R}^{n+1} \to \mathbb{R}^N$ defined by $\mathcal{V}(x) = (x^\alpha)_{|\alpha|=d}$ is a well-known object in algebraic geometry, where it receives the name of *Veronese embedding*. We note here that the feature map Φ defined above is related to $\mathcal{V}$ since for every $x \in X$, $\Phi(x) = D\mathcal{V}(x)$ where D is the diagonal matrix with entries $(C^d_\alpha)^{1/2}$. The image of $\mathbb{R}^{n+1}$ by the Veronese embedding is an algebraic variety called the *Veronese variety*, which is known (cf. §4.4 of [35]) to be non-degenerate, i.e. to span all of $\mathbb{R}^N$. This implies that $\mathcal{H}_K = \mathcal{H}_d$ as vector spaces. We will now see that they are actually the same inner product space.

By definition of the inner product in H_0, for all $x, t \in X$,

$$\langle K_x, K_t \rangle_{H_0} = K(x, t) = \sum_{|\alpha|=d} C^d_\alpha x^\alpha t^\alpha.$$

On the other hand, since $K_x(w) = \sum_{|\alpha|=d} C^d_\alpha x^\alpha w^\alpha$, we have that the Weyl inner product of K_x and K_t satisfies

$$\langle K_x, K_t \rangle_{\mathcal{H}_d} = \sum_{|\alpha|=d} (C^d_\alpha)^{-1} C^d_\alpha x^\alpha C^d_\alpha t^\alpha = \sum_{|\alpha|=d} C^d_\alpha x^\alpha t^\alpha.$$

We conclude that, since the polynomials K_x span all of H_0, the inner product in $\mathcal{H}_K = H_0$ is the Weyl inner product. □

The discussion above extends to arbitrary, i.e. not necessarily homogeneous, polynomials. Let $\mathcal{P}_d = \mathcal{P}_d(\mathbb{R}^n)$ be the linear space of polynomials of degree d in $x_1, \ldots, x_n$. A natural isomorphism between $\mathcal{P}_d$ and $\mathcal{H}_d$ is the "homogenization"

$$\begin{array}{ccc} \mathcal{P}_d & \to & \mathcal{H}_d \\ \displaystyle\sum_{|\alpha|\le d} w_\alpha x^\alpha & \mapsto & \displaystyle\sum_{|\alpha|\le d} w_\alpha x_0^{d-|\alpha|} x^\alpha. \end{array}$$

Here, $\alpha = (\alpha_1, \ldots, \alpha_n) \in \mathbb{N}^n$ is a "multi-index" and $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$. The inverse of the homogenization is obtained by setting $x_0 = 1$. Through this isomorphism we can endow $\mathcal{P}_d$ as well with the Weyl inner product.

Let

$$\begin{array}{rcl} K : \mathbb{R}^n \times \mathbb{R}^n & \to & \mathbb{R} \\ (x,t) & \mapsto & (1 + \langle x, t\rangle)^d \end{array}$$

and $\Phi : \mathbb{R}^n \to \mathbb{R}^N$ given by $\Phi(x) = \big(1, x^\alpha (C^d_\alpha)^{1/2}\big)$. Then, one has $\langle \Phi(x), \Phi(t)\rangle = K(x,t)$.

Remark 5. Note again that the reproducing kernel Hilbert structure on $\mathcal{H}_d$ for $K(x,t) = \langle x, t\rangle^d$ is precisely the Weyl one.

5. Covering numbers on reproducing kernel Hilbert spaces

The goal of this section is to estimate the covering number $\mathcal{N}(\overline{I_K(B_R)}, \eta)$ for $R, \eta > 0$ as promised.

Theorem D. *Let $K : X \times X \to \mathbb{R}$ be a $\mathcal{C}^\infty$ Mercer kernel and $\mathcal{H}_K$ its corresponding RKHS. Then the inclusion $I_K : \mathcal{H}_K \hookrightarrow \mathcal{C}(X)$ is compact and its entropy numbers satisfy $e_k(I_K) \le C'_h k^{-h/2n}$, for all $h > n$, where C'_h is independent of k. Consequently, for $h > n$, $\eta > 0$, and $R > 0$,*

$$\ln \mathcal{N}\left(\overline{I_K(B_R)}, \eta\right) \le \left(\frac{RC_h}{\eta}\right)^{\frac{2n}{h}}$$

where C_h is a constant slightly larger than C'_h.

Lemma 2. *Let $0 < r < s$ and $a \in \mathcal{L}^2_\mu(X)$. Suppose there exists $C > 0$ such that, for all $R > 0$,*

$$\min_{b \text{ s.t.} \|b\|_s \le R} \|b - a\| \le C \left(\frac{1}{R}\right)^{\frac{r}{s-r}}.$$

Then, for all $\delta > 0$, $\|a\|_{r-\delta} \le c_\delta C^{\frac{s-r}{s}}$.

Proof. See e.g. Theorem 2 in [38] (take $E = \mathcal{L}^2_\mu(X)$, $\mathcal{H} = H^s(X)$ and $\theta = r/s$). If for all $R > 0$

$$\min_{b \text{ s.t. } \|b\|_s \le R} \|b - a\| \le C \left(\frac{1}{R}\right)^{\frac{r}{s-r}},$$

then $\|a\|_{r/s,\infty} \le 2C^{\frac{s-r}{s}}$ where $\|\ \|_{r/s,\infty}$ denotes a norm in an interpolation space whose precise definition will not be needed here. Actually, in [4], pages 46 and 55, equation (1x), it is proved that, for all $\delta > 0$, there exists a constant C_δ such that, for all a in this interpolation space,

$$\|a\|_{r-\delta} \le C_\delta \|a\|_{r/s,\infty}.$$

The proof now follows by taking $c_\delta = 2C_\delta$. ∎

Lemma 3. *Let K be a $\mathcal{C}^\infty$ Mercer kernel. Then, the image of L_K is included in $H^\tau(X)$ for all $\tau \ge 0$. Considered as a linear map from $\mathcal{L}^2_\mu(X)$ to $H^\tau(X)$, L_K is bounded.*

Proof. For $f \in \mathcal{L}^2_\mu(X)$,

$$\begin{aligned} \|L_K f\|_\tau^2 &= \int_{x\in X} \sum_{|\alpha|\le\tau} (D^\alpha(L_K f)(x))^2 = \int_{x\in X} \sum_{|\alpha|\le\tau} \left(\int_{t\in X} D_x^\alpha K_t(x) f(t)\right)^2 \\ &\le \int_{x\in X} \sum_{|\alpha|\le\tau} \int_{t\in X} (D_x^\alpha K_t(x))^2 \int_{t\in X} f(t)^2 \\ &\le \|f\|_0^2 \mu(X) \sum_{|\alpha|\le\tau} \sup_{x,t\in X} (D_x^\alpha K_t(x))^2 \end{aligned}$$

where the first inequality follows from the Cauchy-Schwartz inequality. □

Proof of Theorem D. Let $f \in \mathcal{H}_K$ and $R > 0$. By Theorem 3 (2) in Chapter II with $A = L_K$, $s = 1$, $r = 1/2$ and $a = f$, we have

$$\min_{g \text{ s.t. } \|L_K^{-1} g\| \le R} \|g - f\| \le \frac{1}{R}\|L_K^{-1/2} f\|^2 = \frac{1}{R}\|f\|_K^2.$$

Let $\tau > 0$ and $\mathbf{c}_\tau = \|L_K\|$ for $L_K : \mathcal{L}^2_\mu(X) \to H^\tau(X)$. By Lemma 3 above,

$$\min_{g \text{ s.t. } \|g\|_\tau \le \frac{R}{\mathbf{c}_\tau}} \|g - f\| \le \frac{1}{R}\|f\|_K^2$$

or replacing $R/\mathbf{c}_\tau$ by R,

$$\min_{g \text{ s.t. } \|g\|_\tau \le R} \|g - f\| \le \frac{\mathbf{c}_\tau}{R}\|f\|_K^2.$$

Since this inequality holds for all $R > 0$ we can apply Lemma 2. We do so with $s = \tau = 3h/2$, $r = 3h/4$, $\delta = h/4$, and $C = \mathbf{c}_\tau \|f\|_K^2$ to obtain

$$\|f\|_{h/2} \le C'\|f\|_K \tag{10}$$

where $C' = c_\delta \sqrt{\mathbf{c}_{3h/2}}$.

Inequality (10) proves the existence of a bounded embedding $\mathcal{H}_K \hookrightarrow H^{h/2}$. Also, since $h > n$, the Sobolev Embedding Theorem and Rellich's Theorem apply to yield a compact embedding $H^{h/2} \hookrightarrow \mathcal{C}(X)$. From this we deduce the following factorization

$$\begin{tikzcd} \mathcal{H}_K \arrow[r, "I_K"] \arrow[dr, "\mathcal{J}"'] & \mathcal{C}(X) \\ & H^{h/2} \arrow[u, "J_{h/2}"'] \end{tikzcd}$$

which shows that I_K is compact.

In addition, by Edmunds and Triebel's bound (inequality (4) in Chapter I), we have $e_k(J_{h/2}) \le C\left(\frac{1}{k}\right)^{h/2n}$ for a constant C independent of k. Therefore

$$e_k(I_K) = e_k(J_{h/2}\mathcal{J}) \le e_k(J_{h/2})\|\mathcal{J}\| \le C'C\left(\frac{1}{k}\right)^{h/2n},$$

which proves the first statement in the theorem by taking $C_h' = C'C$.

The second statement follows by using that $\mathcal{N}(\overline{I_K(B_R)}, \eta) \le 2^k - 1$ if and only if $e_k(I_K) \le \eta/R$ and solving for k. □

6. On the minimizer of $\mathcal{E}_{\mathbf{z}}(f) + \gamma\|f\|_K^2$

Let $X, \mathcal{L}^2_\nu(X), K, \|\ \|_K$ and $\mathcal{H}_K$ be as in Section 1. We now abandon the setting of a compact hypothesis space adopted in Chapter I and slightly change the perspective. In what follows, we take $\mathcal{H} = \mathcal{H}_K$, i.e. $\mathcal{H}$ is a whole linear space, and we consider the *regularized error* $\mathcal{E}_\gamma$ defined by

$$\mathcal{E}_\gamma(f) = \int_Z (f(x) - y)^2 + \gamma\|f\|_K^2$$

for a fixed $\gamma \geq 0$. For a sample $\mathbf{z}$, the *regularized empirical error* $\mathcal{E}_{\gamma,\mathbf{z}}$ is defined in Proposition 8 below. One may consider a target function f_γ minimizing $\mathcal{E}_\gamma(f)$ over $\mathcal{H}$. But since $\mathcal{H}$ is no longer compact, the existence of such a target function is not immediate. Our next result proves that f_γ exists and is unique.

Proposition 7. *For all $\gamma > 0$ the function $f_\gamma = (\mathrm{Id} + \gamma L_K^{-1})^{-1} f_\rho$ is the unique minimizer of $\mathcal{E}_\gamma$ over $\mathcal{H}$.*

Proof. Apply Theorem 3 (1) of Chapter II with $H = \mathcal{L}^2_\nu(X)$, $s = 1$, $A = L_K^{1/2}$, and $a = f_\rho$. Since for all $f \in \mathcal{H}_K$, $\|f\|_K = \|L_K^{-1/2} f\|_\nu$, the expression $\|b - a\|^2 + \gamma\|A^{-s}b\|^2$ is in our case $\mathcal{E}_\gamma(b)$. Thus, f_γ is the $\hat{b}$ in Theorem 3 and the proposition follows. ☐

For the following result we have followed [17] and its references. See also the earlier paper [10].

Proposition 8. *Let $\mathbf{z} \in Z^m$ and $\gamma \in \mathbb{R}$, $\gamma > 0$. The empirical target, i.e. the function $f_{\gamma,\mathbf{z}} = f_{\mathbf{z}}$ minimizing the regularized empirical error*

$$\frac{1}{m}\sum_{i=1}^m (y_i - f(x_i))^2 + \gamma\|f\|_K^2$$

over $f \in \mathcal{H}_K$, may be expressed as

$$f_{\mathbf{z}}(x) = \sum_{i=1}^m a_i K(x, x_i)$$

where $a = (a_1, \ldots, a_m)$ is the unique solution of the well-posed linear system in $\mathbb{R}^m$

$$(\gamma m \,\mathrm{Id} + K[\mathbf{x}])a = \mathbf{y}.$$

Here, we recall, $K[\mathbf{x}]$ is the $m \times m$ matrix whose (i, j) entry is $K(x_i, x_j)$, $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$, and $\mathbf{y} = (y_1, \ldots, y_m) \in Y^m$ such that $\mathbf{z} = ((x_1, y_1), \ldots, (x_m, y_m))$.

Proof. Let $H(f) = \dfrac{1}{m}\displaystyle\sum_{i=1}^m (y_i - f(x_i))^2 + \gamma\|f\|_K^2$ and write, for any $f \in \mathcal{H}_K$, $f = \displaystyle\sum_{k=1}^\infty c_k\phi_k$. Recall that $\|f\|_K^2 = \displaystyle\sum_{k=1}^\infty \frac{c_k^2}{\lambda_k}$.

For every $k \geq 1$, $\dfrac{\partial H}{\partial c_k} = \dfrac{1}{m}\displaystyle\sum_{i=1}^m -2(y_i - f(x_i))\phi_k(x_i) + 2\gamma\frac{c_k}{\lambda_k}$. If f is a minimum of H, then, for each k, we must have $\frac{\partial H}{\partial c_k} = 0$ or, solving for c_k,

$$c_k = \lambda_k \sum_{i=1}^m a_i\phi_k(x_i)$$

where $a_i = \frac{y_i - f(x_i)}{\gamma m}$. Thus,

$$\begin{aligned} f(x) &= \sum_{k=1}^{\infty} c_k \phi_k(x) = \sum_{k=1}^{\infty} \lambda_k \sum_{i=1}^{m} a_i \phi_k(x_i) \phi_k(x) \\ &= \sum_{i=1}^{m} a_i \sum_{k=1}^{\infty} \lambda_k \phi_k(x_i) \phi_k(x) = \sum_{i=1}^{m} a_i K(x_i, x). \end{aligned}$$

Replacing $f(x)$ in the definition of a_i above, we obtain

$$a_i = \frac{y_i - \sum_{i=1}^{m} a_i K(x_i, x)}{\gamma m}.$$

Multiplying both sides by γm and writing the result in matrix form, we obtain $(\gamma m \text{ Id} + K[\mathbf{x}])a = \mathbf{y}$. And this system is well-posed since $K[\mathbf{x}]$ is positive and the addition of a positive matrix and the identity is strictly positive. $\square$

Proposition 8 yields an algorithm which outputs an approximation of the target function, working in the infinite dimensional function space $\mathcal{H}_K$. We won't pursue the implications of that result here, but see [17] and its references for some indications. Moreover, we have not given a bias-variance estimate based on the parameter γ. That would be useful since a good choice of γ is important in choosing an algorithm. The framework developed here suggests approaches to this problem. But it is time for us to end this modest contribution to the foundations.

Appendix A: Entropy numbers and eigenvalues

The entropy numbers of a compact operator $T : \mathbb{E} \to \mathbb{E}$ are closely related to the eigenvalues of T. If $|\lambda_1| \geq |\lambda_2| \geq \ldots$ are these eigenvalues, then $|\lambda_k| \leq \sqrt{2} e_k(T)$. This inequality is due to B. Carl and is proved, for instance, on page 512 of [24]. An inequality in the opposite direction is proved in Proposition 9 below. Related material can be found in [29].[4]

Proposition 9. *Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n \geq \ldots \geq 0$ be a sequence of real numbers. Consider the diagonal linear operator defined by*

$$\begin{aligned} L : \ell^2 &\to \ell^2 \\ (w_n) &\mapsto (\lambda_n w_n). \end{aligned}$$

If $\lambda_k \leq C k^{-\ell}$ for some C, ℓ and all $k \geq 1$, then

$$\varepsilon_k(L) \leq C_L (\ln k)^{-\ell}, \qquad \text{and, if } k \geq 2, \qquad e_k(L) \leq 2 C_L k^{-\ell}.$$

Here $C_L = 6 C \ell^{\ell}$.

[4] Many of the analysis references used for our paper deal with the case $\dim X = 1$ and that case is not useful in learning theory. Thus some care must be taken in depending on the literature. It is useful to quote Pietsch ([29], page 252) in this respect:

> [...] Moreover the situation is even worse, since these authors have very often omitted proofs claiming that they can be adapted step by step from the scalar-valued setting. Thus [...] we are not in a position to recommend any rigorous reference. On the other hand, it would be beyond the scope of this book to provide all necessary details. This section is therefore in striking contrast to the rest of the book. It presents the most beautiful applications of the abstract theory of eigenvalue distributions to integral operators, but requires a lot of blind confidence on the part of the reader. Nevertheless, I bet my mathematical reputation (but not my car!) that all the statements are correct.

Proof. By Lemma 4 below,

$$\begin{aligned} \varepsilon_k(L) &\leq 6 \sup_{n\in\mathbb{N}} k^{-\frac{1}{n}} (\lambda_1\lambda_2\dots\lambda_n)^{\frac{1}{n}} \\ &\leq 6C \sup_{n\in\mathbb{N}} k^{1/n} \left(\frac{1}{n!}\right)^{\ell/n} \leq 6C \sup_{n\in\mathbb{N}} k^{1/n} \left(\frac{e}{n}\right)^{\ell} \\ &= 6Ce^{\ell} \sup_{n\in\mathbb{N}} k^{1/n} \left(\frac{1}{n}\right)^{\ell}, \end{aligned}$$

the last by Stirling's inequality. Letting $x = n^{-\ell}$ and looking at the zero of the derivative of $f(x) = xk^{-x^{1/\ell}}$, we see that the maximum of f is reached when $x = (\ell/\ln k)^{\ell}$. Therefore, the supremum of the expression above is bounded by its value at $n = \ln k/\ell$, i.e.,

$$\varepsilon_k(L) \leq 6Ce^{\ell} k^{-\frac{\ell}{\ln k}} \left(\frac{\ell}{\ln k}\right)^{\ell} = 6C\left(\frac{\ell}{\ln k}\right)^{\ell}.$$

Moreover

$$e_k(L) = \varepsilon_{2^k-1}(L) \leq 6C\left(\frac{\ell}{\ln(2^k-1)}\right)^{\ell} \leq (12C\ell^{\ell})k^{-\ell}.$$

□

The following result is taken from [9] (see Proposition 1.3.2 there).

Lemma 4. *In hypothsesis of Proposition 9, for every $k \geq 1$,*

$$\sup_{n\in\mathbb{N}} k^{-\frac{1}{n}} (\lambda_1\lambda_2\dots\lambda_n)^{\frac{1}{n}} \leq \varepsilon_k(L) \leq 6 \sup_{n\in\mathbb{N}} k^{-\frac{1}{n}} (\lambda_1\lambda_2\dots\lambda_n)^{\frac{1}{n}}.$$

□

Appendix B: The least squares algorithm

Recall that $f_{\mathbf{z}}$ is the function minimizing in $\mathcal{H}$ the empirical error $\mathcal{E}_{\mathbf{z}}$. In Chapter I we focused on the confidence of having a small sample error $\mathcal{E}_{\mathcal{H}}(f_{\mathbf{z}})$. The problem of actually computing $f_{\mathbf{z}}$ was however ignored. We now shift our attention to that, for the case of $\mathcal{H}$ a finite dimensional full linear space.

Let $\phi_1, \dots, \phi_N$ be a basis of $\mathcal{H}$. Then, each function $f \in \mathcal{H}$ can be written in a unique way as

$$f = \sum_{i=1}^{N} w_i \phi_i$$

with $w_i \in \mathbb{R}$ for $i = 1, \dots, N$.

For a sample $\mathbf{z} \in Z^m$, $\mathbf{z} = ((x_1, y_1), \dots, (x_m, y_m))$, to minimize the empirical error $\mathcal{E}_{\mathbf{z}}$ means to find $f \in \mathcal{H}$ minimizing

$$\sum_{j=1}^{m} (f(x_j) - y_j)^2$$

where we suppose $m > N$. Thus one finds $w \in \mathbb{R}^N$ minimizing

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{N} (w_i\phi_i(x_j)) - y_j\right)^2.$$

Let $a_{ij} = \phi_i(x_j)$ and A be the $m \times N$ matrix with entries a_{ij}. Our problem —in the sequel the *least squares problem*—now becomes that of, given A and y, minimizing over $w \in W = \mathbb{R}^N$:

$$\sum_{j=1}^{m} \left(\sum_{i=1}^{N} a_{ij} w_i - y_j \right)^2 = \sum_{j=1}^{m} ((Aw)_j - y_j)^2 = \|Aw - y\|^2.$$

Note that in our situation, since $m > N$, the system $Aw = y$ is likely to have no solutions. A point w minimizing $\|Aw - y\|^2$ is called a *least squares solution.*

The idea to "solve" an overdetermined system of equations $Aw = y$ by finding a point minimizing $\|Aw - y\|^2$ goes back to Gauss and Legendre.[5] The motivation was to find a function fitting a certain amount of astronomical data. The y values of these data were obtained by measurements and thus contaminated with small errors. Laplace had suggested minimizing $\sum_{j=1}^{m} |(Aw)_j - y_j|$ with the additional restriction $\sum_{j=1}^{m} ((Aw)_j - y_j) = 0$, and he had proved that the solution w thus found satisfied n of the m equalities in $Aw = y$. But Gauss argued that such a solution was not consistent with the laws of probability since greater or smaller errors are equally probable in all of the m equations. Additionally, Gauss proved that, contrary to Laplace's suggestion, the least squares solution enjoys remarkable statistical properties (cf. Theorem 6 below).

Let's now discuss how to find a w minimizing $\|Aw - y\|^2$. By abuse of notation let's denote also by A the linear map from $\mathbb{R}^N$ to $\mathbb{R}^m$ whose matrix is A. Let $\mathrm{Im}(A) \subset \mathbb{R}^m$ be the image of A, and $\mathbf{c} \in \mathrm{Im}(A)$ be the point whose distance to y is minimal. Then

$$S = \{w \in \mathbb{R}^N \mid Aw = c\}$$

is an affine subspace of $\mathbb{R}^N$ of dimension $N - \dim(\ker(A))$. In particular, the least squares problem has a unique solution $w \in \mathbb{R}^N$ if and only if A is injective. The next result is immediate.

Proposition 10. *Let $A : \mathbb{R}^N \to \mathbb{R}^m$ be injective and $y \in \mathbb{R}^m$. Then the solution of the least squares problem is given by $w = A^\dagger y$ where $A^\dagger = (A_{|\,\mathrm{Im}(A)})^{-1}\pi$. Here $\pi : \mathbb{R}^m \to \mathrm{Im}(A)$ is the orthogonal projection onto $\mathrm{Im}(A)$.* □

Recall that the orthogonal complement to $\mathrm{Im}(A)$ in $\mathbb{R}^m$ is $\ker(A^*)$ where A^* denotes the adjoint of A. Thus, for every $w \in \mathbb{R}^N$,

$$\begin{aligned} w = A^\dagger y \quad &\Longleftrightarrow \quad Aw = \pi y \iff Aw - \pi y \in \mathrm{Im}(A)^\perp \\ &\Longleftrightarrow \quad A^*(Aw - y) = 0 \iff w = (A^*A)^{-1}A^*y. \end{aligned}$$

The map $A^\dagger = (A^*A)^{-1}A^*$ is called the *Moore-Penrose inverse* of the injective map A. So, we have shown that $w = A^\dagger y$. In particular, w is a linear function of y. For the rest of this section, assume A is injective.

To compute w the main algorithmic step is to solve the linear system $Sw = b$ with $S = A^*A$ and $b = A^*y$. The field of Numerical Linear Algebra provides us with an important collection of algorithms for doing so as well as results about their complexity and stability properties.

---

[5]In *Nouvelle méthodes pour la determination des orbites des comètes*, published in 1805, Legendre writes "Of all the principles that can be proposed [...], I think that there is none more general, more exact, and more easy to apply, than that consisting of minimizing the sum of the squares of the errors."

Perturbation results for least squares follow from perturbation theory for linear equation solving. Recall that the *condition number* of A is defined to be $\kappa(A) = \|A\|\|A^\dagger\|$. A proof of the following can be found in [7].

Theorem 5. *Let A be an injective $m \times N$ matrix, $y \in \mathbb{R}^m$ and $w = A^\dagger y$. Let δA be an $m \times N$ matrix such that* $\operatorname{rank}(A + \delta A) = \operatorname{rank}(A)$ *and let $\delta y \in \mathbb{R}^m$. Suppose $\varepsilon_A, \varepsilon_y \geq 0$ such that*

$$\frac{\|\delta A\|}{\|A\|} \leq \varepsilon_A \qquad \text{and} \qquad \frac{\|\delta y\|}{\|y\|} \leq \varepsilon_y. \tag{11}$$

Define $\delta w = (A + \delta A)^\dagger (y + \delta y) - w$.

If $\kappa(A)\varepsilon_A < 1$, then

(12)

$$\|\delta w\| \leq \frac{\kappa(A)}{1 - \kappa(A)\varepsilon_A} \left(\varepsilon_A \|w\| + \varepsilon_y \frac{\|y\|}{\|A\|} + \varepsilon_A \kappa(A) \frac{\|y - Aw\|}{\|A\|} \right) + \varepsilon_A \kappa(A) \|w\|.$$

□

Thus, Theorem 5 says that if A and y have relative errors bounded as in (11), then the error in the solution of the least squares problem is given by (12). We note the role of the condition number of A in this estimate.

If A is not injective, one can find a solution $w \in S$ by considering a maximal rank restriction of A and solving the problem for this restriction.

Before finishing this section we state Gauss' result on a statistical property of least squares. First some definitions.

Definition 1. Let Y be a probability space and $\mathbf{y} = (y_1, \ldots, y_m) : Y \to \mathbb{R}^m$ be a random vector. A function $g : \mathbb{R}^m \to \mathbb{R}^N$ is an *unbiased estimate* of a vector $v \in \mathbb{R}^N$ if $\mathbf{E}(g(\mathbf{y})) = v$.

We say that g^* is a *minimum variance estimate* (in a class C of functions from $\mathbb{R}^m$ to $\mathbb{R}^N$) of v if $\mathbf{E}(g(\mathbf{y})) = v$ and $\sum_{i=1}^N \sigma^2(g_i(\mathbf{y}))$ is minimized over all the functions $g \in C$.

Theorem 6 (Gauss). *Let $A \in \mathbb{R}^{m \times N}$ be injective, $y^* \in \mathbb{R}^m$, and $w^* \in \mathbb{R}^N$ such that $Aw^* = y^*$. Consider the random vector $\mathbf{y}$ such that, for $j = 1, \ldots, m$, $y_j = y_j^* + \varepsilon$ where ε is a random variable with mean 0 and variance σ^2. The minimum variance estimate of w^* in the class of all unbiased linear estimators is $w = A^\dagger \mathbf{y}$, i.e. the least squares solution of $Aw = \mathbf{y}$.*

In our case, Gauss' Theorem would say that if for every $x \in X$ the probability measures $\rho(y|x)$ are identical, then the following holds. Let $w^* \in \mathbb{R}^N$ such that

$$f_{\mathcal{H}} = \sum_{i=1}^N w_i^* \phi_i.$$

For all samples $\mathbf{z} \in Z^m$, the least squares solution w of

$$\sum_{j=1}^m (f_w(x_j) - y_j)^2$$

is the one minimizing the variance (in the sense of Definition 1) among all linear maps $g : \mathbb{R}^m \to \mathbb{R}^N$ such that, for $i = 1, \ldots, N$,

$$\int_{\mathbf{y} \in Y^m} g_i(\mathbf{y}) = w_i^*.$$

Generalizations of Gauss' Theorem (among many other results on least squares) can be found in [7]. See also [13].

Remark 6. This paper can be thought of as a contribution to the solution of Problem 18 in [37].

Index

References

[1] L.V. Ahlfors, *Complex analysis*, 3rd ed., McGraw-Hill, 1978. MR **80c**:30001

[2] N. Aronszajn, *Theory of reproducing kernels*, Transactions of the Amer. Math. Soc.**68** (1950), 337–404. MR **14**:479c

[3] A.R. Barron, *Complexity regularization with applications to artificial neural networks*, Nonparametric Functional Estimation (G. Roussa, ed.), Kluwer, Dordrecht, 1990, pp. 561–576. MR **93b**:62052

[4] J. Bergh and J. Löfström, *Interpolation spaces. an introduction*, Springer-Verlag, 1976. MR **58**:2349

[5] M.S. Birman and M.Z. Solomyak, *Piecewise polynomial approximations of functions of the classes* W_p^α, Mat. Sb. **73** (1967), 331–355; English translation in *Math. USSR Sb.* (1967), 295–317. MR **36**:576

[6] C.M. Bishop, *Neural networks for pattern recognition*, Cambridge University Press, 1995. MR **97m**:68172

[7] A. Björck, *Numerical methods for least squares problems*, SIAM, 1996. MR **97g**:65004

[8] L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and real computation*, Springer-Verlag, 1998. MR **99a**:68070

[9] B. Carl and I. Stephani, *Entropy, compactness and the approximation of operators*, Cambridge University Press, 1990. MR **92e**:47002

[10] P. Craven and G. Wahba, *Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation*, Numer. Math. **31** (1979), 377–403. MR **81g**:65018

[11] C. Darken, M. Donahue, L. Gurvits, and E. Sontag, *Rates of convex approximation in non-Hilbert spaces*, Construct. Approx. **13** (1997), 187–220. MR **98c**:41051

[12] L. Debnath and P. Mikusiński, *Introduction to Hilbert spaces with applications*, 2nd ed., Academic Press, 1999. MR **99k**:46001

[13] J.-P. Dedieu and M. Shub, *Newton's method for overdetermined systems of equations*, Mathematics of Computation **69** (2000), 1099–1115. MR **2000j**:65133

[14] R.A. DeVore and G.G. Lorentz, *Constructive approximation*, Grundlehren der mathematischen Wissenschaften, vol. 303, Springer-Verlag, 1993. MR **95f**:41001

[15] J. Duchon, Spline minimizing rotation-invariant semi-norms in Sobolev spaces, *Constructive theory of functions on several variables* (W. Schempp and K. Zeller, eds.), Lecture Notes in Math. **571**, Springer-Verlag, Berlin, 1977. MR **58**:12146

[16] D.E. Edmunds and H. Triebel, *Function spaces, entropy numbers, differential operators*, Cambridge University Press, 1996. MR **97h**:46045

[17] T. Evgeniou, M. Pontil, and T. Poggio, *Regularization Networks and Support Vector Machines*, Advances in Computational Mathematics **13** (2000), 1–50. MR **2001f**:68053

[18] D. Haussler, *Decision theoretic generalizations of the PAC model for neural net and other learning applications*, Information and Computation **100** (1992), 78–150. MR **93i**:68149

[19] H. Hochstadt, *Integral equations*, John Wiley & Sons, 1973. MR **52**:11503

[20] A.N. Kolmogorov and S.V. Fomin, *Introductory real analysis*, Dover Publications Inc., 1975. MR **51**:13617

[21] A.N. Kolmogorov and V.M. Tikhomirov, *ε-entropy and ε-capacity of sets in function spaces*, Uspecki **14** (1959), 3–86. MR **22**:2890

[22] W.-S. Lee, P. Bartlett, and R. Williamson, *The importance of convexity in learning with squared loss*, IEEE Transactions on Information Theory **44** (1998), 1974–1980. MR **99k**:68160

[23] P. Li and S.-T. Yau, *On the parabolic kernel of the Schrödinger operator*, Acta Math. **156** (1986), 153–201. MR **87f**:58156

[24] G.G. Lorentz, M. Golitschek, and Y. Makovoz, *Constructive approximation; advanced problems*, Springer-Verlag, 1996. MR **97k**:41002

[25] W.S. McCulloch and W. Pitts, *A logical calculus of the ideas immanent in nervous activity*, Bulletin of Mathematical Biophysics **5** (1943), 115–133. MR **6**:12a

[26] J. Meinguet, *Multivariate interpolation at arbitrary points made simple*, J. Appl. Math. Phys. **30** (1979), 292–304. MR **81e**:41014

[27] M.L. Minsky and S.A. Papert, *Perceptrons*, MIT Press, 1969.

[28] P. Niyogi, *The informational complexity of learning*, Kluwer Academic Publishers, 1998.

[29] A. Pietsch, *Eigenvalues and s-numbers*, Cambridge University Press, 1987. MR **88j**:47022b

[30] A. Pinkus, *N-widths in approximation theory*, Springer-Verlag, New York, 1986. MR **86k**:41001
[31] T. Poggio and C.R. Shelton, *Machine learning, machine vision, and the brain*, AI Magazine **20** (1999), 37–55.
[32] D. Pollard, *Convergence of stochastic processes*, Springer-Verlag, 1984. MR **86i**:60074
[33] G.V. Rozenblum, M.A. Shubin, and M.Z. Solomyak, *Partial differential equations vii: Spectral theory of differential operators*, Encyclopaedia of Mathematical Sciences, vol. 64, Springer-Verlag, 1994. MR **95j**:35156
[34] I.J. Schoenberg, *Metric spaces and completely monotone functions*, Ann. of Math. **39** (1938), 811–841.
[35] I.R. Shafarevich, *Basic algebraic geometry*, 2nd ed., vol. 1: Varieties in Projective Space, Springer-Verlag, 1994. MR **95m**:14001
[36] S. Smale, *On the Morse index theorem*, J. Math. and Mech. **14** (1965), 1049–1056, With a Corrigendum in J. Math. and Mech. **16**, 1069–1070, (1967). MR **31**:6251; MR **34**:5108
[37] ______, *Mathematical problems for the next century*, Mathematics: Frontiers and Perspectives (V. Arnold, M. Atiyah, P. Lax, and B. Mazur, eds.), AMS, 2000, pp. 271–294. CMP 2000:13
[38] S. Smale and D.-X. Zhou, *Estimating the approximation error in learning theory*, Preprint, 2001.
[39] M.E. Taylor, *Partial differential equations i: Basic theory*, Applied Mathematical Sciences, vol. 115, Springer-Verlag, 1996. MR **98b**:35002b
[40] L.G. Valiant, *A theory of the learnable*, Communications of the ACM**27** (1984), 1134–1142.
[41] S. van de Geer, *Empirical processes in m-estimation*, Cambridge University Press, 2000.
[42] V. Vapnik, *Statistical learning theory*, John Wiley & Sons, 1998. MR **99h**:62052
[43] P. Venuvinod, *Intelligent production machines: benefiting from synergy amongst modelling, sensing and learning*, Intelligent Production Machines: Myths and Realities, CRC Press LLC, 2000, pp. 215–252.
[44] A.G. Vitushkin, *Estimation of the complexity of the tabulation problem*, Nauka (in Russian), 1959, English Translation appeared as *Theory of the Transmission and Processing of the Information*, Pergamon Press, 1961.
[45] G. Wahba, *Spline models for observational data*, SIAM, 1990. MR **91g**:62028
[46] R. Williamson, A. Smola, and B. Schölkopf, *Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators*, Tech. Report NC2-TR-1998-019, NeuroCOLT2, 1998.

Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

E-mail address: `macucker@math.cityu.edu.hk`

Department of Mathematics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

Current address, S. Smale: Department of Mathematics, University of California, Berkeley, California 94720

E-mail address: `masmale@math.cityu.edu.hk, smale@math.berkeley.edu`