Statistical learning theory : a primer

Louis Wehenkel
University of Liège - Institut Montefiore
Department of Electrical Engineering and Computer Science
Email : L.Wehenkel@uliege.be

February 1, 2018

Abstract

The purpose of this document is to provide a first insight into the main results from statistical learning theory. On the way of our exploration we will also make links with classical notions from estimation theory.

Statistical learning theory aims at extending the principles of non-parametric statistics to the problem of estimating functions from input/output examples. There are mainly two results: the first concerns the conditions under which the so-called empirical risk minimization principle remains consistent; the other provides confidence intervals on the actual risk depending on the complexity of the function space used for learning.

Contents

1 Framework **3**

1.1 Data generation model 3

1.1.1 Illustration 4

1.2 Learning model 5

1.3 Examples 6

1.3.1 Trivial learning problem: $\mathcal{F}$ is a singleton 6

1.3.2 A simple, though non trivial case: $\mathcal{F}$ has finite cardinality 6

1.4 Interpretation in the space of random variables 10

1.5 Geometric interpretation and orthogonality (*) 11

1.5.1 Hilbert spaces 11

2 Main theorems of statistical learning theory **12**

2.1 Consistency of the ERM principle 13

2.1.1 Uniform convergence theorems 14

2.1.2 Consistency of the ERM principle over finite hypothesis spaces 15

2.1.3 VC-Entropy of a set of functions 15

2.1.4 Interpretation for binary classification problems 17

2.1.5 Comments 18

2.2 Distribution-independent fast asymptotic rate of convergence of the ERM principle for binary classification problems 18

2.3 VC dimension of a set of indicator functions 19

2.3.1 Examples of VC-dimensions for indicator functions 20

2.3.2 VC-dimension for regression functions 22

2.4 Bounds on the generalization ability of learning machines 23

2.4.1 Comments 25

2.4.2 Practical error bounds (for binary classification problems) 25

2.4.3 Regression problems 26

2.5 Structural risk minimization principle 26

2.5.1 Defining admissible structures 26

3 Concluding remarks **27**

1 Framework

The theoretical framework at which we are looking in this document is composed of two components: a *data generation model* and a *learning model*.

1.1 Data generation model

Our data generation model is intuitively described as follows (see also figure 1).

1. An environment viewed as a random generator of input vectors x in some input space $\mathcal{X}$ (we will focus on Euclidean input spaces in what follows) according to a probability measure $P_{\mathcal{X}}(\cdot)$ defined on $\mathcal{X}$;

2. A stochastic system viewed as a device reacting in a specific way to the environment by producing outputs y belonging to some output space $\mathcal{Y}$, modeled by the conditional probability law $P_{\mathcal{Y}|\mathcal{X}}(\cdot|\cdot)$ defined on $\mathcal{Y}$ for (almost) every x in $\mathcal{X}$;

3. Neither $P_{\mathcal{X}}(\cdot)$ nor $P_{\mathcal{Y}|\mathcal{X}}(\cdot|\cdot)$ are known.

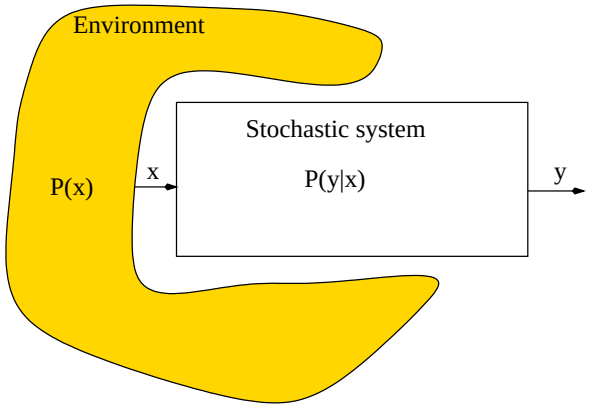

Figure 1: Data generation model : intuitive view

A more formal way to look at our problem is as follows (see figure 2)

1. First define a probability space $(\Omega, \mathcal{E}, P_{\Omega}(\cdot))$, where Ω denotes the sample space, $\mathcal{E}$ a σ-algebra of events, and $P_{\Omega}(\cdot)$ a probability measure defined on all events of $\mathcal{E}$,

2. Then define two measurable functions on Ω [1]

$$\mathcal{X}(\cdot) \quad : \quad \omega \in \Omega \rightarrow \mathcal{X}(\omega) \in \mathcal{X} \tag{1}$$
$$\mathcal{Y}(\cdot) \quad : \quad \omega \in \Omega \rightarrow \mathcal{Y}(\omega) \in \mathcal{Y}. \tag{2}$$

3. The probability measure $P_{\mathcal{X},\mathcal{Y}}(\cdot)$ on $\mathcal{X} \times \mathcal{Y}$, as well as the two probability measures $P_{\mathcal{X}}(\cdot)$ and $P_{\mathcal{Y}|\mathcal{X}}(\cdot|\cdot)$, are then "inherited" automatically from the assumed $P_{\Omega}(\cdot)$.

[1] We leave implicit here the definition of σ-algebras on $\mathcal{X}$, $\mathcal{Y}$, and $\mathcal{X} \times \mathcal{Y}$.

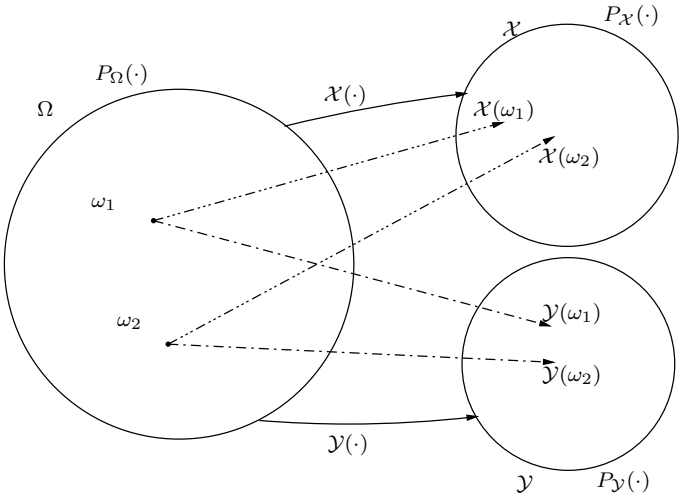

Figure 2: Data generation model : formal view

1.1.1 Illustration

Assuming that Ω is finite makes the interpretations above simpler, without imposing any deep restriction from the conceptual point of view.

Indeed, let us assume that $\Omega = \{\omega_1, \ldots, \omega_n\}$. In this case the following happens

- $P_\Omega(\cdot)$ is entirely defined by the n numbers

$$p_i = P_\Omega(\{\omega_i\}); i = 1, 2, \ldots, n. \tag{3}$$

- A random variable (say $\mathcal{X}(\cdot)$) is entirely defined by the n values

$$X_i = \mathcal{X}(\omega_i); i = 1, 2, \ldots, n. \tag{4}$$

- Given a subset $A \subset \mathcal{X}$ we have

$$P_{\mathcal{X}}(A) = P_\Omega(\{\omega \ : \mathcal{X}(\omega) \in A\}) \tag{5}$$

  or equivalently

$$P_{\mathcal{X}}(A) = \sum_{\{i \ : \mathcal{X}(\omega_i) \in A\}} p_i. \tag{6}$$

In other words, a random variable defined from Ω into $\mathcal{X}$ is nothing more than a vector (n-tuple) chosen in $\mathcal{X}^n$. In particular, if we focus on real-valued random variables (as we will do in the sequel, most of the time), we are actually looking at vectors from $\mathbb{R}^n$. Clearly, finiteness of Ω does not imply finiteness of the set of random variables which may be defined on Ω.

Carrying our analysis of the finite case a little further, let us have a look at the notion of conditional distributions. More precisely, let us suppose that $x \in \mathcal{X}$ suchthat the set $\mathcal{X}^{-1}(\{x\}) = \{\omega_i \ : \mathcal{X}(\omega_i) = x\}$ has a non-zero probability, which means that the value x is "probabilistically" relevant.

Then the observation $\mathcal{X}(\omega) = x$ induces a conditional probability distribution on Ω defined by the n numbers

$$p_{i|x} = P_{\Omega|x}(\{\omega_i\}) \triangleq \frac{P_\Omega(\{\omega_i\} \cap \mathcal{X}^{-1}(\{x\}))}{P_\Omega(\mathcal{X}^{-1}(\{x\}))}; i = 1, 2, \ldots n, \tag{7}$$

which means in very simple words that, the p_i are replaced by zero for those elements of Ω which are incompatible with the observation (i.e. for which $\mathcal{X}(\omega) \neq x$) while the others are kept unchanged except that they are normalized so as to yield a total probability of one for Ω.

$$p_{i|x} = \begin{cases} 0 & \text{if} \quad \mathcal{X}(\omega_i) \neq x \\ \frac{p_i}{\sum_{\{i \,:\mathcal{X}(\omega_i)=x\}} p_i} & \text{if} \quad \mathcal{X}(\omega_i) = x \end{cases} . \tag{8}$$

Then, given this new probability assignment on Ω the conditional probability distributions for each random variable defined on Ω are again inherited in a straightforward manner. In particular, the new probability distribution on $\mathcal{X}$ will be such that the value x has a probability mass of one and all other values a probability mass of zero. However, for other random variables defined on Ω the conditional distribution does not necessarily reduce to this "uncertain" case, at least if the set $\{i \,:\mathcal{X}(\omega_i) = x\}$ does not reduce to a singleton.

We will use this simple "finite" case as our running example in what proceeds.

1.2 Learning model

The learning model is composed of the following

1. A hypothesis space $\mathcal{F}$ of functions $f(x, \lambda)$ where λ varies in some set of parameters Λ. Each such function maps $\mathcal{X}$ on $\mathcal{Y}$ in some way dependent on the parameter λ.

2. A loss function defined on $\mathcal{Y}$ which measures how close two values in $\mathcal{Y}$ are

$$\ell(\cdot, \cdot) \,: \mathcal{Y} \times \mathcal{Y} \to [0; \infty[, \tag{9}$$

   suchthat $\ell(y, y')$ measures the distance (or dissimilarity) of y and y'.[2]

3. A learning sample, namely a set[3] of pairs $LS = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N$, which are supposed to be generated independently and identically distributed according to $P_{\mathcal{X},\mathcal{Y}}$.[4] Thus the probability to obtain a particular learning sample is obtained by

$$Pr(LS) = \prod_{i=1}^{N} P_{\mathcal{X},\mathcal{Y}}(x_i, y_i), \tag{10}$$

4. A learning algorithm which chooses, according to some procedure, a function in $\mathcal{F}$ as close as possible to $\mathcal{Y}(\cdot)$, but using as input only the learning sample LS. In other words, a sequence of rules

$$\lambda_N(\cdot) \,: (\mathcal{X} \times \mathcal{Y})^N \to \Lambda \tag{11}$$

   for selecting a value of λ given a sample of size N.

[2] The precise definition of this kind of function depends on the structure (symbolic or numerical) of $\mathcal{Y}$. E.g. in the case of regression problems $\mathcal{Y} \subset \mathbb{R}$ and we typically use the square loss: $\ell(y, y') = (y - y')^2$.

[3] Strictly speaking the LS is an N-tuple and not a set.

[4] As if we had observed N successive occurrences of ω drawn according to P_Ω and had recorded only the corresponding values of x and y.

Clearly, the objective of a learning algorithm is to produce functions which are close to the objective $\mathcal{Y}$. Therefore, we define the actual risk associated to a function in $\mathcal{F}$ by[5]

$$R(f(\cdot,\lambda)) \stackrel{\triangle}{=} E\{\ell(f(\mathcal{X},\lambda),\mathcal{Y})\} = \sum_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \ell(f(x,\lambda),y) P_{\mathcal{X},\mathcal{Y}}(x,y), \tag{12}$$

and our objective - if it is reachable - would be to design a "universal" learning algorithm which will always find a function which minimizes the expected risk.

Because the risk function in most practical cases defines a distance (in the mathematical sense) among random variables, we will use in the sequel the terms *risk* and *distance* in an interchangeable manner. In the sequel we will also use the terms *expected* or *actual* risk in order to stress the difference with the *empirical* risk defined in the next sections.

1.3 Examples

Before discussing more the above learning problem let us look at some examples.

1.3.1 Trivial learning problem: $\mathcal{F}$ is a singleton

In this problem, the function space reduces to a singleton, i.e. $\mathcal{F}$ contains only one single function. It is of course easy to see that any learning algorithm (i.e. any algorithm which behaves as specified above) is perfectly optimal, in the sense that it must always produce the optimal function in $\mathcal{F}$, since there is only one possible choice.

Thus, if the set of candidate function $\mathcal{F}$ is extremely small, the learning problem is extremely easy.

1.3.2 A simple, though non trivial case: $\mathcal{F}$ has finite cardinality

Let us see now what happens when the hypothesis space becomes larger, while still being of finite cardinality. Let us suppose that $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$. Thus the learning problem is to select one of these m functions on the basis of the LS.

Let us furthermore suppose that the random variable which we want to guess is binary : $\mathcal{Y} = \{0,1\}$; hence we will assume also that the values taken by each one of the candidate functions are binary[6], i.e.

$$\forall\omega\in\Omega\ ,\forall i\leq m : f_i(\mathcal{X}(\omega))\in\{0,1\}. \tag{13}$$

Thus $\mathcal{Y}$ as well as each one of the candidate functions induces a binary partition on Ω and our objective is (naturally) to find a function $f_*\in\mathcal{F}$ such that the binary partition it induces on Ω is as close as possible to the one induced by $\mathcal{Y}$. Hence, we will use as loss function the discrete distance defined by

$$\ell(y,y') = 1-\delta_{y,y'} = \left\{\begin{array}{lll} 0 & \text{if} & y=y' \\ 1 & \text{if} & y\neq y' \end{array}\right. . \tag{14}$$

It follows from (12) that

$$R(f_i) = P_\Omega[f_i(\mathcal{X}(\omega))\neq\mathcal{Y}(\omega)], \tag{15}$$

[5] $E\{\cdot\}$ denotes the "expectation" operator on $(\Omega,\mathcal{E},P_\Omega)$.

[6] We will also use the term "indicator function" to denote such binary valued functions.

i.e. it is the probability that $f_i(\mathcal{X})$ makes a prediction that is different from $\mathcal{Y}$, i.e. the *true error rate* of f_i.

Now let us denote by $R_e^N(f_i)$ the empirical risk of f_i computed on a learning sample of size N, namely

$$R_e^N(f_i) = \frac{1}{N}\sum_{j=1}^{N} \ell(f_i(x_j), y_j). \tag{16}$$

In other words $R(f_i)$ is the expected value of the loss function and $R_e^N(f_i)$ is the sample average value of the loss function; in the present context, the latter is equal to the number of misclassifications over the LS divided by the size N of this sample, i.e. the *resubstitution error rate*.

Introducing the empirical risk minimization (ERM) principle. The empirical risk minimization principle suggests that a learning algorithm should select a function in $\mathcal{F}$ which minimizes the empirical risk R_e^N. Let us denote by f_*^N a function so selected by this empirical risk minimization strategy given a learning sample of size N. In the case of a finite set $\mathcal{F}$, this amounts to

$$f_*^N \in \arg\min_{f\in\mathcal{F}} R_e^N(f), \tag{17}$$

where ties are broken arbitrarily if more than one element of $\mathcal{F}$ realizes the minimum.[7]

Similarly, let us denote by f^* a function minimizing the actual risk, i.e.

$$f_* = \arg\min_{f\in\mathcal{F}} R(f), \tag{18}$$

and by R_* the corresponding minimal value of R, i.e.

$$R_* = \min_{f\in\mathcal{F}} R(f). \tag{19}$$

Note that in general it is not necessarily possible to realize this value in $\mathcal{F}$, i.e. it is not necessarily possible to prove that there exists a function $f_* \in \mathcal{F}$ such that[8]

$$R(f_*) = R_*. \tag{20}$$

In other cases, their may be several functions in $\mathcal{F}$ which realize this infimum. However, when $\mathcal{F}$ is finite, as in the present section, their exists always at least one element of $\mathcal{F}$ that satisfies (20).

The ERM principle will be further studied in depth below. But, let us see what we can say about this strategy in this simple example. First of all, we notice that since the learning sample LS is randomly generated, $R_e^N(f_i)$, is a random variable, and hence the function selected by the ERM principle will also be random to a certain extent. Thus, it seems that we have lost the possibility of making interesting "deterministic" statements about our learning algorithm.

[7] In the context of a finite set $\mathcal{F}$, it is in principle easy to implement the ERM principle.

[8] The theory which is developed below however does not depend on this latter possibility (it is valid even if it is not possible to realize the minimal risk in $\mathcal{F}$). However, in our intuitive discussion we will assume that such a function f_* indeed exists. Therefore, we will use the notation $R(f_*)$ instead of R_*, unless we want to stress the difference between these two concepts.

This is a general fact in automatic learning : as soon as the learning problem is non trivial only probabilistic claims may be formulated about properties of any learning algorithm. This is also a general characteristic of inductive reasoning, which can not be proven to be "correct" in the case of finite evidence. Nevertheless, it is possible to study the behaviour of our learning algorithms by exploiting the laws of large numbers.

Law of large numbers and the question of consistency of the ERM principle. Let us denote by $R_e^N(f_i)$ the random variable corresponding to the empirical risk of function f_i computed on the basis of a random learning samples of size N generated by our data generation model.

Thus, for $N = 1, 2, \ldots$ this defines a sequence of random variables for each function $f_i \in \mathcal{F}$. For any such function $f_i \in \mathcal{F}$, the (weak) law of large numbers guaranties that the sequence $R_e^N(f_i)$ converges in probability to $R(f_i)$. More precisely, this is expressed by

$$\forall i, \epsilon, \eta > 0 : \exists N_i(\epsilon, \eta) \left| \left[N \geq N_i(\epsilon, \eta) \Rightarrow Pr\left\{ |R_e^N(f_i) - R(f_i)| > \epsilon \right\} < \eta \right] . \tag{21}$$

When $\mathcal{F}$ is *finite*, it is therefore also true that

$$\forall \epsilon, \eta > 0 \ : \exists N_0(\epsilon, \eta) \left| \left[N \geq N_0(\epsilon, \eta) \Rightarrow (\forall f_i \in \mathcal{F} : Pr\left\{ |R_e^N(f_i) - R(f_i)| > \epsilon \right\} < \eta) \right] , \tag{22}$$

where $N_0(\epsilon, \eta) = \max\{N_1(\epsilon, \eta), \ldots, N_m(\epsilon, \eta)\}$, and is independent of the function f_i.

Intuitively, this means that for large enough learning samples the empirical risk of any function selected by any kind of learning algorithm using a finite set of candidate functions $\mathcal{F}$, will indeed reflect the actual (expected risk) of this function with high probability and high accuracy.[9]

Notice, however, that the above does not directly (i.e. without further arguments) tell us anything about the fact that, looking at the random sequence of models $f_*^N \in \mathcal{F}$ generated by the ERM principle, we should have

$$\forall \epsilon, \eta > 0, \exists N_1 \ : N \geq N_1 \Rightarrow Pr\left\{ |R(f_*^N) - R(f_*)| > \epsilon \right\} < \eta, \tag{24}$$

or

$$\forall \epsilon, \eta > 0, \exists N_1 \ : N \geq N_1 \Rightarrow Pr\left\{ |R_e^N(f_*^N) - R(f_*)| > \epsilon \right\} < \eta. \tag{25}$$

These two properties, if satisfied by a learning algorithm, are called **consistency properties**, and it is precisely the objective of statistical learning theory to state conditions under which this is indeed the fact. We wait until section 2 for these statements. Before that, let us carry our example a little further, and let us see what we can infer in the finite sample case in terms of error bounds.

[9] We further notice that in any case have

$$\forall \epsilon, \eta > 0, \exists N_0(\epsilon, \eta) \left| \left[N \geq N_0(\epsilon, \eta) \Rightarrow Pr\left\{ |R_e^N(f_*) - R(f_*)| > \epsilon \right\} < \eta \right] , \tag{23}$$

where me merely apply (22) to the best possible function $f_* \in \mathcal{F}$.

Finite sample error bounds. Let us see what we can say in the finite sample case. Therefore we need a stronger formulation of the law of large numbers. This is provided by the following inequality (known as Bernoulli's inequality [Ber93])

$$Pr\{|R_e^N(f_i) - R(f_i)| > \epsilon\} \leq 2e^{-N\epsilon^2/2}. \tag{26}$$

Notice that this condition actually allows one to prove condition (21).

From this condition, we can infer a stronger condition concerning the whole set of functions $\mathcal{F}$. Indeed, let us denote by $A_i (i = 1, \ldots, m)$ the event that $|R_e^N(f_i) - R(f_i)| > \epsilon$. Then from the fact that $Pr(A_i) \leq \alpha$ we can infer directly that

$$Pr(A) = Pr(\bigcup_{i=1}^{m} A_i) \leq \sum_{i=1}^{m} Pr(A_i) \leq m\alpha, \tag{27}$$

in other words

$$Pr\{(\exists f_i \in \mathcal{F}) : |R_e^N(f_i) - R(f_i)| > \epsilon\} \leq 2me^{-N\epsilon^2/2} (= 2e^{\frac{-N\epsilon^2 + 2\ln m}{2}}). \tag{28}$$

This latter condition is very general, in the sense that it may be applied to any model selection strategy. It means that, whatever the principle or algorithm that you use to select a function in a finite set of candidate function $\mathcal{F}$, you can infer something about the expected risk of this function from its empirical risk. It therefore makes again sense to select the function minimizing the empirical risk. However, it shows explicitly how the size of the hypothesis space $\mathcal{F}$ influences your bounds.

Of course, the bounds might be refined in order to take into account not only the number of candidate hypotheses, but also their diversity. For example if several functions in $\mathcal{F}$ classify the learning set in a very similar way, then the sets A_i would overlap significantly and the error bound would be quite conservative. In addition, if the hypothesis space is infinite (as in many applications) the above reasoning is not applicable anymore. The introduction of a measure of the "expressiveness" of the hypothesis space $\mathcal{F}$ is therefore the first part of the work to be done in the section 2. However, intuitively, the larger the hypothesis space the easier it will be to find a function with a small empirical risk; but, as the size increases, the right member of (28) will increase. Or, if the right member of (28) is kept constant, the value of ϵ must increase, leading to a large confidence interval, and hence less precision about the statement concerning the actual risk.

Considering the influences of m, N and ϵ we see that they are combined according to formula

$$-N\epsilon^2 + 2\ln m \tag{29}$$

which represents at least qualitatively the typical influences

- N : the larger the sample size w.r.t. ϵ^2 the tighter the bound
- ϵ^2 : the smaller the (squared) width of the confidence interval, the larger the required sample size
- $\ln m$: the larger the "entropy" of the hypothesis space $\mathcal{F}$, the larger the required sample size required to satisfy a given confidence bound ϵ.

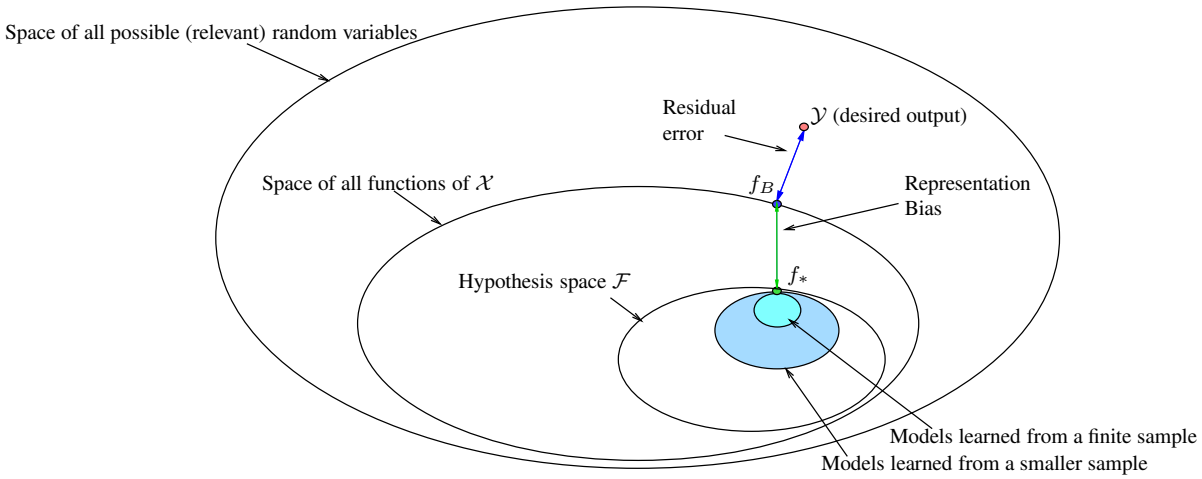

Figure 3: Learning problem in the space of random variables

The main conclusions we can infer from our examples : (i) as soon as the learning problem is non trivial the guaranties we seek must be formulated in probabilistic terms; (ii) at least for finite hypothesis spaces, probability theory allows us to find error bounds which depend on the size of the hypothesis space itself, and will be weaker when this size is big compared to the sample size. The basic contribution of V. Vapnik was to characterize the learning problems for infinite size hypothesis spaces, as we will see in section 2.

1.4 Interpretation in the space of random variables

Anticipating on the precise statement of the theorems of statistical learning theory let us summarize what we have learned up to now.

Figure 3 represents the situation in the space of all random variables from Ω into the set of possible output values $\mathcal{Y}$. Notice that the space of all functions of the input $\mathcal{X}$ does not necessarily contain the goal $\mathcal{Y}$. Furthermore, the hypothesis space $\mathcal{F}$ used by a particular learning algorithm (say the space of all neural networks with one hidden layer and up to 10 neurons in this layer), is in general a small subset of the latter.

Figure 3 allows us to discuss the following concepts

Bayes model f_B. By definition this is the best (or a best, if there are several) model possible : it is the input/output function which truly minimizes the Risk. Under few restrictive assumptions this function indeed exists and may be defined in a point-wise fashion by

$$f_B(x) = \arg\min_{y'} \{E_{y|x}\{\ell(y, y')\}. \tag{30}$$

In particular, if ℓ is the square error criterion (y being real-valued), the Bayes model is identical to the conditional expectation. If y is binary and the discrete loss function is used, the Bayes model corresponds to the locally most probable output value (given the value of $\mathcal{X}$).

The Bayes model does only depend on the chosen loss function and on the conditional probability distribution $P_{\mathcal{Y}|\mathcal{X}}$. It is thus a joint property of the system and of the used loss function ℓ. The Risk of the Bayes model, also called *residual error*, depends also and on the environment via $P_{\mathcal{X}}$.

Optimal hypothesis f_*. This is the target of the learning algorithm. Since the hypothesis space[10] generally does not "contain" the Bayes model, the optimal model is different from the latter. The risk of the optimal model is also sometimes called the representation bias, but in what follows we call representation bias the distance between the optimal model and the Bayes model.

Learned models. Figure 3 also represents the typical behavior of learning algorithms : for a given sample size they will produce a set of possible models which are random in nature and are distributed around some average model. In the statistical literature the error of this average model with respect to the Bayes model is called *bias*). On the other hand, the average (squared) distance between the models produced by the algorithm and this average model, is called the *variance*.

Consistency. Intuitively, when the sample size increases, it is desired that the distribution of learned models converges to the optimal model, which implies that variance vanishes and that the avreage model converges to the optimal model.

Error bounds. These will take into account all the effects. One general fact is that when the hypothesis space grows in size, then variance increases, while bias decreases (or remains constant, in the worst case).

1.5 Geometric interpretation and orthogonality (*)

Let us make a couple of restrictions in order to make a geometrical interpretation which will allow us to state the orthogonality principle well known in estimation theory.

We assume that all random variables are real-valued variables and that we use the square error distance as a risk function

$$d^2(\mathcal{X}, \mathcal{Y}) = \int_\Omega |\mathcal{X}(\omega) - \mathcal{Y}(\omega)|^2 dP_\Omega. \tag{31}$$

Note that this distance is actually defined via a norm, which itself is defined by a scalar product

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \int_\Omega \mathcal{X}(\omega)\mathcal{Y}(\omega) dP_\Omega. \tag{32}$$

Note that if Ω is finite (say containing n different objects), then our random variables are vectors from $\mathbb{R}^n$ and the scalar product is nothing else than

$$\langle \mathcal{X}, \mathcal{Y} \rangle = \sum_{i=1}^{n} p_i X_i Y_i. \tag{33}$$

where X_i (resp. Y_i) denotes $\mathcal{X}(\omega_i)$ (resp. $\mathcal{X}(\omega_i)$).

1.5.1 Hilbert spaces

The space of all[11] random variables is a Hilbert space with respect to the above defined scalar product. This means essentially that most of the geometric intuition that we have from the

[10] Strictly speaking, the closure of the hypothesis space.

[11] all random variables of finite norm, strictly speaking

Euclidean space can be applied also in this space, in particular orthogonality and projection concepts. Let us call this space $\mathcal{L}^2$, and let us call its elements (i.e. the random variables defined on $(\Omega, \mathcal{E}, P_\Omega)$) vectors.

Then, given a set of input random variables (say $\mathcal{X} = \{\mathcal{X}_1, \ldots, \mathcal{X}_n\}$), the space of all real valued functions[12] of these latter is a sub Hilbert space[13] of $\mathcal{L}^2$. Let us call it $\mathcal{L}^2_{f(\mathcal{X})}$.

Furthermore, given a finite set of random variables, the space of all linear combinations of these latter is again a sub Hilbert space of $\mathcal{L}^2_{f(\mathcal{X})}$. Let us call it $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$ and let us assume that this is our hypothesis space $\mathcal{F}$. Thus our learning machine is learning linear input/output models.

Hence, the situation is as follows (see figure 4).

- The Bayes model f_B is the vector of $\mathcal{L}^2_{f(\mathcal{X})}$ which is closest to $\mathcal{Y}$. In other words it may be obtained by projecting $\mathcal{Y}$ on $\mathcal{L}^2_{f(\mathcal{X})}$.

  Furthermore, the random variable $\mathcal{Y} - f_B$ is orthogonal to f_B, and the square norm $d^2(\mathcal{Y} - f_B)$ is the residual error.

- The optimal model f_* is the vector of $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$ which is closest to $\mathcal{Y}$. It may also be obtained by projecting $\mathcal{Y}$ on $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$. But since, $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$ is also a subspace of $\mathcal{L}^2_{f(\mathcal{X})}$, this vector may also be obtained by projecting f_B on $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$.

- The learned models lie somewhere in $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$, and are in general different from f_*. Let, for a given sample size N, $P_{(\mathcal{X},\mathcal{Y})^N}$ be the probability distribution of the learning samples induced from P_{Ω^N}, and let f_N denote the random variable representing the vectors learned for different samples of size N and $\bar{f}_N$ be the average of all these models[14], then we have

$$E_{P_{(\mathcal{X},\mathcal{Y})^N}}\left\{d^2(f_N, \mathcal{Y})\right\} = d^2(\mathcal{Y} - f_B) + d^2(f_B - f_*) + d^2(f_* - \bar{f}_N) + E_{P_{(\mathcal{X},\mathcal{Y})^N}}\left\{d^2(f_N, \bar{f}_N)\right\}. \tag{34}$$

This situation is graphically depicted in Figure 4. The last term in (34) is the model *variance*, which results from the fact that the learning sample is random. The first term in (34) is the *residual error*, also called *noise* by some researchers. Finally, the two central terms combined represent what is called in the litterature the *bias* : the first part of it is the *representation bias* which results only from the choice of the hypothesis space; the second part is a characteristic of the learning algorithm itself, and we could call it *search bias*. Often, but not always this last term is equal to zero.

Again consistency of the learning algorithm means that both variance and search bias should vanish when N becomes very large.

2 Main theorems of statistical learning theory

The objective of the present section is to *state* precisely the main theorems proven by V. Vapnik [Vap95, Vap98]. We will discuss the meaning of these theorems in intuitive terms, and we refer the interested reader to the two above references for details and proofs. Furthermore, since this document is intended as a first introduction to statistical learning theory, we will avoid

[12] all square integrable functions, strictly speaking

[13] it contains all its linear combinations, and it is closed

[14] Because $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$ is a closedlienar space, this model actually belongs to $\mathcal{L}^2_{\text{lin}(\mathcal{X})}$.

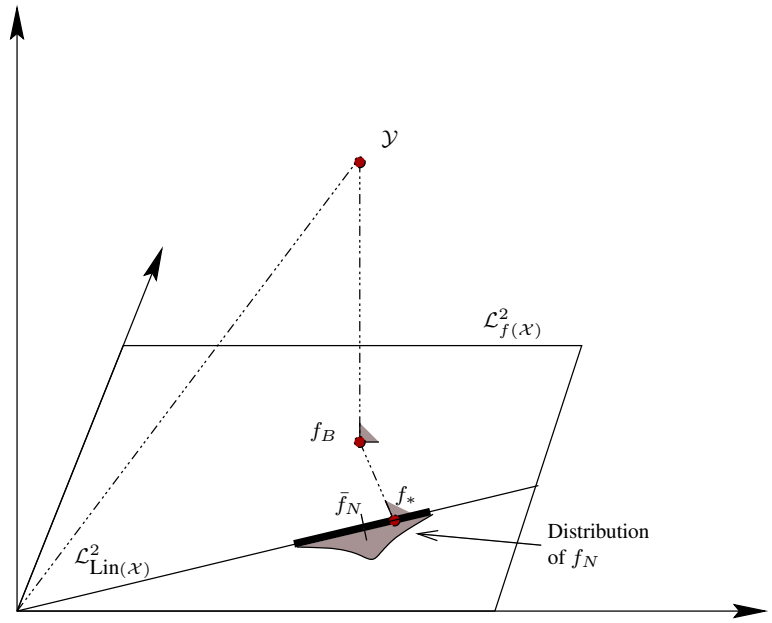

Figure 4: Geometrical interpretation of the learning problem

mathematical rigor each time that we feel that it is not strictly necessary. In what follows, we will address the following items :

- Consistency of the ERM principle (asymptotic "large sample" behaviour).
- Rate of convergence of the ERM principle (large sample behaviour).
- Error bounds on the actual Risk (finite sample behaviour)
- How to control the generalization ability of learning algorithms (structural risk minimization principle).

2.1 Consistency of the ERM principle

In this part we will construct conditions for asymtotic behaviour of the ERM principle which are dependent on the probability distribution $P_{\mathcal{X},\mathcal{Y}}$. These conditions will be extended in the next section into convergence bounds independent of this particular probability distribution.

We need first to state clearly what me mean by consistency.

Definition 1 (Consistency of the ERM principle) *We say that the ERM principle is consistent if the following two properties hold :*

$$\forall \epsilon, \eta > 0, \exists N_1 \ : N \geq N_1 \Rightarrow Pr\left\{|R(f_*^N) - R(f_*)| > \epsilon\right\} < \eta, \tag{35}$$

and

$$\forall \epsilon, \eta > 0, \exists N_2 \ : N \geq N_2 \Rightarrow Pr\left\{|R_e^N(f_*^N) - R(f_*)| > \epsilon\right\} < \eta. \tag{36}$$

In other words, we require that the expected and the empirical risks of the sequence of models produced by the ERM principle converge both in probability to the best achievable expected risk in our hypothesis space $\mathcal{F}$.

Thus consistency means that in the large sample case the ERM principle provides "the best" solution to the learning problem stated in the introduction. Figure 5 shows the typical behaviour that one could observe in such a situation for a particular sequence of samples of growing size.

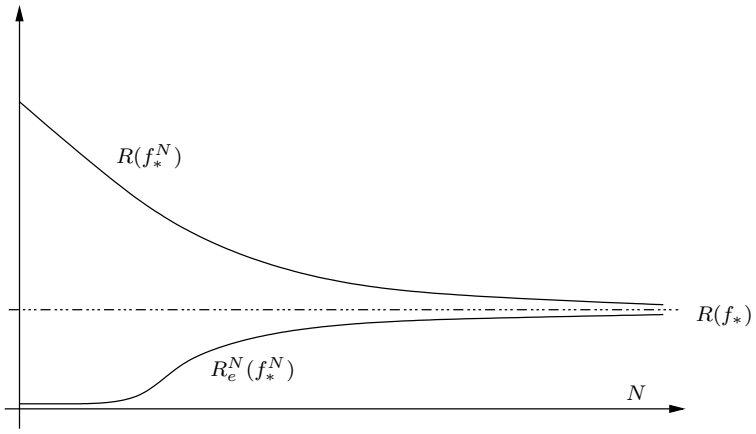

Figure 5: How empirical and expected risks should ideally behave in practice

2.1.1 Uniform convergence theorems

In section 1 we we have shown that the law of large numbers guarantees that for each function $f \in \mathcal{F}$, we have convergence of its empirical risk to its expected risk and we mentioned that this does not by itself guarantee consistency of the ERM principle.

Before stating the fundamental theorem of consistency of the ERM principle, let us introduce two slightly different notions of uniform convergence that we will use in the sequel.

Definition 2 (Uniform one-sided convergence over $\mathcal{F}$) *Uniform one-sided convergence over $\mathcal{F}$ of the empirical risk to the actual risk means that*

$$\forall \epsilon, \eta > 0, \exists N_0 \ : N \geq N_0 \Rightarrow Pr\left\{\sup_{f \in \mathcal{F}}(R(f) - R_e^N(f)) > \epsilon\right\} < \eta. \tag{37}$$

Definition 3 (Uniform (two-sided) convergence over $\mathcal{F}$) *Uniform two-sided convergence over $\mathcal{F}$ of the empirical risk to the actual risk means that*

$$\forall \epsilon, \eta > 0, \exists N_0 \ : N \geq N_0 \Rightarrow Pr\left\{\sup_{f \in \mathcal{F}}|R(f) - R_e^N(f)| > \epsilon\right\} < \eta. \tag{38}$$

Note that there is an obvious relationship between the two conditions. Indeed, we have

$$\left\{\sup_{f \in \mathcal{F}}|R(f) - R_e^N(f)| > \epsilon\right\} \Leftrightarrow \left\{\sup_{f \in \mathcal{F}}(R(f) - R_e^N(f)) > \epsilon\right\} \textit{ or } \left\{\sup_{f \in \mathcal{F}}(R_e^N(f) - R(f)) > \epsilon\right\},$$

and hence

$$Pr\left\{\sup_{f \in \mathcal{F}}|R(f) - R_e^N(f)| > \epsilon\right\} \geq Pr\left\{\sup_{f \in \mathcal{F}}(R(f) - R_e^N(f)) > \epsilon\right\}.$$

Thus uniform two-sided convergence implies uniform one-sided convergence, but the converse is not true. Now let us state the main theorem for the consistency of the ERM principle and its corollary (without proof).

Theorem 1 (Uniform one-sided convergence) *One-sided uniform convergence in probability over the set of candidate hypotheses $\mathcal{F}$ is a necessary and sufficient condition for consistency of the ERM principle applied to $\mathcal{F}$.*

Corollary 1 (Uniform convergence) *Two-sided convergence in probability is a sufficient condition of consistency of the ERM principle.*

Why is two-sided convergence not a necessary condition for consistency ? The reason comes from the asymetrical nature of the empirical risk minimization principle. The one-sided convergence studies the behaviour of those functions within $\mathcal{F}$ which appear on finite samples to be better than in reality. If there would remain such functions for very large sample sizes then the ERM principle could be mislead. On the other hand, the functions which appear to be worse from the empirical point of view than in reality are not likely to be close to the actual optimal function, so they will not confuse the ERM principle.

2.1.2 Consistency of the ERM principle over finite hypothesis spaces

Let us return to our example of a finite hypothesis space and show that in this case the ERM principle is always consistent.

Indeed, given $\epsilon, \eta > 0$, we know that for every function f_i in $\mathcal{F}$ we have

$$\exists N_i \ : N \geq N_i \Rightarrow Pr\left\{|R(f_i) - R_e^N(f_i)| > \epsilon\right\} < \frac{\eta}{m}, \tag{39}$$

where $m = \#\mathcal{F}$ denotes the (finite) number of functions in $\mathcal{F}$. On the other hand we have

$$\left\{\sup_{f\in\mathcal{F}} |R(f) - R_e^N(f)| > \epsilon\right\} \Leftrightarrow \left\{|R(f_1) - R_e^N(f_1)| > \epsilon\right\} or \cdots or \left\{|R(f_m) - R_e^N(f_m)| > \epsilon\right\}$$

hence,

$$Pr\left\{\sup_{f\in\mathcal{F}} |R(f) - R_e^N(f)| > \epsilon\right\} \leq \sum_{i=1}^{m} Pr\left\{|R(f_i) - R_e^N(f_i)| > \epsilon\right\}.$$

Thus, if $N \geq \sup_{i=1}^{m} N_i$ this latter condition implies that

$$Pr\left\{\sup_{f\in\mathcal{F}} |R(f) - R_e^N(f)| > \epsilon\right\} < \eta.$$

□

Thus, in the finite hypothesis space case the law of large numbers implies uniform two-sided convergence and therefore consistency of the ERM principle.

2.1.3 VC-Entropy of a set of functions

In order to generalize to infinite hypothesis spaces, we first introduce a stochastic measure of the expressive power of a set of functions. We proceed in three steps :

- Introduce the VC-entropy[15] measure for the general regression problem (which covers also the classification case);
- Formulate the consistency condition in terms of this measure;
- Provide an interpretation in the case of indicator functions used for binary classification problems;

[15] VC stands for Vapnik and Chervonenkis

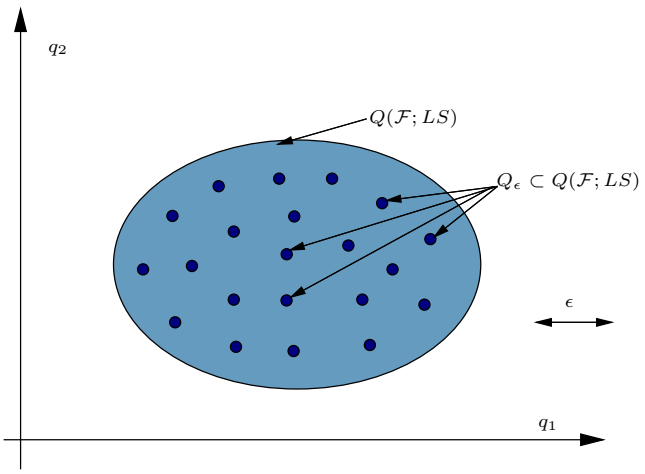

Figure 6: ϵ-net of $Q(\mathcal{F}; LS)$

Definition 4 (VC-entropy of $\mathcal{F}$ w.r.t. $P_{\mathcal{X},\mathcal{Y}}$ (and ℓ)) *Let $LS = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a sample of size N and $\mathcal{F}$ a hypothesis space. For a given function $f \in \mathcal{F}$, we denote by $q(f; LS)$ the N-dimensional approximation vector*

$$q(f; LS) = (\ell(f(x_1), y_1), \ldots, \ell(f(x_N), y_N)) .$$

and we consider the set $Q(\mathcal{F}; LS)$ of such vectors obtained when f varies in $\mathcal{F}$:

$$Q(\mathcal{F}; LS) = \{q(f; LS) \ : f \in \mathcal{F}\}.$$

This set is potentially of infinite cardinality (if $\mathcal{F}$ is itself of infinite cardinality). However, we may define (see figure 6), $\forall \epsilon > 0$, the notion of an ϵ-net of $Q(\mathcal{F}; LS)$: a subset Q_ϵ of $Q(\mathcal{F}; LS)$ which approximates $Q(\mathcal{F}; LS)$ well enough in some sense. More precisely we require that

$$\forall q \in Q(\mathcal{F}; LS) \ : \left\{ \exists q' \in Q_\epsilon \ : \sup_{i=1}^{N} |q_i - q'_i| < \epsilon \right\} .$$

Note that if the set $Q(\mathcal{F}; LS)$ is bounded (which we suppose true here) then, $\forall \epsilon > 0$, there exists a finite ϵ-net; hence the number of vectors of the smallest ϵ-net of $Q(\mathcal{F}; LS)$ is well defined under this condition : let us denote this lower bound by $N^{\mathcal{F}}(\epsilon; LS)$. We then define the VC-entropy of $\mathcal{F}$ by

$$H^{\mathcal{F}}(\epsilon; N) = E_{LS}\{\ln N^{\mathcal{F}}(\epsilon; LS)\}, \tag{40}$$

where the expectation is taken with respect to $(P_{\mathcal{X},\mathcal{Y}})^N$, the probability distribution of i.i.d. samples of size N drawn according to $P_{\mathcal{X},\mathcal{Y}}$.

Note that in the definition of the VC-entropy the functions which are considered are actually the functions $\ell(f(x), y)$ which generate the approximating vectors, and not directly the functions $f(x)$. We will see that in the case of (binary) classification problems these two concepts are actually identical. The VC-entropy quantifies the average diversity of approximations of the output values of samples of size N that can be achieved with the set of functions $\mathcal{F}$. Clearly, for fixed loss-function ℓ and data-generating distribution $P_{\mathcal{X},\mathcal{Y}}$, the VC-entropy increases when ϵ decreases. It also increases when the size of the learning sample increases. This measure allows us to characterize uniform two-sided convergence in terms of properties of the hypothesis space $\mathcal{F}$, as a function of the data generation distribution $P_{\mathcal{X},\mathcal{Y}}$ and loss-function ℓ, by the following theorem.

Theorem 2 (Necessary and sufficient conditions for uniform two-sided convergence)

$$\lim_{N\to\infty} \frac{H^{\mathcal{F}}(\epsilon; N)}{N} = 0, \forall \epsilon > 0. \tag{41}$$

In other words, if the expected diversity of the approximation vectors that can be produced by the functions in $\mathcal{F}$ grows slowly enough with N then the ERM principle is consistent. Let us notice that this property may be refined in order to provide conditions of one-sided uniform convergence; we refer the interested reader to [Vap98], where these latter (asymetric and less stringent) conditions are stated.

2.1.4 Interpretation for binary classification problems

The above notions and conditions have been introduced in the general case, so they are valid in particular for binary classification problems where the functions f and $\mathcal{Y}$ are indicator functions and where we use the discrete 0-1 loss function.

Now, in this discrete case the value of $\ell(f(x_i), y_i)$ may take only the two values 0 or 1. Hence, the set $Q(\mathcal{F}; LS)$ is a subset of the set of all N-dimensional bit-vectors. This latter set is finite and of size 2^N, and thus the size of $Q(\mathcal{F}; LS)$ and also $N^{\mathcal{F}}(\epsilon; LS)$ can not be larger than 2^N. Actually, for $\epsilon < 1$ we now have necessarily

$$Q(\mathcal{F}; LS) = Q_\epsilon$$

and hence we also have

$$N^{\mathcal{F}}(\epsilon; LS) = \#Q(\mathcal{F}; LS),$$

where the symbol $\#$ denotes the cardinality (number of elements) of a set.

Furthermore, if we define

$$y(f; LS) = (f(x_1), \ldots, f(x_n))$$

and

$$Y(\mathcal{F}; LS) = \{y(f; LS) \ : f \in \mathcal{F}\},$$

it is easy to see that

$$\#Q(\mathcal{F}; LS) = \#Y(\mathcal{F}; LS).$$

This provides another interpretation of the VC-entropy. Indeed, the right member of the latter equation is equal to the number of ways the set $\mathcal{F}$ of indicator functions may dichotomize the set of input vectors contained in the LS. Thus, in the binary classification problem, the VC-entropy measures the average logarithm of the number of ways that the hypothesis space can dichotomize samples of size N. If the hypothesis space contains all possible dichotomies of the input space then

$$\#Y(\mathcal{F}; LS) = 2^N, \forall N$$

and

$$H^{\mathcal{F}}(\epsilon; N) = N \ln 2.$$

Hence two-sided convergence can not hold anymore, and therefore the ERM principle can not anymore be guaranteed to be consistent.

2.1.5 Comments

The previous derivations are rather general. In particuler no restriction was made on the structure of the input space, which can be a very small or a very large set.

Actually, the structure of the input space is only indirectly taken into account via the measure of the complexity of the set of functions defined on this space (and the used ℓ). For example, if the input space is extremely simple (say $x_i = cst$), this imposes restrictions on the set of functions which can be defined on this space. Actually only a set of constant functions may be defined on this particular input space. In the classification problem this means that the largest $\mathcal{F}$ will contain only two functions. Hence $\#Y(\mathcal{F}; LS) \leq 2, \forall LS$. Therefore, the VC-entropy is also finite (equal to $\ln 2$, at most) and the ERM principle can be used in order to learn the best function (i.e. identify the majority class).

However, even if the input space is very large (say $\mathbb{R}^n$, with $n \gg 1000$) the used function space $\mathcal{F}$ may still be simple enough to guarantee consistency. Nevertheless, the condition (41) has several weaknesses from the practical viewpoint :

- It is probability distribution dependent : for a given hypothesis space, learnability in the limit depends on the probability distribution $P_{\mathcal{X},\mathcal{Y}}$. Since the latter is normally unknown, the condition can not be checked in practice.
- Even if we assume that $P_{\mathcal{X},\mathcal{Y}}$ is known, there is no easy way to compute the value $H^{\mathcal{F}}(\epsilon; N)$.
- Even if we were able to check the condition (41), it is a very weak condition. It does not for instance guarantee that the convergence rate of the ERM principle is fast.

All these reasons lead to an alternative formulation of the dimension of set of hypotheses, which is distribution independent, computable, and at the same time provides convergence rates and error bounds for finite sample sizes. The unavoidable price to pay for this is that the conditions obtained may be overestrictive and overly pessimistic in terms of convergence rates.

2.2 Distribution-independent fast asymptotic rate of convergence of the ERM principle for binary classification problems

A property of a learning machine (i.e. a set of functions and a learning principle) is said to be distribution-independent if it holds true for any possible probability distribution $P_{\mathcal{X},\mathcal{Y}}$. Such a property may be checked without knowing the probability distribution and has a certain universal character. In automatic learning it is very important to define distribution-independent properties since in most practical cases $P_{\mathcal{X},\mathcal{Y}}$ is not known.

Let us first define what we mean by fast asympotic rate of convergence of the ERM principle.

Definition 5 (Fast asymptotic rate of convergence of the ERM principle) *We say that the rate of convergence of the sequence of models selected by the ERM principle is fast if*

$$\exists c, N_0 > 0 \ : \left\{ \forall N > N_0 \ : Pr\{|R(f_*^N) - R(f_*)| > \epsilon\} < e^{-c\epsilon^2 N} \right\}. \tag{42}$$

To provide a distribution-independent condition for fast asymptotic convergence, we restrict our discussion to sets of indicator functions, so that we don't need the notion of ϵ-net. As in the preceding section, we denote by $N^{\mathcal{F}}(LS)$ the number of different dichotomies induced on

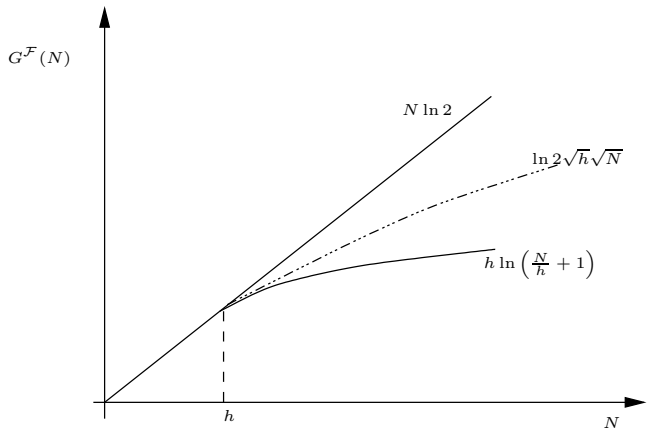

Figure 7: Possible shapes of the Growth function

a particular learning set LS by the set of indicator functions $\mathcal{F}$, and we define a distribution-independent measure of the capacity of a set of indicator functions as follows:

Definition 6 (Growth function $G^{\mathcal{F}}(N)$.) *The growth function is defined as a distribution independent version of the VC-entropy by*

$$G^{\mathcal{F}}(N) = \ln \left\{ \sup_{LS \ \textit{of size} \ N} N^{\mathcal{F}}(LS) \right\}. \tag{43}$$

Clearly, $H^{\mathcal{F}} \leq G^{\mathcal{F}}(N)$. Therefore the condition

$$\lim_{N\to\infty} \frac{G^{\mathcal{F}}(N)}{N} = 0, \tag{44}$$

certainly provides a sufficient condition for consistency of the ERM principle on $\mathcal{F}$ with respect to any possible probability measure. Actually, it turns out that this condition is much stronger; this is stated by the following theorem.

Theorem 3 (Distribution independent fast asymptotic rate of 0-1 ERM learning)

$$\lim_{N\to\infty} \frac{G^{\mathcal{F}}(N)}{N} = 0, \tag{45}$$

is a necessary and sufficient condition for distribution independent consistency of the ERM principle applied to indicator functions and the 0-1 loss function, and it is furthermore also a necessary and sufficient condition for fast asymptotic convergence rate of the ERM principle.

We refer the reader to [Vap95] for the statement of these conditions in the most general setting including regression problems.

2.3 VC dimension of a set of indicator functions

We will now introduce a very synthetic measure of the representation capacity of a set of indicator functions. Later on, we will generalize this measure for real-valued functions and general regression. Figure 7 shows the possible shapes of the growth function $G^{\mathcal{F}}(N)$ introduced in the preceding section. One can show that there are only two possible situations:

- either the growth function is linear in N and equal to $N \ln 2$: this means that for all N there exists a sample of size N which can be dichotomized in all 2^N ways[16] by functions of $\mathcal{F}$;
- or there exists a finite value h such that for $N \leq h$ the growth function is equal to $N \ln 2$ and for $N \geq h$ the growth function verifies

$$G^{\mathcal{F}}(N) \leq h \ln \left(\frac{N}{h} + 1 \right). \tag{46}$$

Thus, the growth function may not behave like the dashed curve of Figure 7.

Definition 7 (VC-dimension of a set of indicator functions) *The VC-dimension of a set of indicator functions is defined as follows*

- *If the growth function is linear for all N, the VC-dimension is said to be infinite.*
- *Otherwise the VC-dimension is finite and equal to h.*

In other words, the VC-dimension is the maximum number N such that there exists a sample of size N which can be dichotomized in all 2^N possible ways by the functions of $\mathcal{F}$.

Theorem 4 (Finite VC-dimension) *Finite VC-dimension implies that (44) holds and infinite VC-dimension implies that it does not hold. Therefore finite VC-dimension is a necessary and sufficient condition for distribution independent consistency of the ERM principle and a sufficient (and also necessary) condition for fast convergence rate.*

The main interest of these results is that the VC-dimension may be computed or at least approximated for many practical hypothesis spaces once and for all (irrespectively of the learning problem). Examples follow immediately.

2.3.1 Examples of VC-dimensions for indicator functions

We have noted above that in the case of indicator functions and using the discrete loss function, the number of different vectors in the set $Q(\mathcal{F}; LS)$ is identical to the number of different vectors in the set $Y(\mathcal{F}; LS)$. Therefore the VC-dimension (and growth function) which is relevant in this case is the VC-dimension of the set of functions $\mathcal{F}$.

Finite hypothesis space. Suppose that $\mathcal{F} = \{f_1, \ldots, f_m\}$. Then the growth function is upper bounded by $\ln m$. Indeed, if the number of functions is bounded by m then $\mathcal{F}$ itself is an ϵ-net for itself and for every $\epsilon > 0$. Hence the growth function $G^{\mathcal{F}}(N)$ is upper bounded by the constant $\ln m$. Thus the VC-dimension is certainly finite. Now, since the VC-dimension is the largest integer N such that $G^{\mathcal{F}}(N) = h \ln N$, the VC-dimension is certainly upper bounded by $\frac{\ln m}{\ln 2} = \log_2 m$. Note that this is consistent with the fact that with m functions it is possible to shatter at most $\log_2 m$ samples. The growth-function and also the VC-dimension of a finite hypothesis space can however be much lower, as will be illustrated in the following example.

[16] We say that the sample can be *shattered* by $\mathcal{F}$.

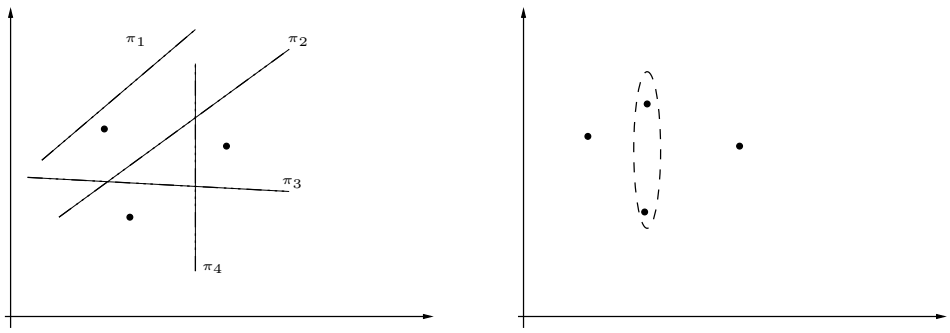

Figure 8: Shattering of points in $\mathbb{R}^{d=2}$

Linear discriminants. The set of linear indicator functions (hyperplanes) has a VC-dimension of $d+1$, where d is dimension of the input space. This is illustrated in figure 8. We observe that if $d = 2$ it is possible to construct a sample of size three which can be dichotomized in all $2^3 = 8$ ways by hyperplanes (each one of the 4 hyperplanes in the left part of figure 8 provides two dichotomies). However, it is impossible to find a sample of size 4 which may be dichotomized in all 16 ways. For example it is impossible to realize the dichotomy described in the right part of figure 8 by using a single hyperplane.

Suppose now that we take a finite subset of cardinality m (however large m is) of the set of linear d-dimensional discriminators. Then, the VC-dimension of this smaller hypothesis set can not be larger than $d+1$, even if $\log_2 m \gg d$. This shows that the VC-dimension of finite hypothesis spaces of size m can be much smaller than $\log_2 m$.

Set of rectangular indicator functions. Let us consider the set of indicator functions defined on a continuous d dimensional input space ($\mathbb{R}^d$) by

$$f(x_1,\ldots,x_d) = 1 \Leftrightarrow \left\{ \sup_{i=1}^{d} |x_i - c_i| \leq w_i \right\} \quad \text{and} \quad f(x_1,\ldots,x_d) = 0 \ \text{ otherwise,} \tag{47}$$

where the $2d$ parameters c_i, w_i may be adjusted without restrictions. One can show that this hypothesis space has a VC-dimension of $2d$, i.e. exactly the number of parameters of the hypothesis space.

Decision trees of bounded complexity. Let us consider a $d-$dimensional continuous input space ($\mathbb{R}^d$) and the set of all *orthogonal* decision trees with at most n terminal nodes which can be defined on this input space, by using tests of the form $x_i \leq c_j$. If we consider a sample of size n such that the input vectors of these learning states differ at least in one coordinate, then it is possible to shatter the sample using the set of decision trees. Hence the VC-dimension is larger or equal to n. The question is whether there are samples of size $N > n$ which can also be shattered by this hypothesis space. Does the answer depend on the dimensionality of the input space ? For instance, when the input space reduces to a single variable ($d = 1$), it is possible to show that the VC-dimension must be equal to n. Is this still valid if the dimensionality of the input space is increased ?

Decision trees of unbounded complexity. The VC-dimension of this hypothesis space is infinite. Why ?

Nearest neighbor classifiers. Let us define by n the number of prototypes used by the nearest neighbor method. Then if n is unbounded (as it is the case in the classical nearest neighbor method, where $n = N$), the VC-dimension of this hypothesis space is infinite.

Feedforward linear threshold networks. The VC-dimension of a set of such networks is upper bounded by

$$2W \log_2(eM), \tag{48}$$

where W denotes the number of parameters of the network and M the number of units.

2.3.2 VC-dimension for regression functions

The relevant VC-entropy, growth function and VC-dimension are those associated to the set $Q(\mathcal{F}; LS)$, or in other words to the set of functions $\mathcal{Q} = \{q(\cdot, \cdot) : q(x, y) = \ell(f(x), y), f \in \mathcal{F}\}$ generated by combining the loss function and the hypothesis space $\mathcal{F}$.

In the case of regression problems, it is not true that these quantities are identical to the same concepts associated to the set $Y(\mathcal{F}; LS)$ which we have called the VC-dimension of $\mathcal{F}$.

However, it is true that the VC-dimension of the set $\mathcal{F}$ provides an upper bound on the VC-dimension, as stated by the following theorem.

Theorem 5 (VC-dimension for regression problems.) *If $\ell(f(x), y)$ is a monotone function of $|f(x) - y|$ (which is the case in the usual formulation of regression problems), then the VC-dimension associated to the set of functions $\mathcal{Q} = \{q(\cdot, \cdot) : q(x, y) = \ell(f(x), y), f \in \mathcal{F}\}$ is not larger than $2h$ where h is the VC-dimension of the set of functions $\mathcal{F}$.*

To use this result, we first need to know how to extend the notion of VC-dimension to sets of real-valued functions used in regression problems (i.e. beyond mere indicator functions). The next definition shows how to extend this notion to a set of real-valued functions $\mathcal{F}$.

VC-dimension of real-valued functions. The extension is as follows. Let $\mathcal{F}$ be set of real-valued functions. Then for each $f \in F$ let us define a set of indicator functions $\mathcal{I}_f$ which is obtained by

$$\mathcal{I}_f = \{g_{\theta,f} \ : g_\theta(x) = 1(f(x) - \theta), \theta \in \mathbb{R}\} \tag{49}$$

and let us consider the union of all these sets as a new hopthesis space

$$\mathcal{I}_\mathcal{F} = \bigcup_{f \in \mathcal{F}} \mathcal{I}_f. \tag{50}$$

Then the VC-dimension of $\mathcal{F}$ is by definition the VC-dimension of the set $\mathcal{I}_\mathcal{F}$. A little thinking shows that this definition is consistent with the previous definition (apply it to a set of indicator functions).

VC-dimension of linear regression functions. It follows from the preceding that the VC-dimension of the set of linear regression functions over an input space of dimension d is equal to $d + 1$ (the same as for linear classifiers). Hence, the *relevant* VC-dimension associated to linear regression functions and a quadratic loss function does not exceed $2d + 2$.

A more general result is given next.

Theorem 6 (VC-dimension of linear hypothesis spaces) *A linear hypothesis space is a space of functions which contains all its finite linear combinations. Such a space may be either of finite or of infinite geometric dimension. Finite geometric dimension means that the space contains a finite basis $\{f_1, \ldots, f_m\} \subset \mathcal{F}$, i.e. a finite number of linearly independent functions such that any function in $\mathcal{F}$ can be written as a linear combination of these latter. The geometric dimension of the space is then by definition the dimension of any such basis, i.e. m. Such a hypothesis space can therefore be defined on a finite set of parameters which represent the coordinates of its functions in the chosen basis. It turns out that the VC-dimension of any linear space of functions which has a finite geometric dimension, is equal to this geometric dimension (which is generally different from the dimension of the input space).*[17] *Thus the relevant VC-dimension is in this case upper bounder by $2m$.*

On the other hand, the VC-dimension of a set of functions non-linearly defined on some finite parameter set may be equal, smaller or larger then the number of parameters.

Definition 8 (Universal representation property) *A set of functions $\mathcal{F}$ has the universal representation property w.r.t. a class of functions $\mathcal{G}$, if for any $\epsilon > 0$ and any $g \in \mathcal{G}$,*

$$\exists f \in F \ : \sup_x |f(x) - g(x)| \leq \epsilon. \tag{51}$$

Then we have the following theorem.

Theorem 7 (VC-dimension of universal approximators) *If $\mathcal{F}$ is a universal approximator of $\mathcal{G}$ then the VC-dimension of $\mathcal{F}$ can not be smaller than the VC-dimension of $\mathcal{G}$.*

For example, single hidden layer neural networks from $\mathbb{R}^d$ into $\mathbb{R}$ with unconstrained number of neurons has the universal approximation property w.r.t. to the set of continuous functions from a fixed compact subset of $\mathbb{R}^d$ into $\mathbb{R}$. The set of regression trees of unconstrained complexity also has the universal approximation property on the same set of functions. The same holds true for most nearest-neighbor type of methods, for projection pursuit regression and for radial basis functions.

Thus all these *large* hypothesis spaces have actually an infinite VC-dimension.

Theorem 8 (VC-dimension of single hidden layer perceptrons) *Let $\mathcal{F}$ denote the set of all single hidden layer perceptrons defined from $\mathbb{R}^d$ into $\mathbb{R}$ and with at most M units in the hidden layer (and one single output neuron). The VC-dimension of this set of functions is upper-bounded by*

This information is missing (52)

2.4 Bounds on the generalization ability of learning machines

To describe the generalization ability of learning machines based on the ERM principle we need to answer the following questions

- What actual risk $R(f_*^N)$ is provided by a function that achieves minimal empirical risk $R_e(f_*^N)$ on a sample of size N ?

[17] As an example of an infinite-dimensional linear space we quote the set of continuous real-valued functions defined on a compact subset of $\mathbb{R}^d$.

- How close is this actual risk to best achievable risk $R(f_*)$ in the given hypothesis set[18] ?

The first question is of very general interest in automatic learning. Actually, the error bounds that we will provide to answer this question are valid even if the function f_N is not selected by the ERM principle. Indeed, even for functions which are not selected in this way it is useful to obtain error bounds, as we will discuss later on.

The second question is interesting in order to determine if the sample size is large enough for the chosen hypothesis space. Unfortunately, the corresponding bounds may be obtained only if the functions are actually selected according to the ERM principle (because only under this condition is it possible to characterize the convergence properties).

Because we are interested in the minimization of the risk, we will answer these questions in the form of one-sided confidence intervals. Then the bounds are as follows (see [Vap98])

Theorem 9 (Bounds for classification problems.) *Let us fix $\alpha \in [0;1]$ (typically α would be close to 0), and define*

$$\mathcal{E} \triangleq 4\frac{G^{\mathcal{F}}(2N) - \ln\left(\frac{\alpha}{4}\right)}{N}, \tag{53}$$

where N is the size of the LS used for learning and for computing empirical risks.

Then

- *for any function $f \in \mathcal{F}$, we have with probability at least $1-\alpha$*

$$R(f) \leq R_e(f) + \frac{1}{2}\sqrt{\mathcal{E}}. \tag{54}$$

- *For a function $f_*^N \in \mathcal{F}$ minimizing the empirical risk over the learning sample of size N, we also have with probability at least $1-2\alpha$*

$$R(f_*^N) \leq R(f_*) + \sqrt{\frac{-\ln\alpha}{2N}} + \frac{1}{2}\sqrt{\mathcal{E}}. \tag{55}$$

The above bounds can be simplified and reformulated in terms of the VC-dimension h (or an upper bound of the latter). Indeed, noting that $G^{\mathcal{F}}(2N) \leq h\ln\left(\frac{2N}{h}+1\right)$, we obtain

- for any function $f \in \mathcal{F}$, we have with probability $1-\alpha$

$$R(f) \leq R_e(f) + \frac{1}{2}\sqrt{4\frac{h\ln\left(\frac{2N}{h}+1\right) - \ln\left(\frac{\alpha}{4}\right)}{N}}. \tag{56}$$

- For the function $f_*^N \in \mathcal{F}$ minimizing the empirical risk, we also have with probability $1-2\alpha$

$$R(f_*^N) \leq R(f_*) + \sqrt{\frac{-\ln\alpha}{2N}} + \frac{1}{2}\sqrt{4\frac{h\ln\left(\frac{2N}{h}+1\right) - \ln\left(\frac{\alpha}{4}\right)}{N}}. \tag{57}$$

If the hypothesis space is finite of size m, we can also use a simpler version of these bounds, noting that $G^{\mathcal{F}}(2N) \leq \ln m$ in this case. This gives the following bounds:

[18] Note that actually we should use R_* here instead of $R(f_*)$.

- for any function $f \in \mathcal{F}$, we have with probability $1 - \alpha$

$$R(f) \leq R_e(f) + \frac{1}{2}\sqrt{2\frac{\ln m - \ln \alpha}{N}}. \tag{58}$$

- For the function $f_*^N \in \mathcal{F}$ minimizing the empirical risk, we also have with probability $1 - 2\alpha$

$$R(f_*^N) \leq R(f_*) + \sqrt{\frac{-\ln \alpha}{2N}} + \frac{1}{2}\sqrt{2\frac{\ln m - \ln \alpha}{N}}. \tag{59}$$

2.4.1 Comments

In practice it is necessary to distinguish between situations where the ERM principle can be truly implemented and more general situations where it is not possible to formulate a learning algorithm which actually implements the ERM principle. We know that for most learning problems their may be several local minima of the empirical risk and we have no algorithms which provide the guarantee that the global optimum is reached. In this case, the bounds which are really useful are those which are independent of the search strategy, i.e. those which are valid for any function in $\mathcal{F}$ which might be produced by a learning algorithm.

We also saw that the determination of the VC-dimension of various hypothesis spaces is not necessarily a trivial problem and for many popular hypothesis spaces the actual VC-dimension is not well characterized at the time of writing this document. Nevertheless, we will reformulate the practical error bounds in the next section in a form useful in practice if the VC-dimension is indeed known.

2.4.2 Practical error bounds (for binary classification problems)

A derivation which we will not reproduce here provides the following three types of error bounds for any function $f \in \mathcal{F}$ of finite VC-dimension h (in the case of binary classification problems).

$$R(f) \leq R_e(f) + \epsilon_1(N, h, \alpha, R_e(f)), \tag{60}$$

where

$$\epsilon_1(N, h, \alpha, R_e(f)) = \frac{\mathcal{E}}{2}\left(1 + \sqrt{1 + \frac{4R_e(f)}{\mathcal{E}}}\right) \tag{61}$$

where

$$\mathcal{E} = 4\frac{h}{N}\ln\left(\frac{2N}{h} + 1\right) - \frac{1}{N}\ln\left(\frac{\alpha}{4}\right). \tag{62}$$

If $R_e(f)$ is close to 0 then this formula can be reshaped into

$$R(f) \leq R_e(f) + \mathcal{E}. \tag{63}$$

This latter formula can be used in practical situations, since in general $R_e(f)$ will be close to zero. Note that this formula is valid for classification problems using the 0-1 loss function and more generally for any problem where the loss function $\ell(\cdot, \cdot)$ belongs to the interval $[0; 1]$. Furthermore, if the loss function is bounded by a constant L, then it is always possible to transform it into a loss function which has this property by multiplying/dividing the suitable terms by L.

2.4.3 Regression problems

In most regression problems it is not possible to ascertain that the loss function is bounded. (Think, for instance, about a problem where there is additive white (Gaussian) noise.) We refer the reader to [Vap98] for the corresponding error bounds. We merely note here that the error bounds in this reference are formulated in terms of the VC-dimension of the space $\mathcal{Q}$ of appromixation errors rather then in terms of the VC-dimension of the hypothesis space $\mathcal{F}$.

2.5 Structural risk minimization principle

When $\frac{N}{h}$ is large then $\mathcal{E}$ is small and the ERM principle is fully justified, since under these conditions the actual risk is then close to the empirical risk with high probability. This for instance fully justifies the use of the ERM principle in linear regression analysis, as soon as the number of samples is larger than, say 1000 times, the dimensionality of the input space.

If, on the other hand, $\frac{N}{h}$ is small, then a small empirical risk does not guarantee a small actual risk. The idea of structural risk minimization then consists of minimizing the sum of the two terms which define the error bound, rather then only the empirical risk. Therefore, one has to make the VC-dimension of the hypothesis space a controlling variable in the learning process. This may be achieved in the following way.

Definition 9 (Structural risk minimization (SRM) principle.) *Let the set of functions $\mathcal{F}$ be decorated with a structure of nested subsets $\mathcal{F}_i$, such that*

$$\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots \subset \mathcal{F}_n \cdots , \tag{64}$$

where the elements of the structure satisfy the following properties

- *The VC dimension h_i of each $\mathcal{F}_i$ si finite, which implies*

$$h_1 \leq h_2 \leq \cdots \leq h_n \cdots . \tag{65}$$

- *Any set $\mathcal{F}_i$ contains only a set of totally bounded functions.*

Then the SRM principle chooses the function $f \in \mathcal{F}_{i_*}$ which minimizes the empirical risk in this set, and where the set itself is chosen so that the obtained error bound is minimal.

Figures 9 illustrate the behaviour of this algorithm. Note that if the sample size increases, the error bounds change and the optimal value of h_* is hence going to increase (at least its expectation). The way the empirical risk itself decreases when h increases is dependent on the complexity of the function which is approximated. The more complex this function, the more beneficial in terms of empirical risk decrease will be the increase in h.

2.5.1 Defining admissible structures

In practice there are two main ways to define admissible structures :

- Constraining the number of parameters (e.g. the number of nodes of a tree, the number of input variables, the number of hidden neurons).
- Constraining the optimization by adding a penalization term in the objective function used for learning (so-called regularization).

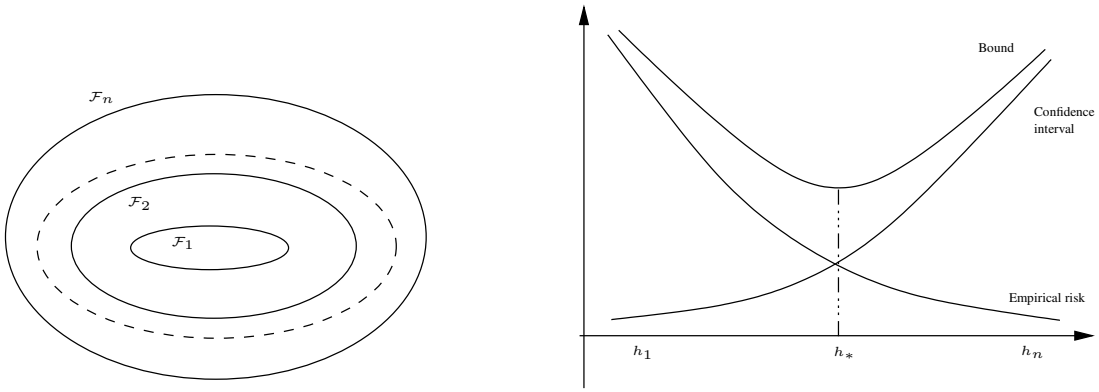

Figure 9: Structural risk minimization principle

3 Concluding remarks

The statistical learning theory depicted in this document was invented by the team of Vladimir Vapnik during the 1970's and was only broadly disseminated during the 1990's to the machine learning community. Since then, it has had a very strong influence on theoretical research in the field and it led also to several major developments in terms of novel machine learning algorithms (such as support vector machines and boosting algorithms).

In parallel, a computational learning theory had also been developed by Leslie Valiant [Val84], where the computational complexity notions are combined with the statistical properties of algorithms, in order to figure out in what conditions computational learning algorithms could be exploitable with reasonable computational resources.

The Valiant and Vapnik theories provide only partial answers to the theoretical analysis of machine learning algorithms. For example, they assume that the learning algorithms select hypotheses according to the ERM applied to a given hypothesis space, or they assume that one can describe in some specific ways the set of target functions that should be learnable.

Still, many state-of-the-art machine learning methods (Random forests, Kernel based methods, Deep-learning) do not fit to such settings. Further research shall thus be necessary (see e.g. [Pog04]) in order to better understand machine learning from the theoretical point of view.

References

[Ber93] F. Bergadano, *Machine learning and the foundations of inductive inference*, Minds and Machines **3** (1993), 31–51.

[Vap95] V. N. Vapnik, *The nature of statistical learning theory*, Springer Verlag, 1995.

[Vap98] V. N. Vapnik, *Statistical learning theory*, Adaptive and learning systems for signal processing, communications and control, Wiley, 1998.

[Val84] L. Valiant, "A theory of the learnable". *Communications of the ACM*. 27 (11): 1134-1142, 1984.

[Pog04] T. Poggio, R. Rifkin, S. Mukherjee and P. Niyogi, "General conditions for predictivity in learning theory". *Nature*, 428(6981), p.419, 2004.