An introduction to statistical learning theory

Gianluca Bontempi

Machine Learning Group

Département d'Informatique

Boulevard de Triomphe - CP 212

http://www.ulb.ac.be/di

From statistics to learning

- Learning is the problem of finding a desired dependence using a limited number of observations.
- The problem of learning is so general that almost any question that has been discussed in statistical science has its analog in learning theory.
- However, some very important general results were first found in the framework of learning theory and then reformulated in the terms of statistics.
- In particular, learning theory for the first time stressed the problem of
  1. small sample statistics,
  2. nonparametric assumptions,
  3. nonlinearity

Definition of learning method

- A learning method is an algorithm (usually implemented in software) that estimates an *unknown mapping* (dependency) between a system's inputs and outputs, from the available data, i.e. *known* input-output samples.
- Once such a dependency has been accurately estimated, it can be used for prediction of system outputs from the input values.
- The goal of learning is the prediction accuracy for future data, also known as *generalization*.

Learning problem

Two are the major problems of learning

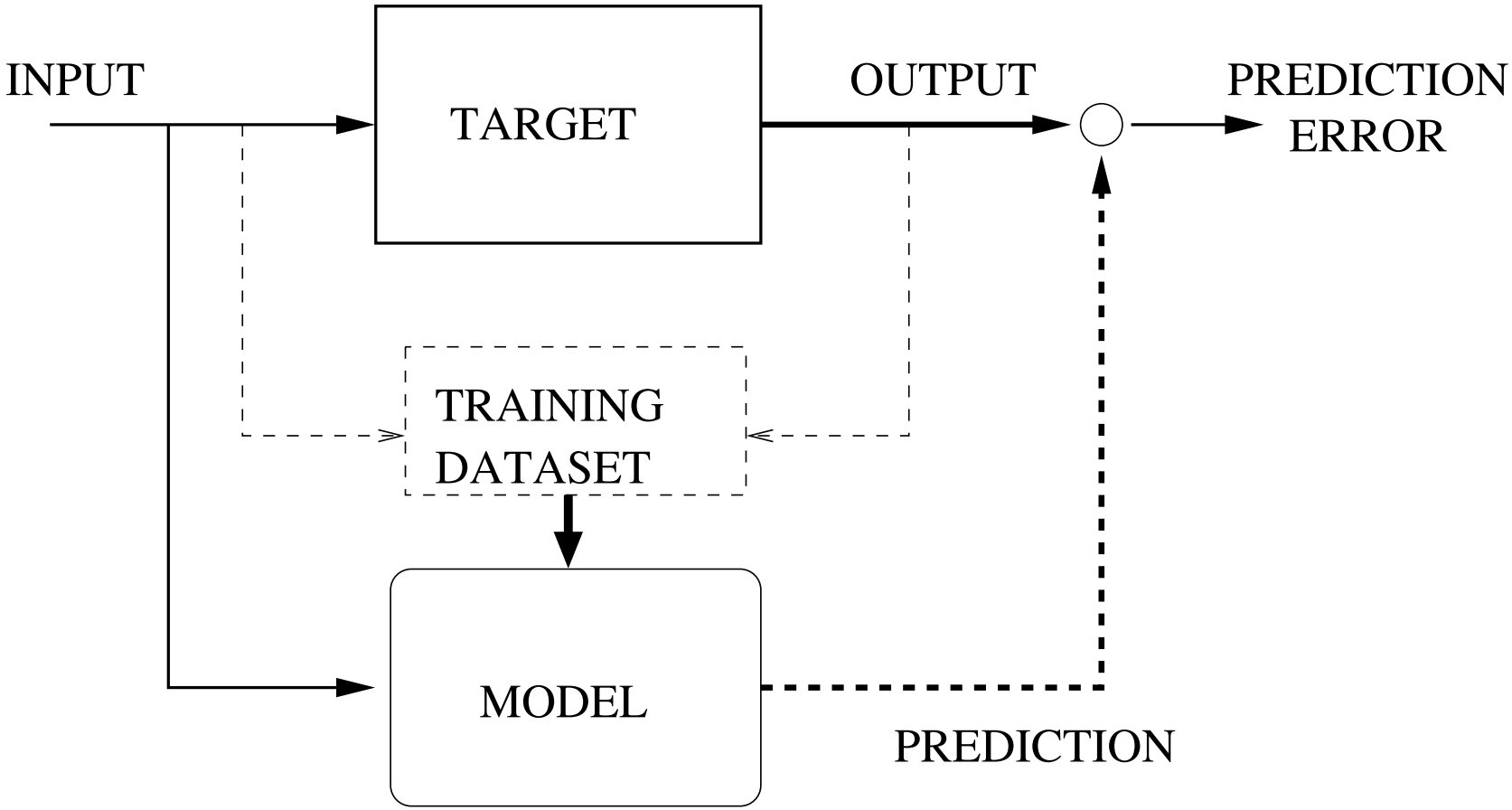
1. To estimate the desired function from a wide set of functions.
2. To estimate the desired function on the basis of a limited number of examples.

The assumption is that the observed data are somehow *representative* of the unknown probabilistic phenomenon underlying the data.

Supervised learning: the actors

- Multidimensional input $x \in \mathbb{R}^n$ and scalar output $y \in \mathbb{R}$.
- A finite number of noisy input/output observations (*training set* D_N).
- No a priori knowledge on the process underlying the data.
- A test set of input values for which an accurate *generalization* or *prediction* of the output is required.

Supervised learning



Statistical formulation of a learning machine

- A data *generator* of i.i.d input vectors $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$ according to some unknown (but fixed) probability distribution $F_{\mathbf{x}}(x)$.
- A *target* operator, which transforms the input $\mathbf{x}$ into the output value $\mathbf{y} \in \mathcal{Y} \subset \Re$ according to some unknown (but fixed) conditional distribution $F_{\mathbf{y}}(y|x)$.
- A *training set* $D_N = \{\langle x_1, y_1 \rangle, \langle x_2, y_2 \rangle, \ldots, \langle x_N, y_N \rangle\}$ made of N pairs $\langle x_i, y_i \rangle \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ independent and identically distributed (i.i.d) according to the joint distribution

$$F_{\mathbf{z}}(z) = F(\langle x, y \rangle) = F_{\mathbf{y}}(y|x) F_{\mathbf{x}}(x)$$

Statistical formulation of a learning machine

A *learning machine* has three components:

1. A class of *hypothesis* functions $h(\cdot, \alpha)$ with $\alpha \in \Lambda$. Think, for example to the class of linear functions $h(\cdot, \alpha) = x^T \alpha$. In the following, we will assume that finding the required function means determining the corresponding value of the parameter α.
2. A *cost* function $C(y, h(x))$. For instance, a quadratic cost function.
3. An *algorithm* $\mathcal{L}$ of parametric identification which takes as input the training set D_N and returns as output one hypothesis function $h(\cdot, \alpha_N)$ where $\alpha_N \in \Lambda$

$$\alpha_N = \alpha(D_N) = \arg\min_{\alpha \in \Lambda} R_{\text{emp}}(\alpha)$$

minimizes the *empirical risk*

$$R_{\text{emp}}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} C(y_i, h(x_i, \alpha))$$

Assessment criteria

Functional risk: it averages over the $\mathcal{X}\mathcal{Y}$-domain the cost C for a given hypothesis $h(\cdot, \alpha_N)$:

$$R(\alpha_N) = E_{\mathbf{xy}}[\mathbf{C}|D_N] = \int_{\mathcal{X},\mathcal{Y}} C(y, h(x, \alpha_N)) dF_{\mathbf{y}}(y|x) dF_{\mathbf{x}}(x)$$

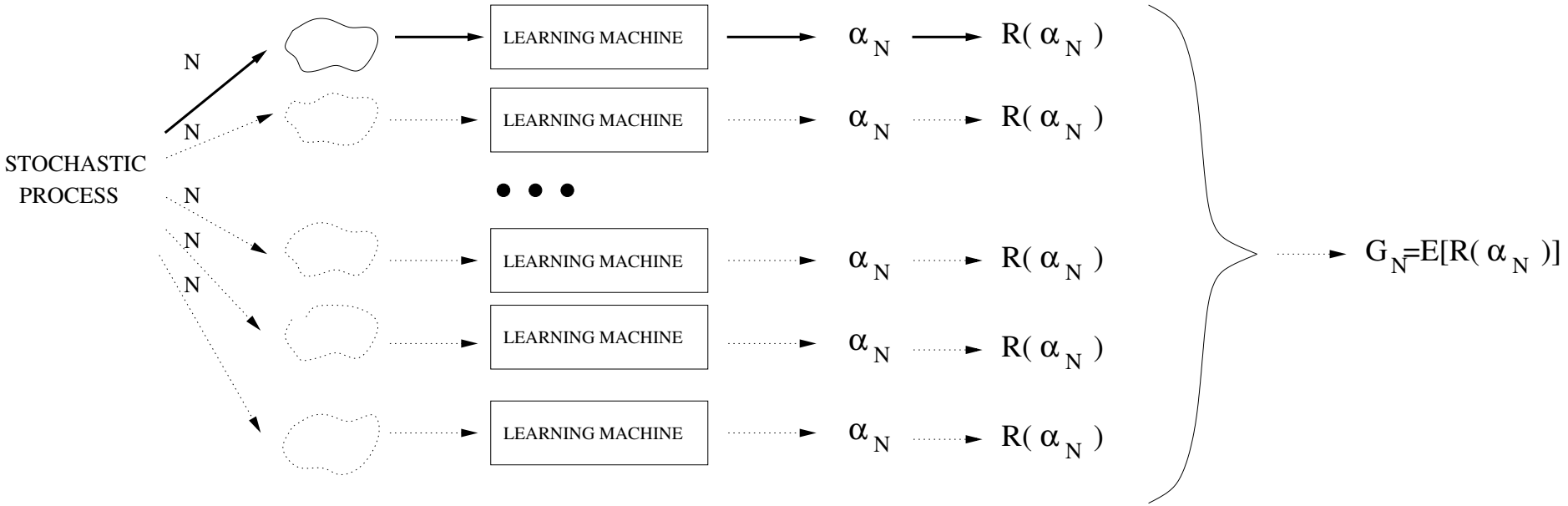
The hypothesis α_N is kept fixed.

MISE: Mean Integrated Squared Error: it is the average of the functional risk $R(\alpha_N)$ over the ensemble of training sets with N samples:

$$\text{MISE} = E_{\mathbf{D}_N}[R(\alpha_N)]$$

In this expression $R(\boldsymbol{\alpha}_N)$ is a function of the random variables $\mathbf{D}_N$.

Functional risk vs. MISE



Three learning problems

The previous formulation is very general and encompasses three well-known problems:

1. Classification or pattern recognition
2. Regression
3. Density estimation

Classification

It is an example of learning problem where

- the target's output y takes only two values $y \in \mathcal{Y} = \{0, 1\}$
- the class of hypothesis is made of *indicator functions*, that are functions which take only two values: zero and one.
- the cost function is

$$C(y, h(x, \alpha)) = \begin{cases} 0 & \text{if } y = h(x, \alpha) \\ 1 & \text{if } y \neq h(x, \alpha) \end{cases}$$

The functional risk in this case represents the probability of different answers between the target and the indicator function $h(x, \alpha)$.
The problem consists then in finding a function that minimizes the probability of classification error when the probability distribution is unknown but a training set is given.

Regression

It is an example of learning problem where

- the target's output y is a real value
- the class of hypothesis is made of real value functions
- the cost function is

$$C(y, h(x, \alpha)) = (y - h(x, \alpha))^2$$

It is well known in statistics that the function that minimize the functional risk is

$$h(x, \alpha^*) = \int y p(y|x) dy = \int y dF_{\mathbf{y}}(y|x)$$

The problem consists then in estimating the regression function when the probability distribution is unknown but a training set is given.

Density estimation

It is an example of learning problem where

- we consider only the input random variable $\mathbf{x}$
- the class of hypothesis is the set of probability densities $h(x, \alpha)$,
- the cost function is

$$C(h(x, \alpha)) = -\log(h(x, \alpha))$$

It is known that the generator density is the one that minimize the functional risk with the above cost function.
Density estimation boils down to minimizing a functional risk when the probability distribution is unknown but a training set is given.

Illustrative example

We consider a supervised learning problem where :

- The input is a scalar random variable $\mathbf{x} \in \Re$ with a uniform probability distribution over the interval $[-2, 2]$.
- The target is distributed according to a Gaussian distribution with mean x^3 and unit variance.
- The training set is made of $N = 100$ pairs.

The learning machine is characterized by the following three components:

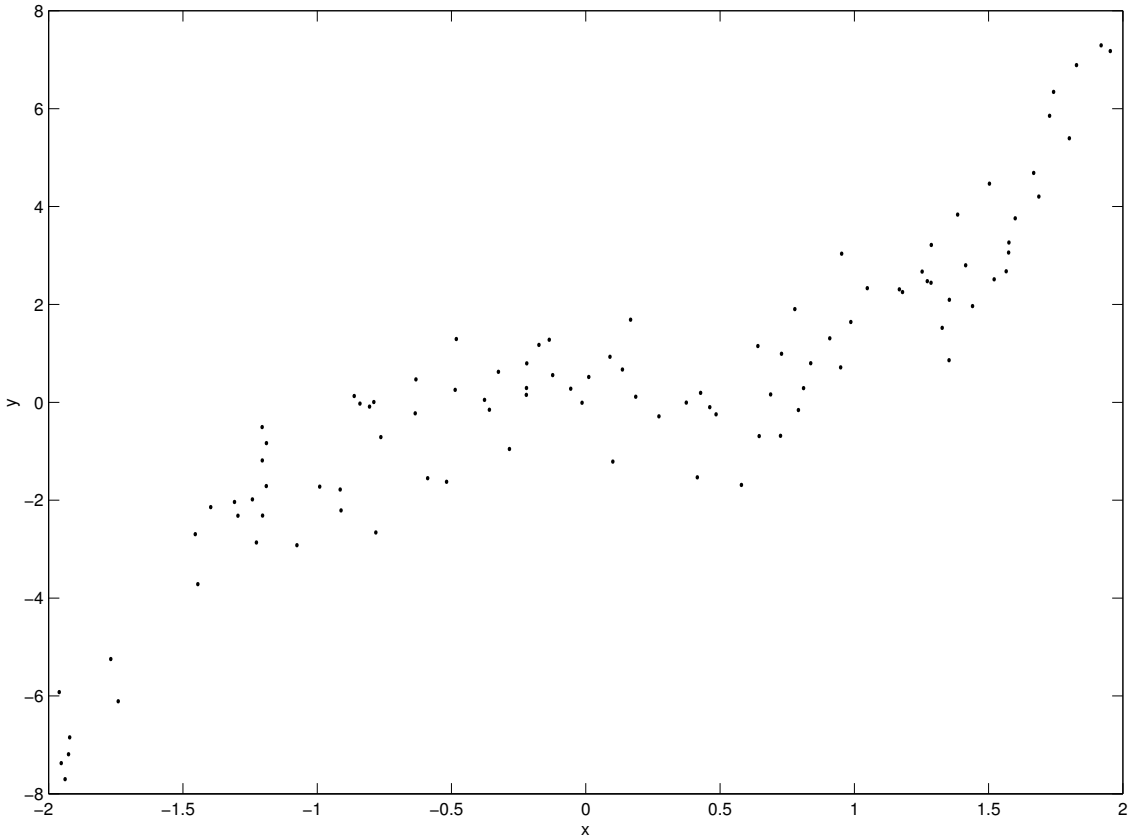
1. A class of hypothesis functions $h(x, \alpha) = \alpha x$ made of all the linear models passing through the origin. The class Λ is then the set of real numbers.
2. A quadratic cost $C(y, h(x)) = (y - h(x))^2$.
3. An algorithm of parametric identification based on the least-squares technique. The empirical risk is the quantity

$$R_{\text{emp}}(\alpha) = \sum_{i=1}^{100} (y_i - \alpha x_i)^2$$

If the knowledge on the joint distribution was available, it would be possible to compute also the risk functional as

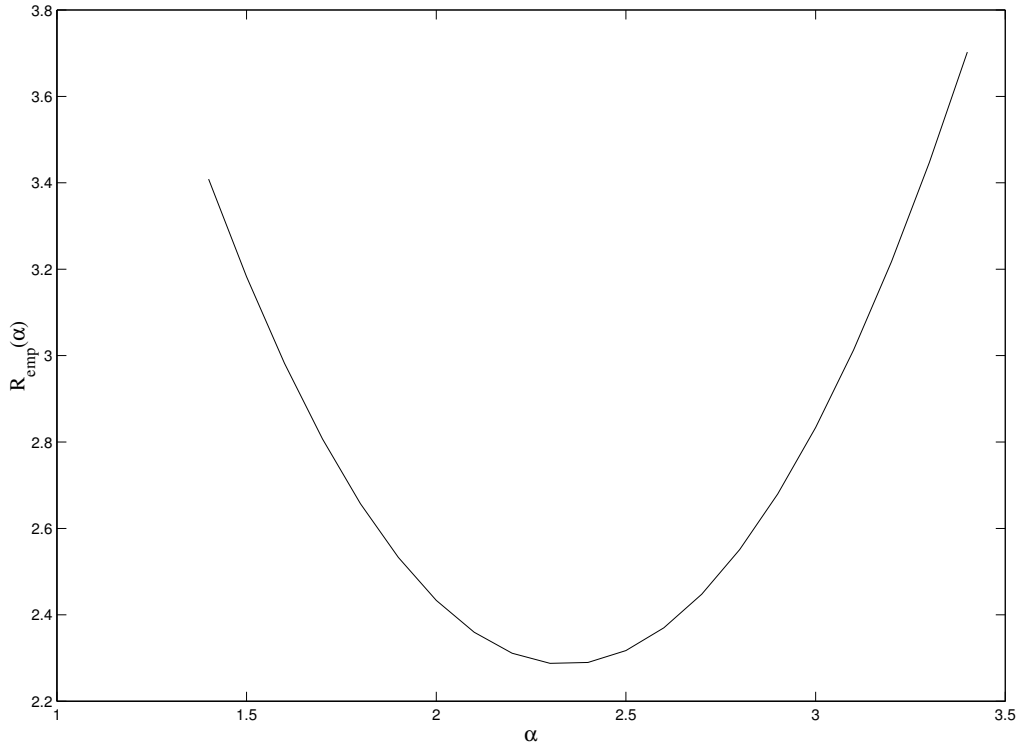
$$R(\alpha) = \frac{1}{4} \int_{-2}^{2} (x^3 - \alpha x)^2 dx + 1$$

Training set



Empirical risk

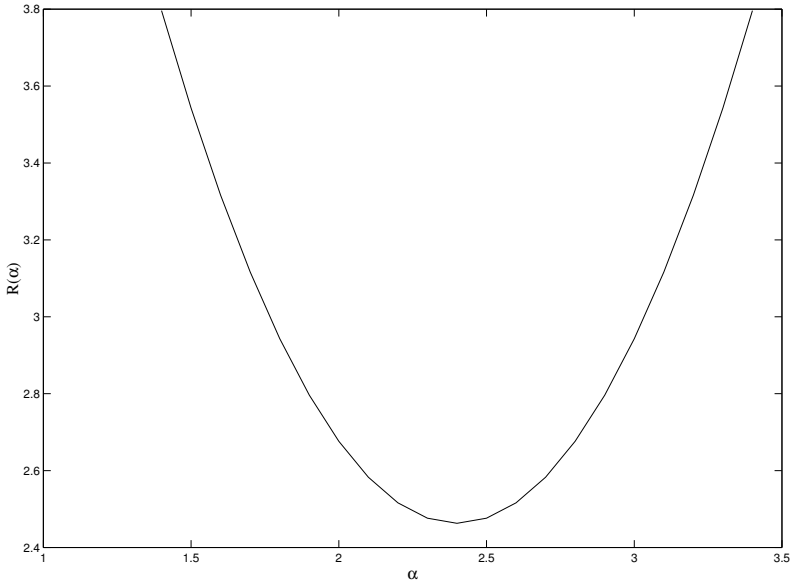
The empirical risk for the training set D_N (y-axis) vs. the parameter value (x-axis).



The minimum of the empirical risk is attained in $\alpha = 2.3272$.

Functional risk

The functional risk (y-axis) vs. the value of parameter α (x-axis).



The minimum of the functional risk is attained in $\alpha = 2.4$

Two interpretations of the learning problem

Hypothesis-based: it is the interpretation proposed by Vapnik. The goal of the *hypothesis-based* approach is to estimate the performance of the selected hypothesis α_N. The main assumption is that averaging over all possible training sets would be unnatural given the single realization available. Since the distribution of the data is not known, *hypothesis-based* methods search for distribution-free bounds. As a drawback, the results might be too conservative for a specific learning problem.

Algorithm-based: it focuses on the estimation of the Mean Integrated Squared Error. A learned hypothesis is seen as a function of the data D_N: since D_N is a random variable, the hypothesis is random as well and must be assessed averaging different realizations. It would be desirable to repeat several times the data generation and to run each time the learning algorithm. since the use of repeated realizations is not viable in a real learning problem resampling methods are employed.

The Vapnik's approach

- The Vapnik's approach proposes a comprehensive theory of the problem of learning and generalization.
- His work returns a large amount of results from the theoretical conditions for consistency of the learning process to constructive methods to select function estimators.
- In the following the functional risk notation is at first rewritten as

$$R(\alpha) = \int C(y, h(x, \alpha) dF_{\langle \mathbf{x}, \mathbf{y} \rangle}(x, y) = \int Q(z, \alpha) dF_{\mathbf{z}}(z) \qquad \alpha \in \Lambda$$

  where $z = \langle x, y \rangle$, $Q(z, \alpha) = C(y, h(x, \alpha))$, the probability measure $F_{\mathbf{z}}(\cdot)$ is unknown but an i.i.d. sample $z_1, \ldots, z_N$ is given.
- In the following the empirical risk is rewritten

$$R_{\text{emp}}(\alpha_N) = \frac{1}{N} \sum_{i=1}^{N} Q(z_i, \alpha_N)$$

Decomposition estimation/approximation

- Let us define with Λ^* the set of **all** possible single valued mappings $f : \mathcal{X} \to \mathcal{Y}$ and consider the quantity

$$\alpha^* = \arg \min_{\alpha \in \Lambda^*} R(\alpha)$$

  Thus, $R(\alpha^*)$ represents the absolute minimum rate of functional risk. In the classification case, the model α^* is called the *Bayes classifier* and $R(\alpha^*)$ the *Bayes error*.

- Let

$$\alpha_0 = \arg \min_{\alpha \in \Lambda} R(\alpha)$$

  be the hypothesis in the class Λ that minimizes the functional risk.

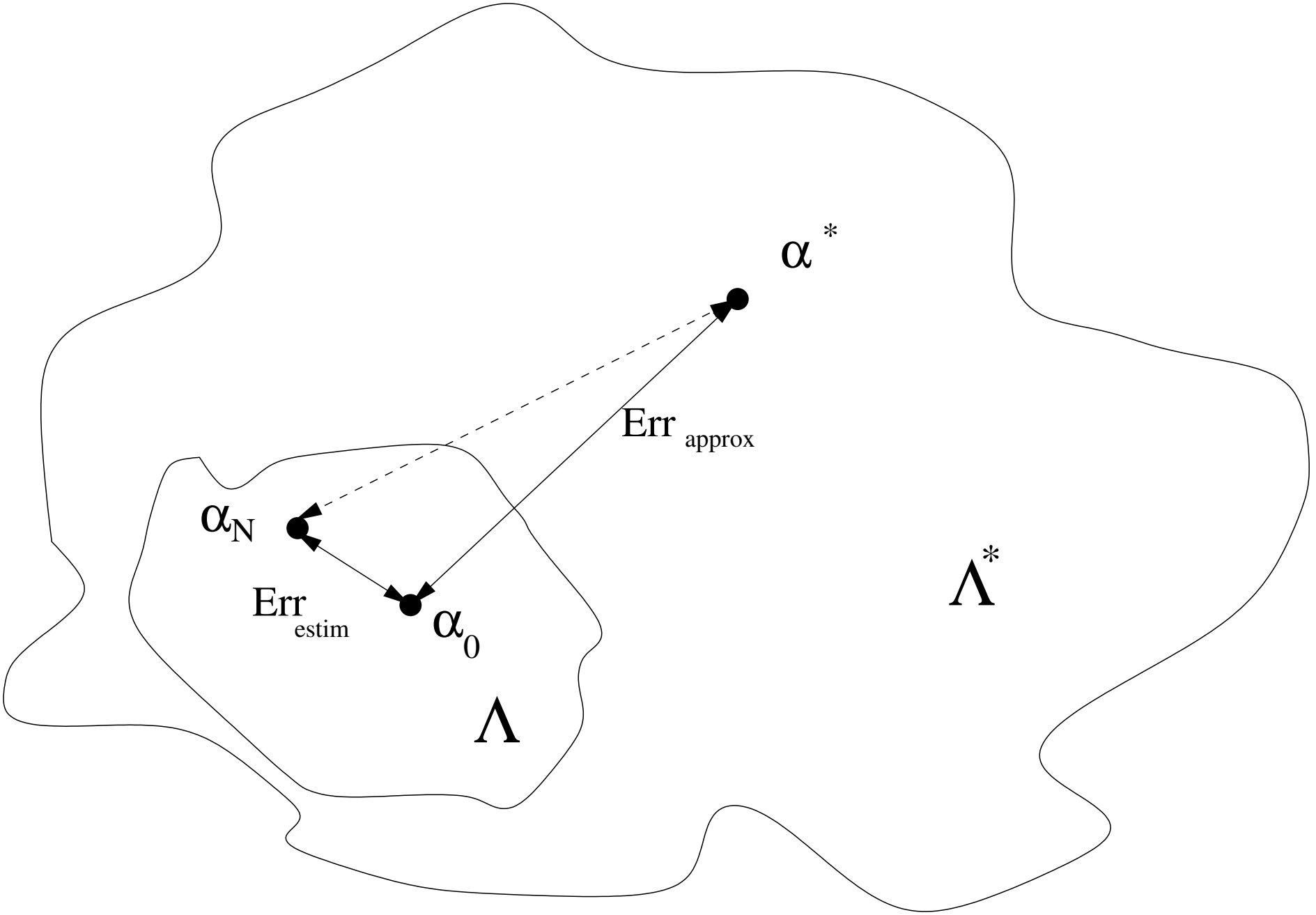
Decomposition estimation/approximation (II)

We can write the equality

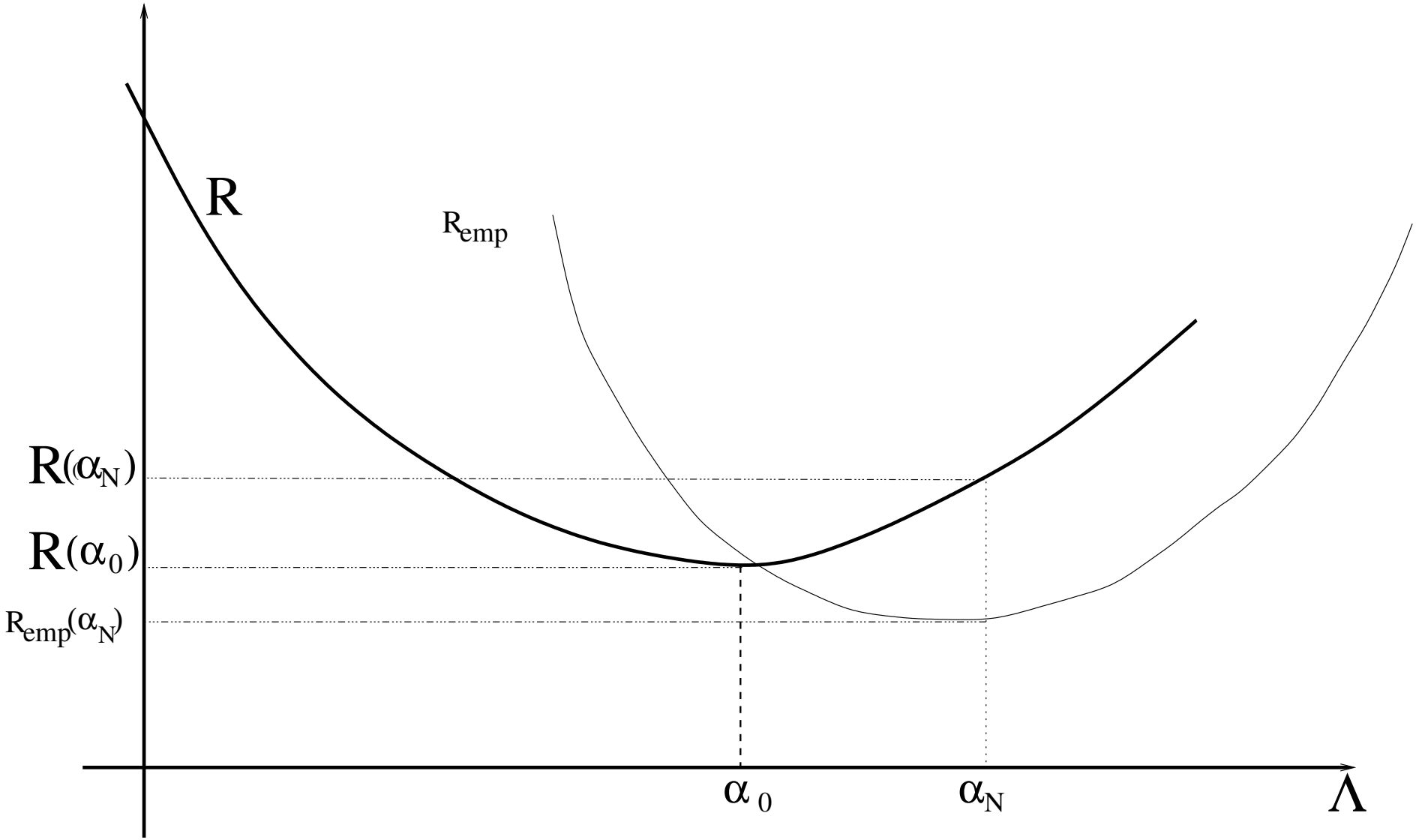
$$R(\alpha_N) - R(\alpha^*) = (R(\alpha_N) - R(\alpha_0)) + (R(\alpha_0) - R(\alpha^*)) =$$
$$= \mathrm{Err}_{\mathrm{estim}}(\alpha_N) + \mathrm{Err}_{\mathrm{approx}}(\alpha_N)$$

- It is common practice to define the first right-hand term as *estimation error* and the second term as *approximation error*. The approximation error is intrinsically related to the approximation capabilities of the class of hypothesis, while the estimation error represents the discrepancy between the best generalization error in the class and what is obtained from D_N.
- The size of Λ is a compromise: when the size of Λ is large, $R(\alpha_0)$ may be close to $R(\alpha^*)$, but the estimation error is probably large as well. If the size of Λ is too small, there is no hope to make the approximation error small

Decomposition estimation/approximation (III)



Functional and empirical risk



The ERM principle

- A key issue is the the convergence of the empirical risk $R_{\text{emp}}(\alpha_N)$ to the best functional risk into the class Λ, or in other terms the consistency of the Empirical Risk Minimization (ERM) principle.
- The ERM principle is consistent for the set of functions $Q(z, \alpha)$ and for the probability distribution $P_{\mathbf{z}}(z)$ if the following two sequences converge in probability to the same limit

$$R(\alpha_N) \xrightarrow[N\to\infty]{P} R(\alpha_0)$$

$$R_{\text{emp}}(\alpha_N) \xrightarrow[N\to\infty]{P} R(\alpha_0)$$

Sufficient condition

The following two relations hold between the quantities presented in the figure

Lemma 1 (Devroye 1988).

$$R(\alpha_N) - \inf_{\alpha \in \Lambda} R(\alpha) = R(\alpha_N) - R(\alpha_0) \leq 2 \sup_{\alpha \in \Lambda} |R_{emp}(\alpha) - R(\alpha)|$$

$$|R_{emp}(\alpha_N) - R(\alpha_N)| \leq \sup_{\alpha \in \Lambda} |R_{emp}(\alpha) - R(\alpha)|$$

From this lemma, we see that upper bounds for $\sup_{\alpha \in \Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$ provide us with upper bounds for three quantity simultaneously

1. the quantity $R(\alpha_N) - \inf_{\alpha \in \Lambda} R(\alpha)$ which returns the sub-optimality of the the model chosen by the ERM principle within the class $\alpha \in \Lambda$
2. the quantity $|R_{\text{emp}}(\alpha_N) - R(\alpha_N)|$ which returns the error committed when the empirical risk is used to estimate the functional risk of the selected model
3. the quantity $|R_{\text{emp}}(\alpha_N) - \inf_{\alpha \in \Lambda} R(\alpha)|$ which returns the error made when the empirical risk is used to estimate the functional risk of the best model in the class Λ

It can be shown that bounding $\sup_{\alpha \in \Lambda} |R_{\text{emp}}(\alpha) - R(\alpha)|$ is not only a sufficient but also a necessary condition for consistency of the ERM principle.

Key theorem of learning

Theorem 1 (Vapnik,Chervonenkis, 1991). *Let* $Q(z,\alpha)\ \alpha \in \Lambda$ *be a set of functions that satisfy the condition*

$$a \leq \int Q(z,\alpha) dP(z) \leq b$$

Condition necessary and sufficient for the ERM principle to be consistent is that the empirical risk $R_{emp}(\alpha)$ *converges uniformly to the actual risk* $R(\alpha)$ *over the set* $Q(z,\alpha)$, $\alpha \in \Lambda$ *that is*

$$\lim_{N\to\infty} \textit{Prob}\left\{\sup_{\alpha\in\Lambda}(R(\alpha) - R_{emp}(\alpha)) > \varepsilon\right\} = 0 \qquad \forall\varepsilon > 0$$

This theorem replaces the problem of consistency with the problem of uniform convergence.

Beyond classical statistics

The uniform convergence is guaranteed by the Law of Large Numbers in the trivial case of the set of functions $Q(z, \alpha)$ containing only one element. For a real-valued bounded function

$$a \leq Q(z, \alpha) \leq b$$

using Hoeffding's inequalities we have

$$\text{Prob}\left\{\left|\int Q(z,\alpha)dP(z) - \frac{1}{N}\sum_{i=1}^{N} Q(z_i,\alpha)\right| > \varepsilon\right\} < \exp\left\{-\frac{2\varepsilon^2 N}{(b-a)^2}\right\}$$

Beyond classical statistics (II)

The generalization is easy for the case where $Q(z, \alpha)$ has a finite number of K elements:

$$\text{Prob}\left\{\sup_{1\leq k\leq K}\left|\int Q(z,\alpha)dP(z) - \frac{1}{N}\sum_{i=1}^{N} Q(z_i,\alpha)\right| > \varepsilon\right\} <$$

$$K\exp\left\{-\frac{2\varepsilon^2 N}{(b-a)^2}\right\} = \exp\left\{\left(\frac{\ln K}{N} - \frac{2\varepsilon^2}{(b-a)^2}\right)N\right\}$$

In order to obtain uniform convergence for any ε, the expression

$$\lim_{N\to\infty}\frac{\ln K}{N} = 0$$

has to be true.

A similar relation will be indicative for uniform convergence also in the case of an infinite set.

Uniform convergence

- Consider the sequence of random variables

$$\xi_N = \sup_{\alpha \in \Lambda} \left(R(\alpha) - R_{\text{emp}}(\alpha) \right) = \sup_{\alpha \in \Lambda} \left(\int Q(z, \alpha) dF(z) - \frac{1}{N} \sum_{i=1}^{N} Q(z_i, \alpha) \right)$$

  where the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, has an infinite number of elements.

- In contrast to the cases with a finite number of elements the sequence of random variables ξ_N for a set with an infinite number of elements *does not necessarily converge to zero*.

- The problem becomes: *To describe the properties of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, under which the sequence of random variables ξ_N converges in probability to zero.*

Entropy of a set of functions

- Let $Q(z, \alpha)$, $\alpha \in \Lambda$, be a set of indicator functions. Consider a sample D_N.Λ contains infinitely many functions, only a finite number of clusters of functions is distinguishable for a given sample D_N.
- The idea is that, even if a set
- Let us characterize the diversity of the set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, on the dataset D_N by the quantity $\mathcal{N}^{\Lambda}(D_N)$ that evaluates how many different separations of the given sample can be done using functions from the family $Q(z, \alpha)$, $\alpha \in \Lambda$.
- Note that $\mathcal{N}^{\Lambda}(\mathbf{D}_N)$ is a random variable since $\mathbf{D}_N$ is a random variable.
- The quantity

$$H^{\Lambda}(N) = E \ln \mathcal{N}^{\Lambda}(\mathbf{D}_N)$$

  is called the *entropy* of the set of functions on the given data.

Entropy and convergence

Theorem 2. *A necessary and sufficient condition for the two-sided uniform convergence of the functional risk to the empirical risk is that*

$$\lim_{N\to\infty} \frac{H^{\Lambda}(N)}{N} = 0$$

- In other words, the ratio of the VC entropy to the number of observations should decrease to zero with increasing number of observations.
- This is a sufficient condition for the consistency of the ERM principle (necessary and sufficient conditions are given by a slightly different construction). Note that this condition depends on the underlying probability distribution $F_{\mathbf{z}}(\cdot)$.
- Note that the entropy of functions has taken the place of the number of function in the finite case.

Other measures of capacity

Consider two new concepts that are constructed on the basis of $\mathcal{N}^{\Lambda}(D_N)$.

Annealed VC entropy: it is defined as $H^{\Lambda}_{ann}(N) = \ln E[\mathcal{N}^{\Lambda}(\mathbf{D}_N)]$.

Growth function: it is defined as

$$G^{\Lambda}(N) = \ln \max_{D_N} \mathcal{N}^{\Lambda}(D_N)$$

It can be shown that

$$H^{\Lambda}(N) \leq H^{\Lambda}_{ann}(N) \leq G^{\Lambda}(N)$$

Rate of convergence

Definition 1 (Fast convergence). *We say that the asymptotic rate of convergence of the empirical risk to the functional risk is* fast *if for any $N > N_0$, the exponential bound*

$$\textit{Prob}\{R(\alpha) - R(\boldsymbol{\alpha}_N) > \varepsilon\} < e^{-2c\varepsilon^2 N}$$

holds true, where $c > 0$ is some constant.

It can be shown that

Theorem 3. *Sufficient condition for a fast rate of convergence is that*

$$\lim_{N\to\infty} \frac{H^{\Lambda}_{ann}(N)}{N} = 0$$

Note that this condition depends on the underlying probability distribution $F_{\mathbf{z}}(\cdot)$.

Universal fast convergence

- The results on consistency and fast convergence presented before are dependent on the (unknown) underlying probability distribution $F(\cdot)$.
- What makes the Vapnik approach a milestone in learning theory was the ability to discover a theoretical result independent of the probability measure(i.e. independent of the problem to be solved). He showed that:

Theorem 4. *Necessary and sufficient conditionfor consistency of ERM for any probability measure is*

$$\lim_{N\to\infty} \frac{G^{\Lambda}(N)}{N} = 0$$

If this condition holds true, then the rate of convergence is fast.

Distribution independent Bound

Vapnik proved that in the pattern recognition case

$$\text{Prob}\left\{\sup_{\alpha\in\Lambda}(R(\alpha)-R_{\text{emp}}(\alpha))>\varepsilon\right\}\leq 4\exp\left\{\left(\frac{G^{\Lambda}(2N)}{N}-\varepsilon^2\right)N\right\}$$

This means that, provided $G^{\Lambda}(N)$ does not grow linearly in N, it is actually possible to make nontrivial statements about the functional risk $R(\alpha_N)$ on the basis of the empirical risk $R_{\text{emp}}(\alpha_N)$.

Confidence interval

- If we specify the probability with which we want the bound to hold, we can get a confidence interval, which tells us how close the risk should be to the empirical risk.
- By setting the right-hand side of the bound to $\delta > 0$, and then solving for ε, we get that with a probability $1 - \delta$.

$$R(\alpha_N) \leq R_{\text{emp}}(\alpha_N) + \frac{\sqrt{\mathcal{E}}}{2}$$

where

$$\mathcal{E} = 4\frac{G^{\Lambda}(2N) - \ln(\delta/4)}{N}$$

- We call the right-hand side, the *guaranteed risk*.

VC dimension

- It summarizes the behaviour of the growth function $G^{\Lambda}(N)$ with a single number.
- If Λ is as rich as possible, so that for any size N, the points can be chosen such that an hypothesis function $h(\cdot, \alpha)$, $\alpha \in \Lambda$ can separate them in all 2^N possible ways, then $G_{\Lambda}(N) = N \ln 2$. If this is the case, the convergence does not take place and the learning is not successfull.
- Vapnik and Chervonenkis showed that either the relation $G^{\Lambda}(N) = N \ln 2$ holds true for all N, or there exists some maximal N for which this relation is satisfied. This maximal N is called the VC dimension and is denoted by h.
- By construction, the VC dimension is the maximal number of points which can be shattered by functions in Λ.

Theorem 5. *Any growth function either satisfies the equality*

$$G^{\Lambda}(N) = N \ln 2$$

or is bounded by the inequality

$$G^{\Lambda}(N) \leq h \left(\ln \frac{N}{h} + 1 \right)$$

where h is an integer such that when $N = h$

$$G^{\Lambda}(h) = h \ln 2,$$

$$G^{\Lambda}(h+1) < (h+1) \ln 2$$

We will say that the VC dimension of a set of indicator functions $Q(z, \alpha)$ is infinite if the growth function is linear.
We will say that the VC dimension of a set of indicator functions $Q(z, \alpha)$ is finite and equals h if the corresponding growth function is bounded by a logarithmic function with coeficient h.

VC dimension and number of parameters

- The finiteness of the the VC dimension is a necessary and sufficient condition for distribution independent consistency of ERM learning machines
- The VC dimension of the set of *linear functions* with $n+1$ parameters is equal to $h = n+1$. Note that for the set of linear functions the VC dimension equals the number of free parameters but that **in the general case this is not true.**
- The VC dimension of the set of functions

$$h(x, \alpha) = \sin \alpha x, \qquad \alpha \in \mathbb{R}$$

  is infinite.
- Generally speaking, the VC dimension of a set of functions can be either larger than or smaller than the number of parameters.
- The VC dimension of the set of functions (rather than the number of parameters) is responsible for the generalization ability of

Structural risk minimization

- Consider the confidence interval for a set of functions $Q(z, \alpha)$, $\alpha \in \Lambda$, having a finite VC dimension h. Then the second summand on the right-hand side of the inequality is

$$\mathcal{E}(h, N, \delta) = 4\frac{h\left(\ln\frac{2N}{h} + 1\right) - \ln(\delta/4)}{N}$$

- When N/h is large, the $\mathcal{E}$ term is small. The actual risk is then close to the empirical risk. This means that a small value of the empirical risk gurantees a small value of the (expected) risk.
- If N/h is small, a small $R_{\text{emp}}(\alpha_N)$ does not guarantee a small value of the actual risk. In this case, to minimize the actual risk $R(\alpha_N)$ one has to minimize the right-hand side of the confidence interval simultaneously over both terms.
- Note that the first term depends on a specific function $h(\cdot, \alpha_N)$ while the second term depends on the VC dimension of the whole set of functions.

Structural risk minimization

- To minimize the right hand side of the bound of risk, one has to make the VC dimension a controlling variable.
- The SRM principle inductive principle is intended to minimize the risk functional with respect to the empirical risk and the confidence interval.
- Let the set S of functions $Q(z, \alpha)$, $\alpha \in \Lambda$ be provided with a *structure* consisting of nested subsets of functions $S_s = \{Q(z, \alpha), \alpha \in \Lambda_s\}$ such that $S_1 \subset S_2 \subset \cdots \subset S_K$ and the VC dimension h_k of each set S_k of functions is finite.
- For a given set D_N, the SRM principle chooses the function $Q(z, \alpha_k)$ minimizing the empirical risk in s_k for which the guaranteed risk (right-hand side) is minimal.
- The SRM principle defines a trade-off between the quality of the approximation of the given data and the complexity of the approximating function.

Structural risk minimization

Let α_N^s be the hypothesis that minimizes the empirical risk in the set $\alpha \in \Lambda_s$:

$$\alpha_N^s = \arg\min_{\alpha \in \Lambda_s} R_{\text{emp}}(\alpha)$$

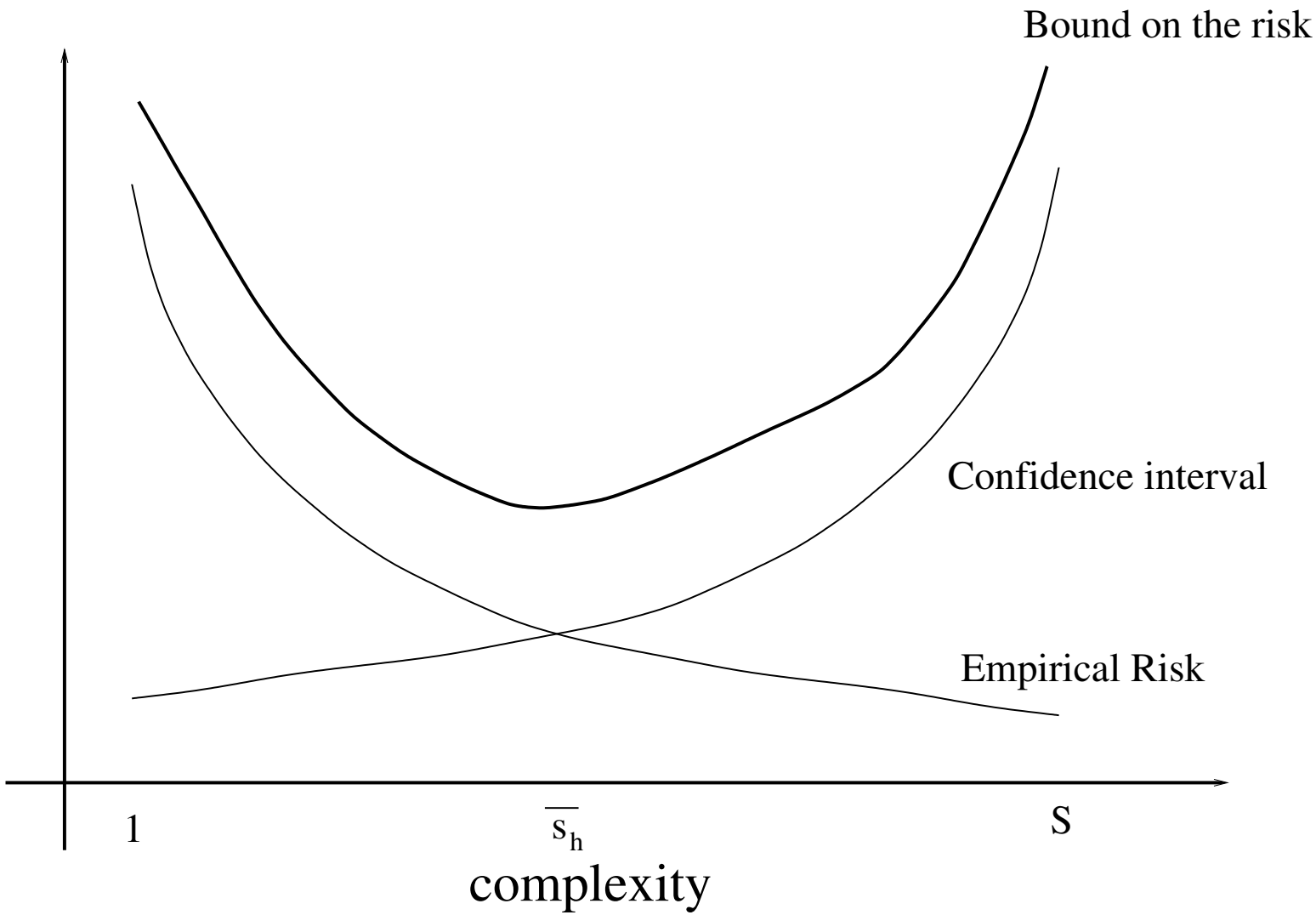
It follows that with probability $1 - \delta$, the risk $R(\alpha_N^s)$ is bounded as

$$R(\alpha_N^s) \leq R_{\text{emp}}(\alpha_N^s) + \mathcal{E}(h, N, \delta) = \text{SRM}(s)$$

The term SRM(s) is made of the sum of the empirical risk and a term which increases as the VC dimension of the class increases.
The SRM method chooses the hypothesis $\alpha_N^{\bar{s}}$ where

$$\bar{s} = \arg\min \text{SRM}(s)$$

Structural risk minimization



Universal consistency of SRM

Definition 2. *An hypothesis α_N is called universally strongly consistent if*

$$\lim_{N\to\infty} R(\alpha_N) = R(\alpha^*)$$

for any distribution of $\langle \mathbf{x}, \mathbf{y} \rangle$.

Theorem 6. *Let $\Lambda_1, \Lambda_2, \ldots$ be a sequence of classes of hypothesis such that for any distribution of $\langle \mathbf{x}, \mathbf{y} \rangle$,*

$$\lim_{s\to\infty} \inf_{\alpha\in\Lambda_s} R(\alpha) = R(\alpha^*)$$

Assume also that the VC dimensions $h_1, h_2, \ldots$ are finite and satisfy

$$\sum_{s=1}^{\infty} e^{-h_s} < \infty$$

then the hypothesis $\alpha_N^{\bar{s}}$ selected by structural risk minimization is strongly universally consistent.

Separating hyperplane

Consider a binary classification problem.

We call a hyperplane

$$\beta^* x - \beta_0 = 0, \qquad \|\beta\| = 1$$

a Δ-margin hyperplane if it classifies vectors x as follows

$$y = \begin{cases} 1 & \text{if } \beta^* x - \beta_0 \geq \Delta \\ -1 & \text{if } \beta^* x - \beta_0 \geq Delta \end{cases}$$

Support Vector Machines provides an algorithm for finding a maximal margin hyperplane in the separable case.

Separating hyperplane

The followin theorem holds

Theorem 7. *Let the input vectors $x \in \mathbb{R}^n$ belong to a sphere of radius R. Then the set of Δ-margin separating hyperplanes has VC dimension h which is bounded by the inequality*

$$h \leq \min\left(\left[\frac{R^2}{\Delta^2}\right], n\right) + 1$$

It follows that

- the VC dimension of a separating hyperplane can be less than $n + 1$.
- the VC dimension of the set of Δ-margin separating hyperplanes with Δ large is small.

Science vs. non science

Question: Is there a formal way to distinguish between scientific and non scientific approaches?
Consider for example, meteorology and astrology. What is the formal difference between them?

- Is it the complexity of the models?
- Is it the predictive ability of their models?
- Is it their use of mathematics?
- Is it in the level of formalism?

None of the aboves can be useful to make a distinction.

Popper's theory

- In the 1930s, K. Poper suggested his famous criterion for demarcation between scientific and nonscientific theories.
- According to Popper, a necessary condition for a scientific discipline is the feasibility of its falsification.
- By falsification, Popper means the existence of a collection of particular assertions which canot be explained by the given theory although they fall into its domain.
- If there is no example that can falsify the theory of astrology, then astrology according to Popper should be considered as non scientific.
- The notion of entropy of of Λ summarizes this notion in the case of statistical learning. If $H^{\Lambda}(N) = 2^N$ then the set of functions Λ is such that almost any sample D_N (of arbitrary size N) can be separated in all possible ways by functions of this set.
- We call this learning machine nonfalsifiable because it can give a