The Bias-Variance Tradeoff

The compromise *bias-variance* express the effect of various possible factors on the final error between the hypothesis chosen by the LM and that which it would have had to choose, the ideal *target function*.

According to the general model of learning from examples, the LM receive from the environment a sample of data $\left\{\mathbf{x}_1,\ldots,\mathbf{x}_m\right\}$ where $\mathbf{x}_i \in \mathcal{X}$. In the absence of additional information on their source, and for reasons of simplicity of modeling and mathematical analysis, one will suppose that these objects are drawn randomly and independently the ones of the others according to a probability distribution $\mathcal{D}_{\mathcal{X}}$ (it is what one calls *the assumption of independently and identically distributed).* Attached with each one of these data $\mathbf{x}_i$ the LM receives in addition one *label* or *supervision* u_i produced according to a functional dependence between $\mathbf{x}$ and u.

We note $\mathcal{S} = \left\{\mathbf{z}_1 = \left(\mathbf{x}_1, u_1\right),\ldots,\mathbf{z}_m = \left(\mathbf{x}_m, u_m\right)\right\}$ the sample of learning made up here of supervised *examples*. To simplify, we will suppose that the functional dependence between an entry $\mathbf{x}$ and its label u takes the form of a function f belonging to a family of functions $\mathcal{F}$. Without loosing the generality we also suppose that there can be erroneous labeling, in the form of a *noise,* i.e. a measurable bias between the proposed label and the true label according to *f*. The LM seeks to find a hypothesis function h, in the space of the functions $\mathcal{H}$ as near as possible to *f*, the target function. We will specify later the concept of proximity used to evaluate the distance between *f* and *h.*

Figure Error! **No text of specified style in document.**-1 illustrates the various sources of error between the target functions *f* and the hypothesis function *h.* We call *total error* the error resulting from the conjunction of these various errors between *f* and *h*. Let us detail them.

- The first source of error comes from the following fact: *nothing* does not allow *a priori* to postulate the equality between the target functions space $\mathcal{F}$ of the Nature and the hypotheses functions space $\mathcal{H}$ realizable by the LM. Of this fact, even if the LM provides an *optimal* assumption h^* (in the sense of the proximity measurement mentioned above), h^* is inevitably taken in $\mathcal{H}$ and can thus be different from the target function *f*. It is the *approximation error* often called *inductive bias* (or simply bias) due to the difference between $\mathcal{F}$ and $\mathcal{H}$.

- Then, the LM does not provide in general h^* the optimal hypothesis in $\mathcal{H}$ but a hypothesis $\hat{h}$ based on the learning sample $\mathcal{S}$. Depending of this sample, the learned hypothesis $\hat{h}$ will be able to vary inside a set of functions that we denote by $\left\{\hat{h}\right\}_S$ to underline the dependence of each one of its elements on the random sample $\mathcal{S}$. The distance between h^* and the estimated hypothesis $\hat{h}$ who depends on the particularities of $\mathcal{S}$ is the *estimating error*. One can show formally that it is the *variance* related on the sensitivity of the calculation of the hypothesis $\hat{h}$ as function of the sample $\mathcal{S}$. More the hypotheses space $\mathcal{H}$ is rich, more, in general, this variance is important.

- Finally, it occurs the *noise* on labeling: because of transmission errors, the label *u* associated to $\mathbf{x}$ can to be inaccurate with respect to *f*. Hence the LM receives a sample of data relative to *disturbed function* $f_b = f + noise$. It is the *intrinsic error* who generally complicates the research of the optimal hypothesis h^*.

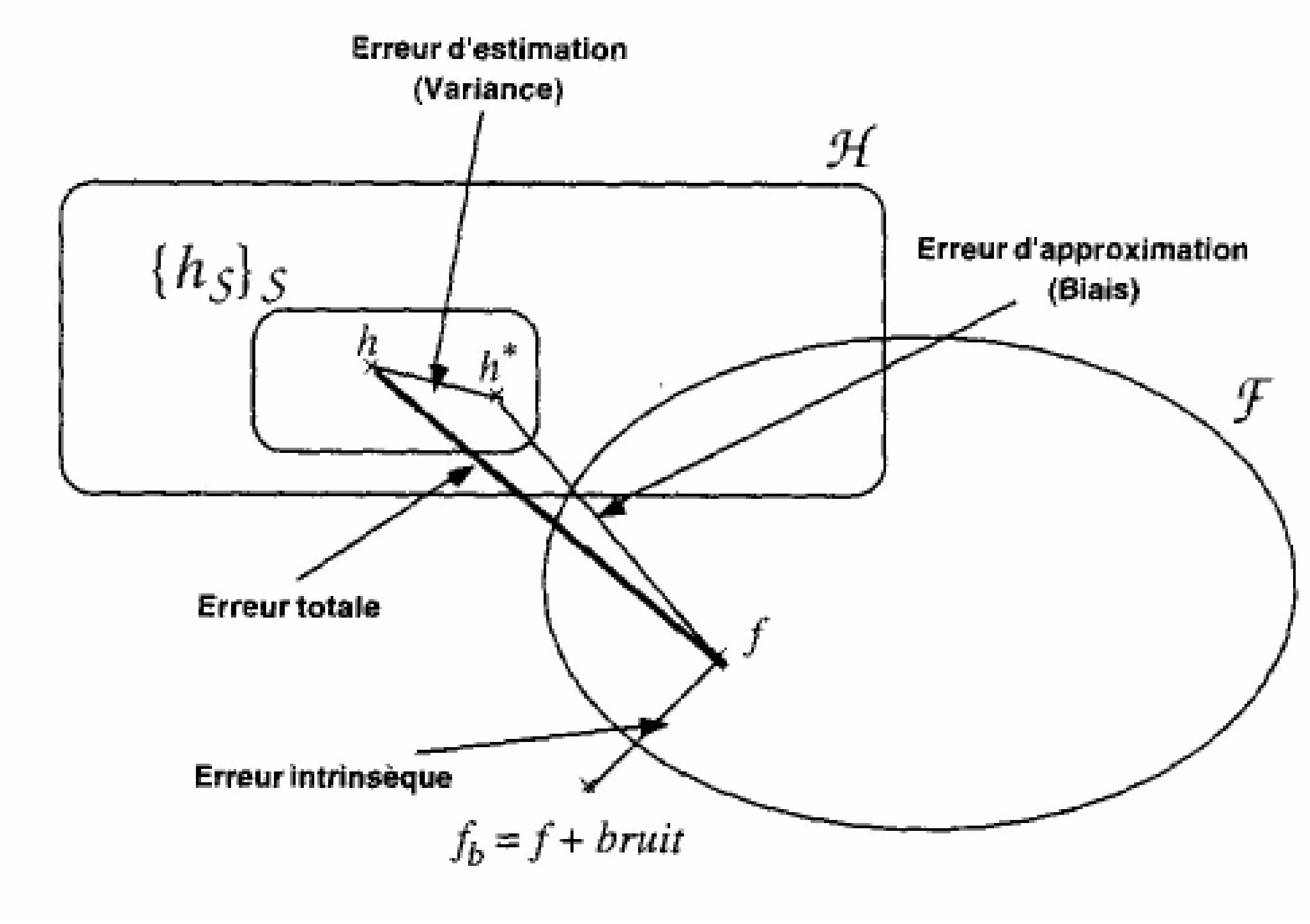

Erreur d'estimation
(Variance)
$\mathcal{H}$
$\{h_S\}_S$
h
h^*
Erreur d'approximation
(Biais)
$\mathcal{F}$
Erreur totale
f
Erreur intrinsèque
$f_b = f + bruit$

Figure Error! No text of specified style in document.**-1** *The various types of errors arising in the estimate of a targets function starting from a learning sample. With a more restricted space of hypotheses, one can reduce the variance, but generally at the price of a greater error of approximation.*

Being given these circumstances, the *bias-variance* trade off can be defined in the following way: to reduce bias due to the bad adequacy of $\mathcal{H}$ to $\mathcal{F}$ it is necessary to increase the richness of $\mathcal{H}$. Unfortunately, this enrichment will be paid, generally, with an increase in the variance. Of this fact, *the total error,* which is the sum of the approximation error and the estimation error, cannot significantly be decreased.

The bias-variance tradeoff should thus be rather called the compromise of the *approximation error/estimation error*. However, the important thing is that it is well a question of making a compromise, since one exploits a sum of terms that vary together in contrary direction. On the other hand the noise, or the *intrinsic error,* can only worsen the things while increasing. The ideal would be to have a null noise and a restricted $\mathcal{H}$ hypotheses space to reduce the variance, but at the same time *well informed,* i.e. containing only functions close to the target function, which would obviously be equivalent to have an *a priori* knowledge on Nature.

Regularization Methods

The examination of the compromise bias-variance and the analysis of the ERM principle by Vapnik have clearly shown that the mean of risk (the real risk) depends at the same time on the empirical risk measured on the learning sample and on the "capacity" of the space of the hypotheses functions. The larger this one is, the more there is a greater chance to find a hypothesis close to the target function (small approximation error), but also the hypothesis minimizing the empirical risk depends on the provided particular learning sample (big estimation error), which prohibits to exploit with certainty the performance measured by the empirical risk to the real risk.

In other words, supervised induction must always face the risk of *over-fitting*. If the space of the assumptions $\mathcal{H}$ is too rich, there are strong chances that the selected hypothesis, whose empirical risk is small, presents a high real risk. That is because several hypotheses can have a small empirical risk on a learning sample, while having very different real risks. It is thus not possible, only based on measured empirical risk, to distinguish the good hypothesis from the bad one. It is thus necessary to restrict as much as possible the richness of the hypotheses space, while seeking to preserve a sufficient approximation capacity.

Tuning the Hypotheses Class

Since one can measure only the empirical risk, the idea is thus to try to evaluate the real risk by correcting the empirical risk, necessarily optimistic, by a *penalization term* corresponding to a measurement of the capacity of $\mathcal{H}$ the used hypotheses space. It is there the essence of all induction approaches, which revise the ERM principle (the adaptation to data) by a *regularization* term (depending on the hypotheses class). This fundamental idea is found in the heart of a whole set of methods like the *regularization theory, Minimum Description Length Principle: (MDLP), the Akaike information criterion* (AIC), and other methods based on complexity measures.

The problem thus defined is known, at least empirically, for a long time, and many techniques were developed to solve there. One can arrange them in three principal categories: methods of models selection, regularization techniques, and average methods.

- In the *methods of models selection*, the approach consists in considering a hypotheses space $\mathcal{H}$ and to decompose it into a discrete collection of nested subspaces $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \ldots \subseteq \mathcal{H}_d \subseteq \ldots$ then, being given a learning sample, to try to identify the optimal subspace in which to choose the final hypothesis. Several methods were proposed within this framework, that one can gather in two types:

– *complexity penalization methods,* among which appear the *structural risk minimization principle (SRM)* of Vapnik, the *Minimum Description Length principle* of Rissanen (1978) and various methods or statistical criteria of selection,

– *methods of validation by multiple learning:* among which appears the *cross validation* and *bootstrapping*.

- The *regularization methods* act in the same spirit as the methods selection models, put aside that they do not impose a discrete decomposition on the hypotheses class. A penalization criterion of is associated to each hypothesis, which, either measure the complexity of their parametric structure, or the global properties of "regularity", related, for example, to the derivability of the hypothesis functions or their dynamics (for example the high frequency functions, i.e. changing value quickly, will be more penalized comparatively to the low frequency functions).

- The *average methods* do not select a single hypothesis in the space $\mathcal{H}$, but choose a weighed combination of hypothesis to form one prediction function. Such a weighed combination can have like effect "to smooth" the erratic hypothesis (as in the methods of *Bayesian average* and in the *bagging methods),*or to increase the capacity of representation of the hypothesis class if this one is not convex (as in the *boosting* methods*).*

All these methods generally led to notable improvements of the performances compared to the "naive" methods. However, they ask to be used carefully. On the one side, indeed, they correspond sometimes to an increase in the richness of the hypotheses space, and to an increased risk of over-fitting. On the other side, they require frequently an expertise to be applied, in particular because additional parameters should be regulated. Some recent work tries, for these reasons, to determine automatically the suitable complexity of the candidate's hypotheses to adapt to the learning data.

Selection of the Models

We will define more formally the problem of the models selection, which is the objective of all these methods.

Let $\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq ...\mathcal{H}_d \subseteq$ be a nested sequence of spaces or classes of hypotheses (or *models)* where the spaces $\mathcal{H}_d$ are of increasing capacity. The target function *f* can or cannot be included in one of these classes. Let h_d^* be the optimal hypothesis in the class of hypotheses $\mathcal{H}_d$ and $R(d) = R_{real}\left(h_d^*\right)$ the associate real risk. We note that the sequence $\left\{R(d)\right\}_{1 \leq d \leq \infty}$ is decreasing since the hypotheses classes $\mathcal{H}_d$ are nested, and thus their approximation capacity of the targets function *f* can only increase.

Using these notations, the problem of the models selection can be defined as follows.

Definition 1.2 The model selection problem *consist to choose, on the basis of a learning sample $\mathcal{S}$ of length m, a class of hypotheses $\mathcal{H}_{d^*}$ and a hypothesis $h_d \in \mathcal{H}_{d^*}$ such that the associate real risk $R_{real}\left(h_d\right)$ is minimal.*

The underlying conjecture is that the real risk associate with the selected hypothesis h_d for each class $\mathcal{H}_d$ present a global minimum for a nontrivial value of *d* (i.e. different from zero and *m*) corresponding to the "ideal" hypothesis space $\mathcal{H}_{d^*}$. (see Figure Error! **No text of specified style in document.**-2).

It is thus a question of finding the ideal hypothesis space $\mathcal{H}_{d^*}$, and in addition to select the best hypothesis h_d in $\mathcal{H}_{d^*}$.The definition say nothing about this last problem. It is generally solved by using the *ERM* principle dictating to seek the hypothesis that minimizes the empirical risk.

For the selection of $\mathcal{H}_{d^*}$, one uses an estimate of the optimal real risk in each $\mathcal{H}_d$ by choosing the best hypothesis according to the empirical risk (the *ERM* method*)* and by correcting the associated empirical risk with a penalization term related to the characteristics of space $\mathcal{H}_d$. The problem of model selection consists then in solving an equation of the type:

$$\begin{aligned} d^* &= \underset{d}{ArgMin}\left\{h_d \in \mathcal{H}_d : R_{real}^{estimated}\left(h_d\right)\right\} \\ &= \underset{d}{ArgMin}\left\{h_d \in \mathcal{H}_d : R_{emp}\left(h_d\right)\right\} + penalization\ term \end{aligned}$$

Let us note that the choice of the best hypotheses space depends on the size *m* of the data sample. The larger this one is, the more it is possible, if necessary, to choose without risk (i.e. with a little variance or confidence interval) a rich hypotheses space making possible to approach as much as possible the targets function *f*.

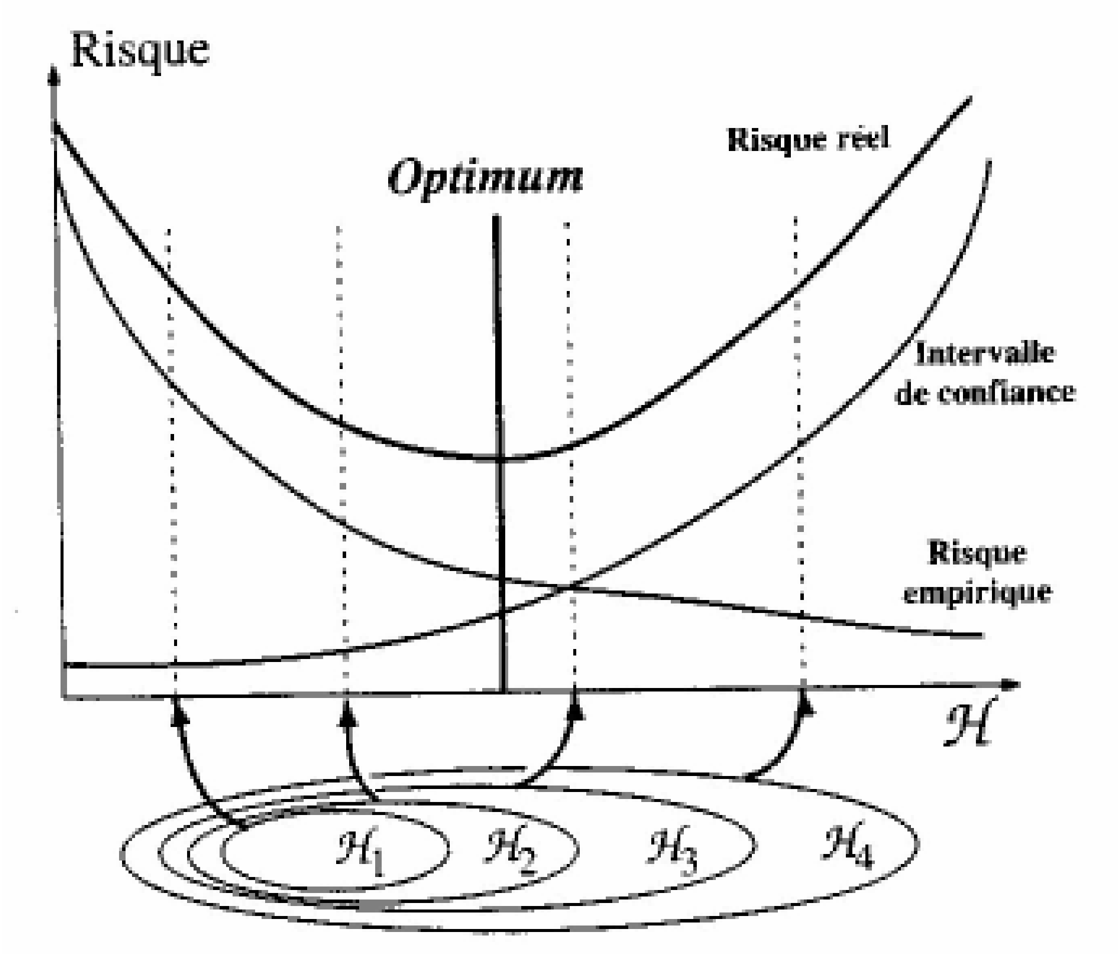


Figure Error! No text of specified style in document.**-2** *The bounds on the real risk results from the sum of the empirical risk and a confidence interval depending on the capacity of the associated hypotheses space. By supposing that one has a nested sequence of hypotheses spaces of increasing capacity and subscripted by* d*, the accessible optimal empirical risk decreases for increasing* d *(corresponding to the bias), while the confidence interval (corresponding to the variance) increases. The minimal bound on the real risk is reached for a suitable hypotheses space* $\mathcal{H}_d$.

Evaluation of the Learning Performances

We have a set of examples and a learning algorithm of which we can tune certain parameters. This algorithm returns a hypothesis. How can we evaluate the performances of this hypothesis

A solution consists in applying the theoretical results that provide probability bounds on the real risk according to the empirical risk. These bounds have the general form:

$$R_{real}(h) = R_{emp}(h) + \Phi\left(d_{VC}(\mathcal{H}), m\right)$$

where Φ is a function of the Vapnik-Chervonenkis dimension of the hypothesis space of $\mathcal{H}$ and *m* is the sample size of training $\mathcal{S}$. If one can obtain in theory asymptotically tightened bounds, the assumptions to be made to compute in practice, $\Phi\left(d_{VC}(\mathcal{H}), m\right)$ imply often such margins that the obtained bounds are too loose and do not allow to estimate the real performance precisely. It is why, except favorable particular cases, the estimate of the learning performance is estimated generally, by empirical measurements.

A Posteriori Empirical Evaluation

We admit most of the time in this paragraph that the optimization algorithm employed for the training functioned perfectly, i.e. it discovered the best hypothesis within the framework of the *ERM* principle. It is an unrealistic simplification, but of no importance for the developments presented here. To return to the practical case, it is enough to remember that in general the empirical risk of the hypothesis found by the algorithm is not minimal.

Let be $h^*_{\mathcal{H}}$ the hypothesis that minimizes $R_{real}(h)$ for $h \in \mathcal{H}$:

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

Its real risk $R_{real}\left(h^*_{\mathcal{H}}\right)$ can be noted more simply $R_{real}(\mathcal{H})$, since this hypothesis depends only on $\mathcal{H}$.

This hypothesis is theoretically that which any training algorithm should seek to approach. Nevertheless, this search is illusory: one cannot actually know if one of it is close or not. The assumption made by the *ERM* method is that one can replace his research by that of the hypothesis $h^*_{\mathcal{S},\mathcal{H}}$ described in the following paragraph.

The empirical risk of $h^*_{\mathcal{H}}$ on the training data can be noted $R_{emp}\left(h^*_{\mathcal{H}}\right)$ but this term is in general not measurable since $h^*_{\mathcal{H}}$ is unknown.

One notes $h^*_{\mathcal{S},\mathcal{H}}$ the hypothesis that minimizes the empirical risk on the training sample

$$h^*_{\mathcal{S},\mathcal{H}} = \underset{h\in\mathcal{H}}{ArgMin}\, R_{emp}\left(h\right)$$

This hypothesis is that which the learning algorithm seeks to find and using $\mathcal{S}$ in accordance to the *ERM* principle. As one is never sure that the selected algorithm finds this hypothesis, one does know neither its empirical risks $R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$ nor its real risk $R_{real}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$

It is noted by $h^*_{a\lg,\mathcal{S},\mathcal{H}}$ the hypothesis found by the training algorithm. It depends of $\mathcal{H}$, $\mathcal{S}$ and the training algorithm. Its empirical risk $R_{emp}\left(h^*_{a\lg,\mathcal{S},\mathcal{H}}\right)$ is measured on the training sample. Its real risk $R_{real}\left(h^*_{a\lg,\mathcal{S},\mathcal{H}}\right)$ can be estimated by methods that we will see below.

As one said, one supposes for the moment that $h^*_{a\lg,\mathcal{S},\mathcal{H}} = h^*_{\mathcal{S},\mathcal{H}}$ i.e. that the algorithm is effective from the *ERM* principle point of view.

However, in reality one has

$$R_{emp}\left(h^*_{a\lg,\mathcal{S},\mathcal{H}}\right) \geq R_{real}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$$

in addition, most of the time $R_{real}\left(h^*_{a\lg,\mathcal{S},\mathcal{H}}\right) \geq R_{real}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$.

Practical Selection of the Model

We seek to approach $h^*_{\mathcal{S},\mathcal{H}}$. The choice of the hypothesis space that one explores being left free, it would be useful to know *a priori* to compare two spaces $\mathcal{H}$ and $\mathcal{H}'$. However, one does not have any indication in general on the subject. On the other side, once selected $\mathcal{H}$ it is often easy to order it partially according to a criterion. One can often index its elements by the order of the algorithmic complexity of the program, which fulfills the function of corresponding decision, and parameterize the learning algorithm according to this index.

It will thus be supposed that one can define a nested sequence of sets $\mathcal{H}_k$ of increasing algorithmic complexity:

$$\mathcal{H}_1 \subset \ldots \subset \mathcal{H}_k \subset \mathcal{H}_{k+1} \subset \ldots \subset \mathcal{H}_\infty$$

We also suppose that the target function finishes by being included in one of these sets of increasing size, i.e. $f \in \mathcal{H}_\infty$.

Let us note:

- $h^*_{\mathcal{H}_k}$ the hypothesis having the smaller real risk (probability of error) of $\mathcal{H}_k$
- $h^*_{\mathcal{S},\mathcal{H}_k}$ the hypothesis having the smaller empirical risk (the apparent error rate) of $\mathcal{H}_k$.

Let us recall that we make the simplifying assumption that the learning algorithm is ideal from the point of view of the *ERM* principle: it is supposed to be able to discover for any training set the hypothesis $h^*_{\mathcal{S},\mathcal{H}_k}$ of $\mathcal{H}_k$.

What can one say relative to the values of $h^*_{\mathcal{H}_k}$ and of $h^*_{\mathcal{S},\mathcal{H}_k}$ for a given *k* and when *k* varies?

***k* is constant.**

One has first of all:

$$R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right) \geq R_{real}\left(h^*_{\mathcal{H}_k}\right)$$

This inequality translates simply the fact that the hypothesis $h^*_{\mathcal{S},\mathcal{H}_k}$, found by the algorithm, is in general not optimal in term of real error, because the training set cannot perfectly summarize the probability distributions of all the data. It is the problem of any generalization.

One has also in general:

$$R_{real}\left(h^*_{\mathcal{H}_k}\right) \geq R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right)$$

This formula expresses the fact that the learned hypothesis being tuned on the learning data, it tends to over-estimate their characteristics with the detriment of a good generalization, in sense of the *ERM* principle.

_k_ increase

One has in general, for all *k :*

$R_{real}\left(\mathcal{H}_k\right)$ decrease when *k* increases

$R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right)$ decrease when *k* increases

Indeed, the apparent error of $h^*_{\mathcal{H}_k}$ decrease when *k* increases, in general until being zero for *k* enough large: in a rather complex hypotheses space, one can learn by heart the sample $\mathcal{S}$

In general the value:

$$R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right) - R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$$

is positive and increases with *k.*

One also has, when *k* increase, till a given[1] k_0

[1] To simplify, it is supposed that there is a single value; actually, it acts of an area around this value. However, that does not change the basic argument.

$$R_{real}\left(h^{*}_{\mathcal{S},\mathcal{H}_1}\right) \geq R_{real}\left(h^{*}_{\mathcal{S},\mathcal{H}_2}\right) \geq \ldots \geq R_{real}\left(h^{*}_{\mathcal{S},\mathcal{H}_{k_0-1}}\right) \geq R_{real}\left(h^{*}_{\mathcal{S},\mathcal{H}_{k_0}}\right)$$

It seems that the increasing of k has a positive effect, since the error probability of the learned hypothesis tends to decrease.

However, exceeding k_0, the inequality is reversed:

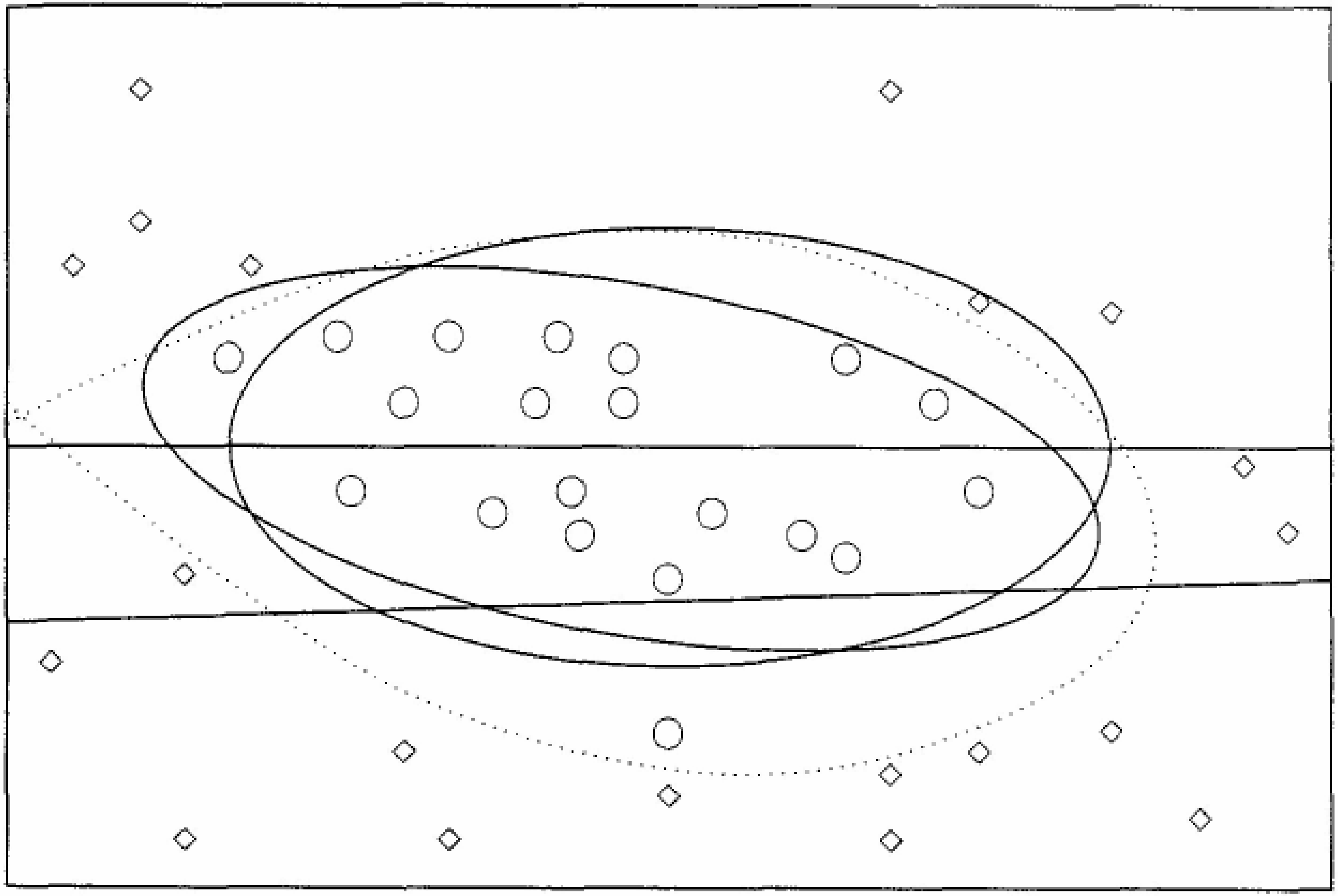
$$k \geq k_0 : R_{real}\left(h^{*}_{\mathcal{H}_k}\right) \leq R_{real}\left(h^{*}_{\mathcal{H}_{k+1}}\right)\ldots$$

Thus, beyond a value k_0 the real performance of the learned hypotheses will decrease.

The last phenomenon is called over-fitting (see Figure Error! **No text of specified style in document.**-2). Intuitively, it means that a hypothesis of too great complexity represents too exactly the training set, i.e. realizes as a matter of fact learning by heart, with the detriment of its quality of generalization. There is consequently a value k_0 of compromise, on which one does not have any *a priori* information, which is the best for a given training sample and a family of hypothesis ordered according to the complexity of *k*. The value k_0 is thus crucial. To estimate it, one can use a sample of validation $\mathcal{V}$.

In the case where the learning algorithm is not optimal from the *ERM* point of view the same phenomenon is met.

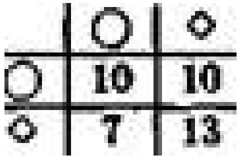
An example



Let be two classes of uniform density, one inside the central ellipse, the other between the external rectangle and this ellipse. The separating optimal curve is here known: it is the central ellipse.

To make the problem a little more difficult, we draw the 40 co-ordinates of the training points according to the uniform distributions, by adding a Gaussian noise. The separating optimal curve remains the ellipse, but the points of the two classes are not more all exactly on both sides of the separating curve. It is noticed that the noise made one of the points ○ outside the ellipse and a point ◇ in its interior. The error R_{B} is not thus not null, because of these noise effects. Its empirical value on the training data is 7.5 % (3 points badly classified on forty: $\frac{3}{40} = 7.5\%$).

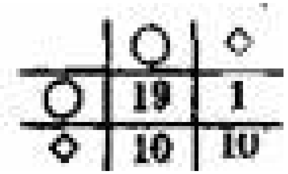
Let $\mathcal{H}_1$ be the lines, $\mathcal{H}_2$ the curves of second degree, $\mathcal{H}_3$ the curves of the third degree, etc. For $k = 1$, the separating optimal curve $h^*_{\mathcal{H}_1}$ is the horizontal straight line in the middle of **Error! Reference source not found.**. For $k = 2$, one has $h^*_{\mathcal{H}_2} = h_{\mathrm{B}}$. The best separating surface thus belongs to $\mathcal{H}_2$. One is certain here only because the data are generated and that h_{B} is known. For $k \geq 3$ one has: $h_3 = h^*_{\mathcal{H}_k} = h_{\mathrm{B}}$ since one can bring back a curve of superior degree to a curve of the second degree by canceling the necessary coefficients. The best line, which minimizes the real error, is noted with $h^*_{\mathcal{H}_1}$. It is calculable since the probability densities of the both classes are known: it is the horizontal median line of the ellipse. Its real error is also calculable: it is $R\left(\mathcal{H}_1\right) = 35\%$ .In our example, its empirical error is worth $\dfrac{10+7}{20+20} = 42.5\%$ since the empirical matrix of confusion is the following:



	○	◇
○	10	10
◇	7	13

For example, the number 7 in this matrix means that seven objects labeled ◇ have been classified like ○.

A good quality-training algorithm finds in $\mathcal{H}_1$ the line $h^*_{\mathcal{S},\mathcal{H}_1}$ that minimizes the empirical error. This one is $\frac{10+1}{20+20} = 27.5\%$ since its matrix of empirical confusion is the following one:



	○	◇
○	19	1
◇	10	10

As the distributions are uniform and the geometry fixed, one can measure $R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_1}\right)$ that is 45 %.

In $\mathcal{H}_2$, this algorithm finds $h^*_{\mathcal{S},\mathcal{H}_2}$, for which one has

$$R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_2}\right) = \frac{1+0}{20+20} = 2.5\%$$

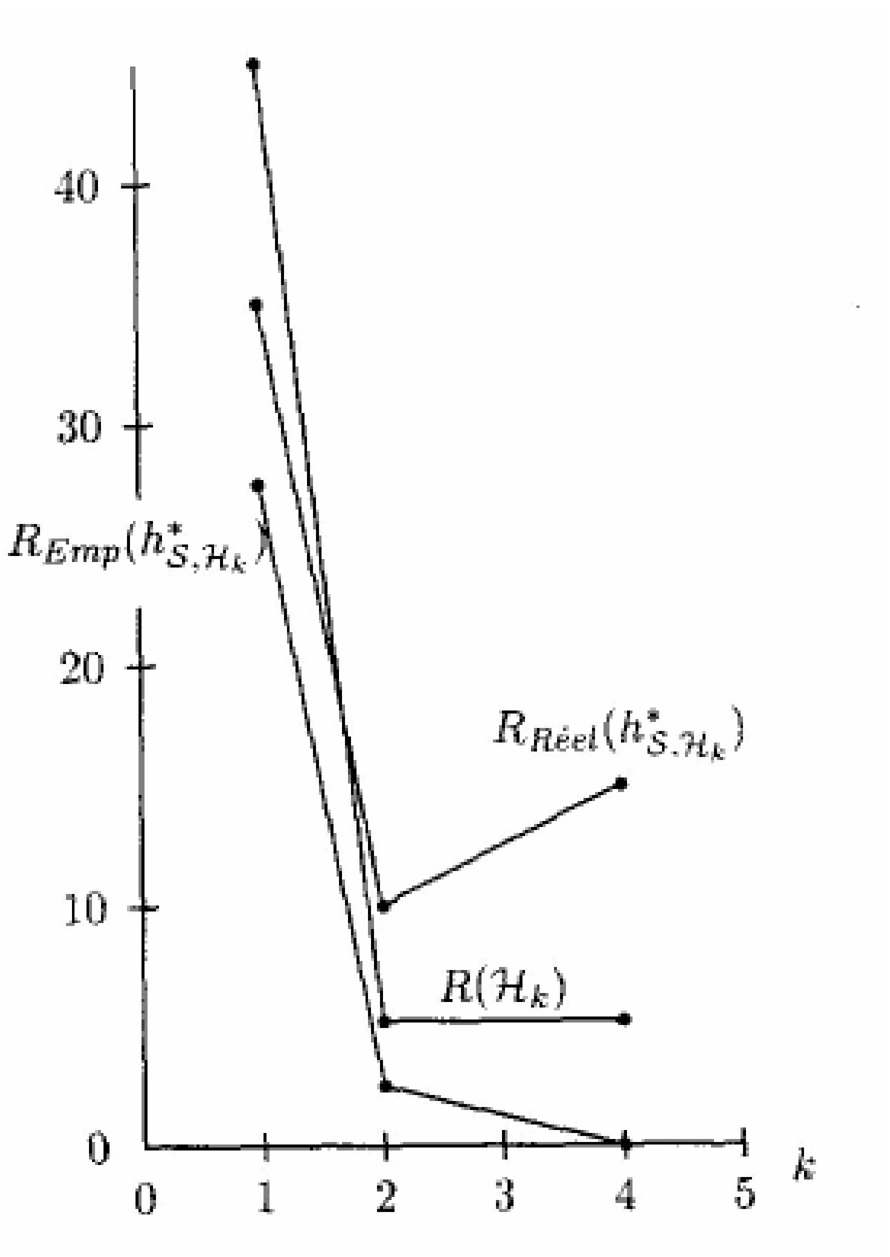
It is the best ellipse than one can trace to separate the training data: it makes only one error. $R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_2}\right)$ is 10 %.

We do not treat the case of $\mathcal{H}_3$ and we pass directly to $\mathcal{H}_4$.

In $\mathcal{H}_4$ one can find $h^*_{\mathcal{S},\mathcal{H}_4}$ such that $0 = R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_4}\right)$ but this time $R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_4}\right)$ has increased until approximately 15 % and exceeds $R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_2}\right)$.

One thus has $k_0 = 2$ in this example. In short:

Hypothesis	Curve	Empirical risk	Real risk
$h^*_{\mathcal{S},\mathcal{H}_1}$	Oblique line	27.5%	45%
$h^*_{\mathcal{S},\mathcal{H}_2}$	Leaning ellipse	2.5%	10%,
$h^*_{\mathcal{S},\mathcal{H}_4}$	Curve of degree 4	0%	15%
$h^*_{\mathcal{H}_1}$	Horizontal line	42.5%	35%
$h^*_{\mathcal{H}_2} = h_{\mathrm{B}}$	Central ellipse	7.5%	5%



***k* is constant.**

$$R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right) \geq R_{real}\left(h^*_{\mathcal{H}_k}\right)$$

$$R_{real}\left(h^*_{\mathcal{H}_k}\right) \geq R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right)$$

***k* increase**

$R_{real}\left(\mathcal{H}_k\right)$ decrease when *k* increases

$R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right)$ decrease when *k* increases

In general the value:

$$R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_k}\right) - R_{emp}\left(h^*_{\mathcal{S},\mathcal{H}}\right)$$

is positive and increases with *k.*

One also has, when *k* increase, till a given k_0

$$R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_1}\right) \geq R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_2}\right) \geq \ldots \geq R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_{k_0-1}}\right) \geq R_{real}\left(h^*_{\mathcal{S},\mathcal{H}_{k_0}}\right)$$

It seems that the increasing of *k* has a positive effect, since the error probability of the learned hypothesis tends to decrease.

However, exceeding k_0, the inequality is reversed:

$$k \geq k_0 : R_{real}\left(h^*_{\mathcal{H}_k}\right) \leq R_{real}\left(h^*_{\mathcal{H}_{k+1}}\right) \ldots$$

Thus, beyond a value k_0 the real performance of the learned hypotheses will decrease.