1.1 General Model of Learning from Examples

A problem of learning is defined by the following components:

1. A set of three actors:
   - *The environment*: it is supposed to be stationary and it generates data $\mathbf{x}_i$ drawn independently and identically distributed (sample i.i.d.) according to a distribution $\mathcal{D}_{\mathcal{X}}$ on the space of data $\mathcal{X}$.
   - The *oracle* or *supervisor* or *professor* or *Nature,* who, for each $\mathbf{x}_i$ return a *desired answer* or *label* u_i in agreement with an unknown conditional probability distribution $F(u \mid \mathbf{x})$.
   - The *learner* or *learning machine* (LM) $\mathcal{A}$ able to fulfill a function (not necessarily deterministic) belonging to a space of functions $\mathcal{H}$ such that the exit produced by LM verifies $y_i = h(\mathbf{x}_i)$ for $h \in \mathcal{H}$.

2. The *learning task*: LM seek in the space $\mathcal{H}$ a function h who as well as possible approximate the desired response of supervisor. In the case of induction, the *distance* between the hypothesis function h and the response of the supervisor is defined by *the mean loss* on the possible situations in $\mathcal{Z} = \mathcal{X} \times \mathcal{U}$. Thus, for each entry $\mathbf{x}_i$ and response of supervisor u_i, one measures the *loss* or *cost* $l\left(u_i, h\left(\mathbf{x}_i\right)\right)$ evaluating the cost to have taken the decision $y_i = h\left(\mathbf{x}_i\right)$ when the desired answer was u_i (one will suppose, without loss of generality, the loss positive or null). The mean cost, or *real risk* is then:

$$R_{real}\left(h\right) = \int_{\mathcal{Z}} l\left(u, h\left(\mathbf{x}_i\right)\right) dF\left(\mathbf{x}, u\right)$$

It is a statistical measurement that is a function of the functional dependence $F\left(\mathbf{x}, u\right)$ between the entries $\mathbf{x}$ and desired exits u.This dependence can be expressed by a density of joint probability definite on $\mathcal{X} \times \mathcal{U}$ who is unknown. In other words, it is a question of finding a hypothesis h near to f in the sense of the loss function, and this is done particularly in the frequently met areas of the space $\mathcal{X}$. As these areas are not know a priori, it is necessary to use the training sample to estimate them, and the problem of induction is thus to seek to minimize the unknown real risk starting from the observation of the training sample $\mathcal{S}$.

3. Finally, an inductive *principle* that prescribe what the sought function h must check, according at the same time to the concept of proximity evoked above and the observed training sample $\mathcal{S} = \{(\mathbf{x}_1, u_1), ..., (\mathbf{x}_m, u_m)\}$, with the aim of minimizing the real risk.

The *inductive principle* dictate what the best assumption must check according to the training sample, the loss function and, possibly, other criteria. It acts of an ideal objective. It should be distinguished from the *learning method* (or algorithm) which describes an effective realization of the inductive principle. For a given inductive principle, there are many learning methods, which result from different choices of solving the computational problems that are beyond the scope of the inductive principle. For example, the inductive principle can prescribe that it is necessary to choose the simplest hypothesis compatible with the training sample. The learning method must then specify how to seek this hypothesis indeed, or a suboptimal hypothesis if it is necessary, by satisfying certain constraints of reliability like computational resources. Thus, for example, the learning method will seek by a gradient method, sub-optimal but easily controllable, the optimum defined by the inductive principle.

The definition given above is very general: in particular, it does not depend on the selected loss function. It has the merit to distinguish the principal ingredients of a learning problem that are often mixed in practical achievements descriptions.

1.1.1 The Theory of Inductive Inference

The inductive principle prescribes which assumption one should choose to minimize the real risk based on the observation of a training sample. However, there is no unique or ideal inductive principle single or ideal. How to extract, starting from the data, a regularity which has chances to have a relevance for the future? A certain number of "reasonable" answers were proposed. We describe the principal ones in a qualitative way here before more formally re-examining them in this and next chapters.

The choice of the hypothesis minimizing the empirical risk *(Empirical Risk Minimization* or the *ERM principle)*. The empirical risk is the average loss measured on the training sample $\mathcal{S}$:

$$R_{emp}(h) = \frac{1}{m}\sum_{i=1}^{m} l\left(u_i, h(\mathbf{x}_i)\right)$$

The idea subjacent of this principle is that the hypothesis, which agrees best to the data, by supposing that those are representative, is a hypothesis that describes the world correctly in general.

The ERM principle was, often implicitly, the principle used in the artificial intelligence since the origin, as well in the connectionism as in the learning symbolic system. What could be more natural indeed than to consider that a regularity observed on the known data will be still verified by the phenomenon that produced these data? It is for example the guiding principle of the

perceptron algorithm like that of the ARCH system. In these two cases, one seeks a coherent hypothesis with the examples, i.e. of null empirical risk. It is possible to refine the principle of the empirical risk minimization while choosing among the optimal hypothesis, either one of most specific, or one of most general.

The choice of the most probable hypothesis being given the training sample. It is the *Bayesian decision principle.* The idea is here that it is possible to define a probability distribution on the hypothesis space and that the knowledge preliminary to the learning can be expressed in particular in the form of an *a priori* probability distribution on the hypotheses space. The sample of learning is then regarded as information modifying the probability distribution on $\mathcal{H}$ (see Figure Error! **No text of specified style in document.**-1). One can then, or to choose the most probable *a posteriori* hypothesis *(the maximum likelihood principle)* or *Maximum A posteriori (MAP),* or to adopt a composite hypothesis resulting from the average of the hypotheses weighed by their *a posteriori* probability *(true Bayesian approach).*

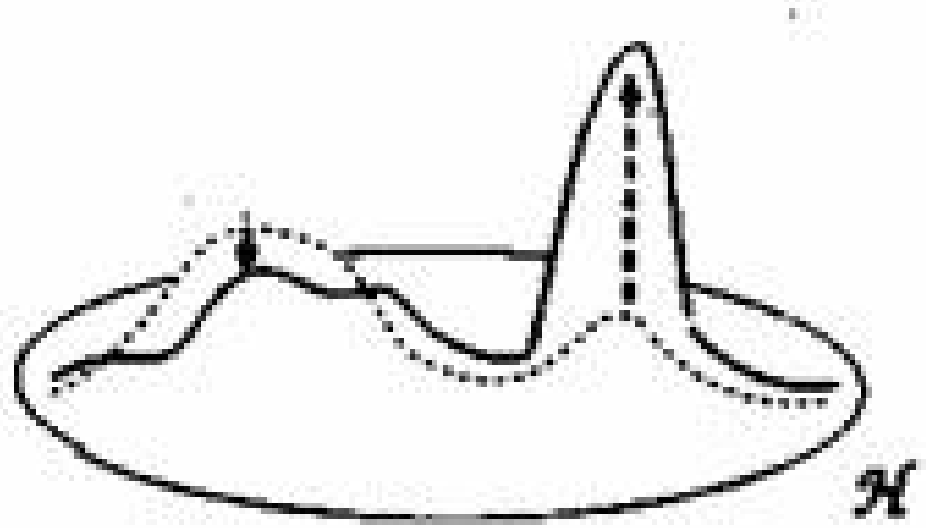

Figure Error! No text of specified style in document.**-1** The space of the assumptions $\mathcal{H}$ is presumably provided with a density of probabilities a priori. The learning consists in modifying this density according to

the learning example.

The choice of a hypothesis that compresses as well as possible the information contained in the training sample. We will call this precept: *the information compression principle*. The idea is to eliminate the redundancies present in the data in order to extract the subjacent regularities allowing an economic description of the world. It is implied that the regularities discovered in the data are valid beyond the data and apply to the whole world.

The question is to know if these ideas intuitively tempting make it possible to learn *effectively*. More precisely, we would like to obtain answers to a certain number of "naive questions ":

- does the application of the selected inductive principle to minimize the real risk indeed?
- what conditions should be checked for that? Moreover, the conditions must be verified on the training sample, or on the target functions, or by the supervisor, or on the hypotheses space.
- how the performance in generalization depends on the information contained in the training sample, or of its size, etc. ?
- which maximum performance is possible for a given learning problem?
- which is the best LM for a given learning problem?

To answer these questions implies choices that depend partly on the type of inductive principle used. It is why we made a brief description of it above.

1.1.3 Validity Conditions for the ERM Principle

In this section, we concentrate on the analysis of the inductive principle ERM who prescribes to choose hypothesis minimizing the empirical risk measured on the learning sample. It is indeed the most employed rule, and its analysis leads to very general conceptual principles. The ERM principle has initially been the subject of an analysis in the worst case, which we describe here. An analysis in the average case, utilizing ideas of statistical physics, also was the object of many very interesting works. It is however technically definitely more difficult.

Let us recall that the learning consists in seeking a hypothesis h such that it minimizes the learning average loss. Formally, it is a question of finding an optimal hypothesis h^* minimizing *the real risk:*

$$h^* = \underset{h \in \mathcal{H}}{\text{ArgMin}}\ R_{real}(h)$$

The problem is that one does not know the real risk attached to each hypothesis h. The natural idea is thus to select hypothesis h in $\mathcal{H}$ who behaves well on the learning data $\mathcal{S}$: it is the inductive principle of the ERM. We will note $\hat{h}_S$ this optimal hypothesis for the empirical risk measured on the sample $\mathcal{S}$:

$$\hat{h}_S = \underset{h \in \mathcal{H}}{\text{ArgMin}}\, R_{emp}(h)$$

This inductive principle will be relevant only if the empirical risk is correlated with the real risk. Its analysis must thus attempt to study **the correlation between the two risks** and more particularly the correlation between the real risk incurred with the selected hypothesis using the ERM principle, $R_{emp}\left(\hat{h}_S\right)$ and the optimal real risk $R_{real}\left(h^*\right)$

This correlation will utilize two aspects:

1. The difference (inevitably positive or null) between the real risk of the hypothesis $\hat{h}_S$ selected using the training sample $\mathcal{S}$ and the real risk of the optimal hypothesis h^*: $R_{real}\left(\hat{h}_S\right) - R_{real}\left(h^*\right)$.

2. The probability that this difference is higher than a given bound ε. Being given indeed that the empirical risk depends on the training sample, the correlation between the measured empirical risk and the real risk depend on the *representativeness* of this sample. This is why also, when the difference $R_{real}\left(h^*\right) - R_{real}\left(\hat{h}_S\right)$ is studied is necessary to take into account the probability of the training sample being given a certain target function. One cannot be a good learner of all the situations, but only for the reasonable one (representative training samples) which are most probable.

Thus, let us take again the question of the correlation between the empirical risk and the real risk. The *ERM* principle is a valid inductive principle if, the real risk computed with the hypothesis $\hat{h}_S$ that minimize the empirical risk, is guaranteed to be close to the optimal real risk obtained with the optimal hypothesis h^*. This closeness must happen in the large majority of the situations that can occur, i.e. for the majority of the samples of learning drawn by chance according to the distribution $\mathcal{D}_{\mathcal{X}}$.

In a more formal way, one seeks under which conditions it would be possible to ensure:

$$(\forall)\ 0 \le \varepsilon, \delta \le 1 : P\left(\left|R_{real}\left(\hat{h}_S\right) - R_{real}\left(h^*\right)\right| \ge \varepsilon\right) < \delta \qquad (1)$$

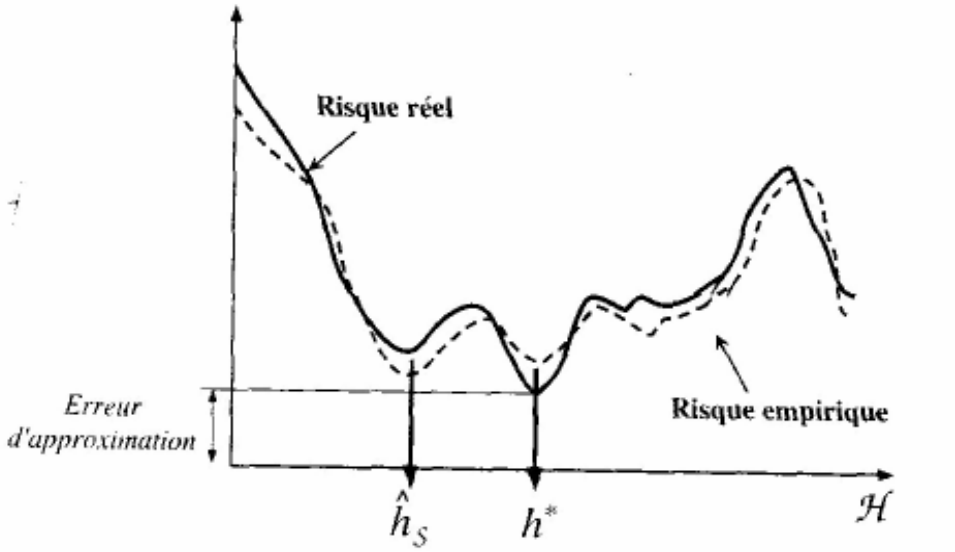

Figure Error! No text of specified style in document.**-2**

It is well understood that the correlation between the empirical risk and the real risk depends on the training sample $\mathcal{S}$ and, since this one is drawn randomly, of its size *m* too. That suggests a natural application of the *law of large numbers* according to which, under very general conditions, the average of a random variable (here $R_{emp}(h)$) converges towards its mean (here $R_{real}(h)$) when grows the size *m* of the sample.

The law of large numbers encourages to want to ensure the inequality (1) by growing the sample size of the training set $\mathcal{S}$ towards ∞ and to ask starting from which size *m* of the training sample drawn randomly (according to an unspecified distribution $\mathcal{D}_{\mathcal{X}}$), the inequality (1) is guaranteed:

$$(\forall)\ 0 \le \varepsilon, \delta \le 1 : (\exists) m \text{ such that } P\left(\left|R_{real}\left(h^*\right) - R_{real}\left(\hat{h}_{S_m}\right)\right| \ge \varepsilon\right) < \delta$$

The Figure Error! No text of specified style in document.**-3** illustrates the desired convergence of the empirical risk towards the real risk.

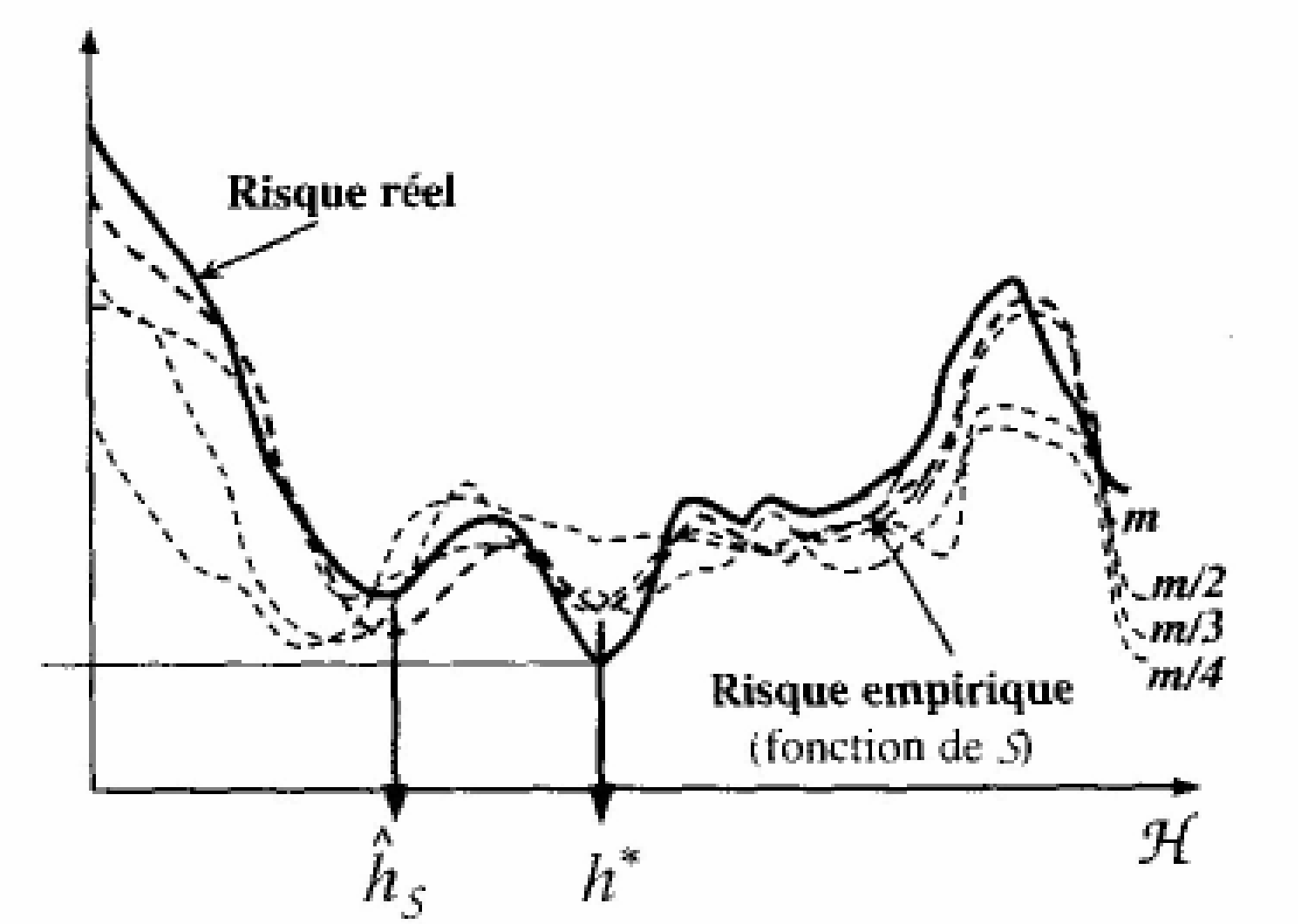

Figure Error! No text of specified style in document.**-3**

Definition 1.1 (The consistency of the *ERM principle)*

It is said that the ERM *principle is consistent if the unknown real risk* $R_{real}\left(\hat{h}_S\right)$ *and the empirical risk* $R_{emp}\left(\hat{h}_S\right)$ *converge towards the same limit* $R_{real}\left(h^*\right)$ *when the size m of the sample tends towards* ∞ *(see* Figure Error! No text of specified style in document.**-4***).*

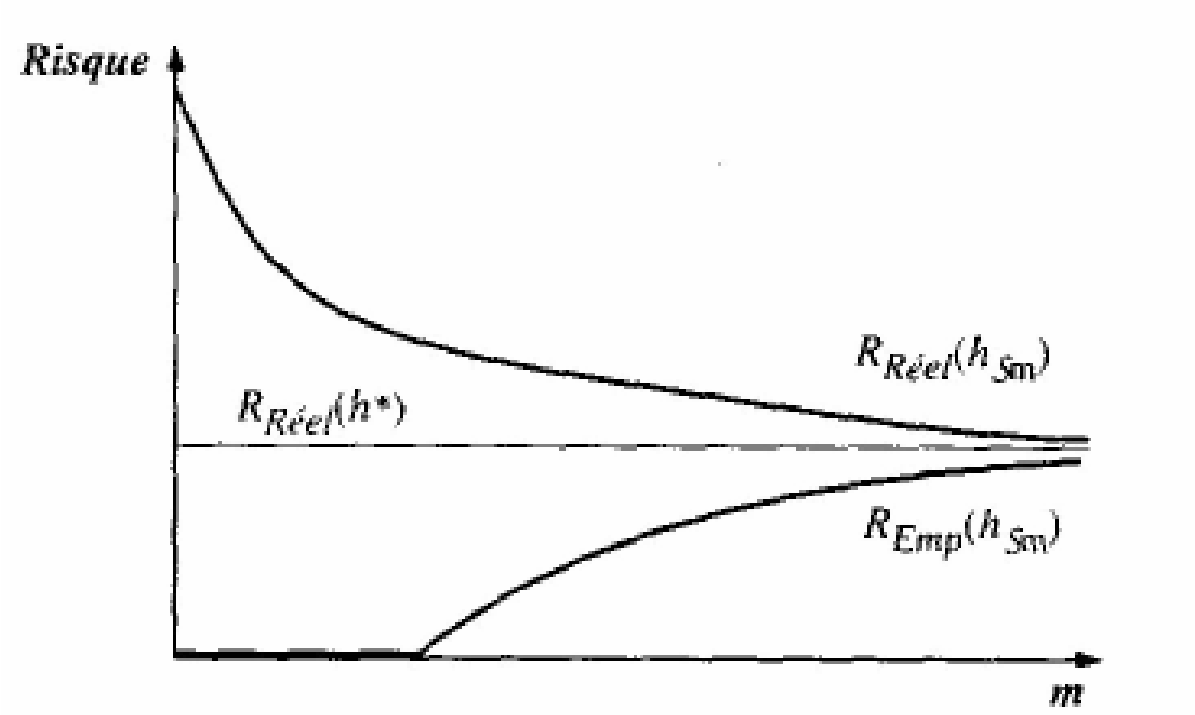

Figure Error! No text of specified style in document.**-4** *Consistency of the* ERM *principle*.

Unfortunately, the law of large numbers is not sufficient for our study. Indeed, what this law affirms, it is that the empirical risk *of* a given *hypotheses* data h converge towards its real risk. However, what we seek is different. We want to ensure that the hypotheses $\hat{h}_{S_m}$ taken in $\mathcal{H}$ and who minimizes the empirical risk *for the sample* $\mathcal{S}$ has an associated real risk that converges towards the optimal real risk obtained for the optimal hypotheses h^* independent of $\mathcal{S}$. It is well necessary to see that in this case the training sample does not play only the role of a test set, but also the role of a set being used for the choice of the hypothesis. One cannot thus take without precaution the performance measured on the learning sample as representative of the real performance.

Indeed one can build hypotheses spaces $\mathcal{H}$ such that it is always possible to find a hypothesis with null empirical risk without that indicate a good general performance. It is sufficient to imagine a hypothesis that agrees to all the learning data and which randomly draws the label from the not sights data. This is why it is necessary *to generalize the law of large numbers.*

This generalization is easy in the case of a finite space of hypotheses functions. It was obtained only recently by Vapnik and Chervonenkis (1971,1989), within the framework of induction, for the case of spaces of infinite size.

The problem, “How do humans generalize?” (What is the model of induction? Why is the rule that is correct for previous observations also correct for future observations?) was discussed in classical philosophy for many centuries.

Now the same question — but posed for the simplest mathematical model of generalization, the pattern recognition problem — became the subject of interest.

If the necessary and sufficient conditions for uniform convergence are not valid, that is, if the VC entropy over the number of observations does not converge to zero,

$$\frac{H^{\Lambda}_{P}(\ell)}{\ell} \longrightarrow c \neq 0,$$

then there exists a subspace X^* of the space R^n whose probability measure is equal to c,

$$P(X^*) = c,$$

such that almost any sample of vectors $x_1^*, \ldots, x_k^*$ of arbitrary size k from the subspace X^* can be separated in all 2^k possible ways by the functions from the admissible set of indicator functions $f(x, \alpha)$, $\alpha \in \Lambda$.

This means that if uniform convergence does not take place then any algorithm that does not use additional prior information and picks up one function from the set of admissible functions cannot generalize.

This, however, leaves an opportunity to use averaging algorithms that possess a priori information about the set of admissible functions. In other words VC theory does not intersect with Bayesian theory