Pattern recognition

Learning theory - basic concepts

- Chapiter 1 -

The knowledge of the probabilistic model is replaced by a training set of data $\mathcal{A}_n$:

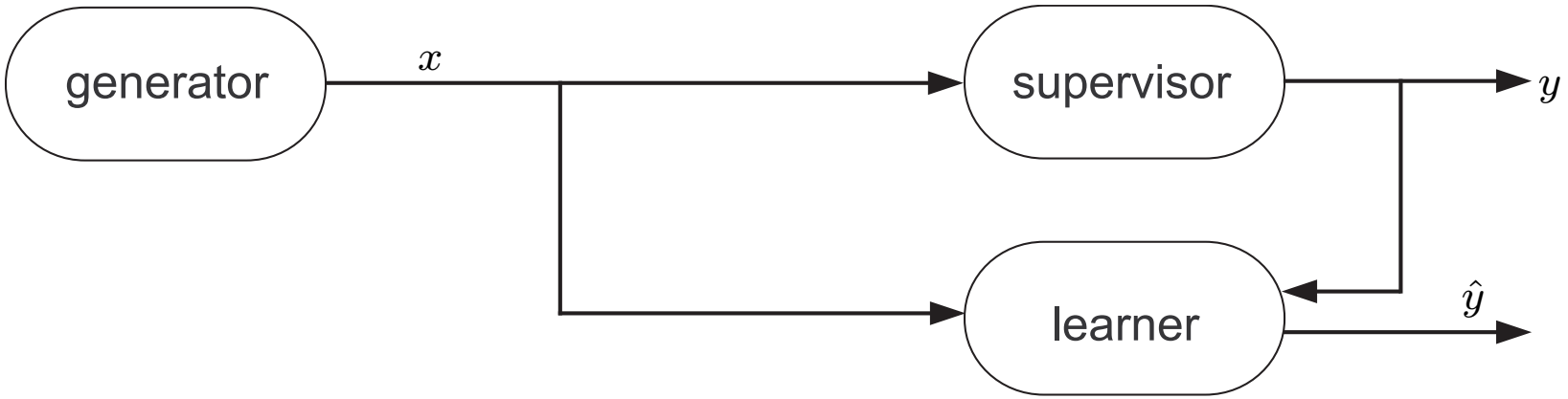
$$\mathcal{A}_n = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}.$$

Building a decision rule consist in searching for a partition of the observation space $\mathcal{X}$. The partition must be optimal according to a chosen performance criteria.

Two main approaches may be found in the literature :

1. Choice of a decision rule structure and optimisation of the characteristic parameters according to a chosen criteria.
2. Direct use of the training set to take the decision.

The training model is composed of 3 elements :



1. Generator : $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^l$, random variables i.i.d.
2. Supervisor : $Y \in \mathcal{Y} \subset \mathbb{R}$, random variables
3. Learner : represented by $d(\boldsymbol{x}; \theta) \in \mathcal{D}$

– **Polynomial of degre p**

$$d(\boldsymbol{x};\boldsymbol{a}) = \sum_{\substack{i_1,\ldots,i_l \in \mathbb{N} \\ i_1+\ldots+i_l \leq p}} a_{i_1,\ldots,i_l}\, x[1]^{i_1} \ldots x[l]^{i_l}$$

..., and other decomposition on Fourier Basis, Harr Basis...

– **Splines**

$$d(\boldsymbol{x};c) \in \mathcal{L}^2(\mathbb{R}^l) \text{ tel que } d' \in \mathcal{L}^2(\mathbb{R}^l), \|d'\|^2 \leq c$$

– **Nadaraya-Watson**

$$d(\boldsymbol{x};\sigma) = \frac{\sum_{i=1}^n y_i\, K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}{\sum_{i=1}^n K_\sigma(\boldsymbol{x},\boldsymbol{x}_i)}$$

– **MLP, RBF, ...**

$$d(\boldsymbol{x};\boldsymbol{a},\boldsymbol{\theta}) = \sum_k a_k\; g_k(\boldsymbol{x};\boldsymbol{\theta}_k)$$

Find a linear classifier :

$$d(\boldsymbol{x}; V, \nu_0) = V^T \boldsymbol{x} + \nu_0 \underset{D_1}{\overset{D_0}{\lessgtr}} 0$$

Problem
Estimation of V and ν_0. ???

Perceptron

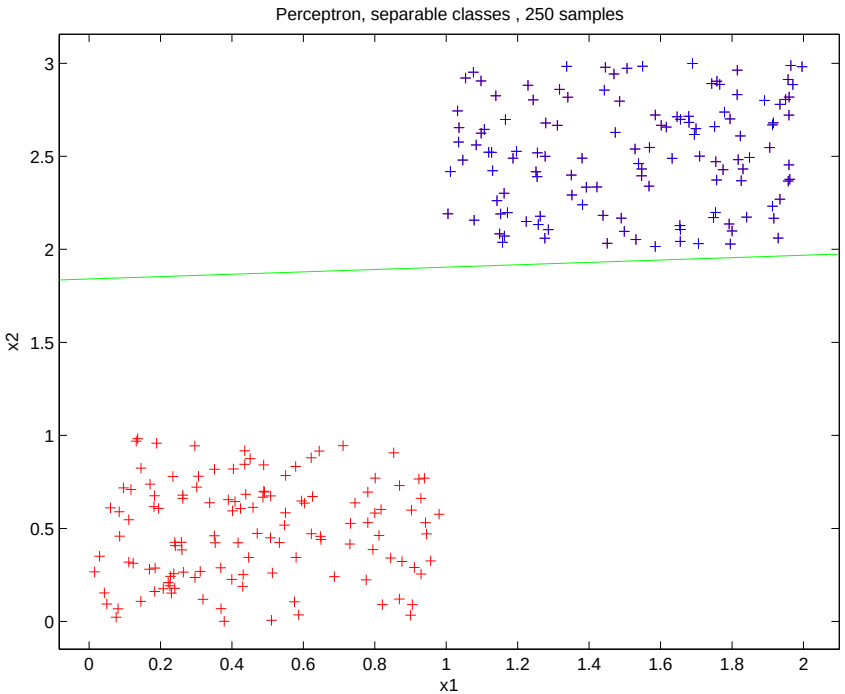
First approach
Let us rewrite $d(\boldsymbol{x}; V, \nu_0)$ as :

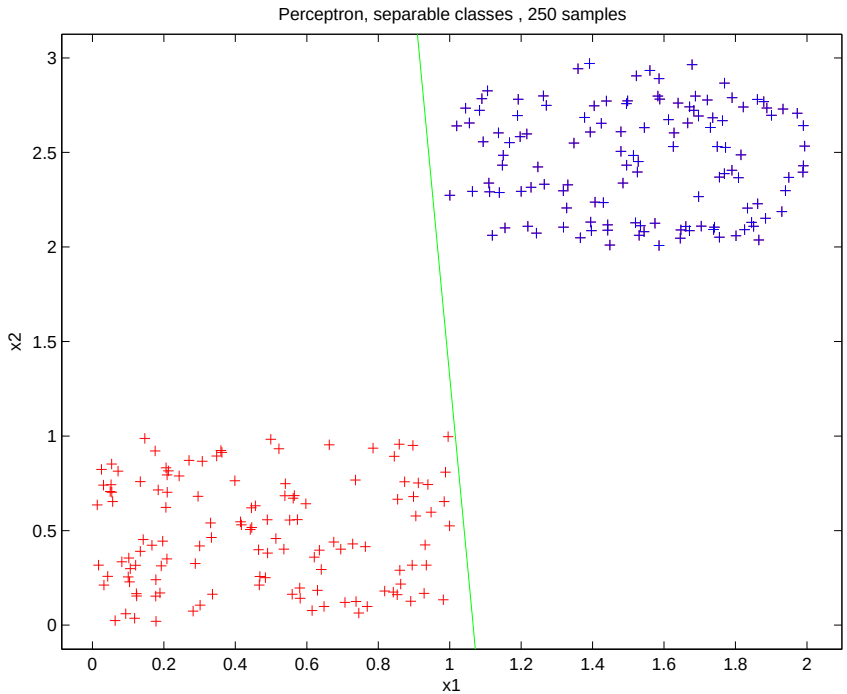
$$d(\boldsymbol{x}; V, \nu_0) = y\left(V^T \boldsymbol{x} + \nu_0\right) > 0 \quad \forall \boldsymbol{x} \in \mathcal{X}$$

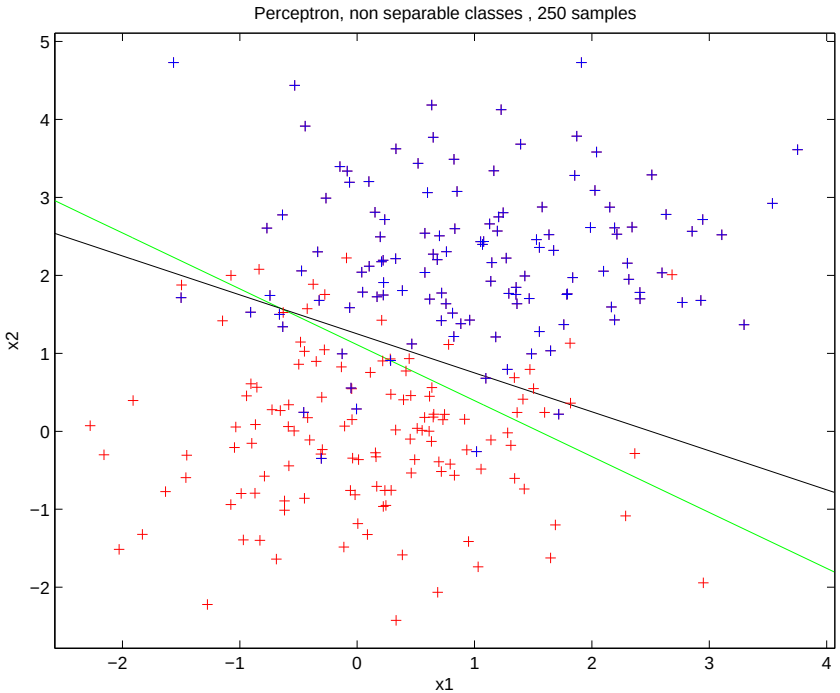
Assuming ω_0 and ω_1 are separable.

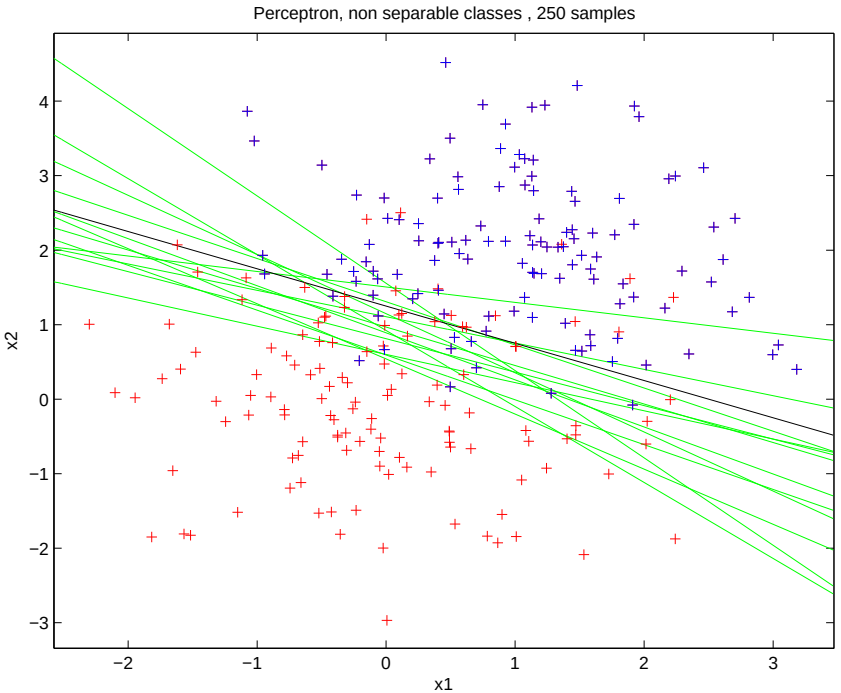
Perceptron - Algorithm

Until V and ν_0 are stable
if $y_i \left(V^T \boldsymbol{x_i} + \nu_0\right) > 0$ do nothing,
if $y_i \left(V^T \boldsymbol{x_i} + \nu_0\right) < 0$ then
$V' = V + c\boldsymbol{x_i} y_i \quad \nu_0' = \nu_0 + c y_i$









Property

Convergence is proved only if classes are separable.

Target

Find within $\mathcal{D} = \{d(\boldsymbol{x}, \theta) \ : \ \theta \in \Theta\}$ the fonction which gives the best approximation of y according to a risk functional which can be expressed as

$$J(d) = \int Q(d(\boldsymbol{x}, \theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

where Q expresses the cost associated to each couple $(\boldsymbol{x}, y)$.

Example of a cost function : error probability

To develop a decision rule minimizing error probability, the risk is expressed as :

$$P_e(d) = \int \mathbb{1}_{d(\boldsymbol{x},\theta)\neq y}\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

where $\mathbb{1}$ is the indicatrice function.

– Quadratic cost

$$Q(\boldsymbol{x}, y) = (y - d(\boldsymbol{x}; \theta))^2 \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{E}(y \,|\, \boldsymbol{x})$$

– Absolute cost

$$Q(\boldsymbol{x}, y) = |y - d(\boldsymbol{x}; \theta)|$$

– Cross Entropy

$$Q(\boldsymbol{x}, y) = -y \log(d(\boldsymbol{x}; \theta)) - (1 - y) \log(1 - d(\boldsymbol{x}; \theta)) \quad \rightarrow \quad d^*(\boldsymbol{x}; \theta) = \mathrm{P}(y = 1 \,|\, \boldsymbol{x})$$

The aim is to minimize the following functional :

$$J(d) = \int Q(d(\boldsymbol{x};\theta), y)\, p(\boldsymbol{x}, y)\, d\boldsymbol{x}\, dy,$$

the density $p(\boldsymbol{x}, y)$ is unknown.

Minimization of empirical risk (MRE)

The minimization of $J(d)$ is done by plugging an estimator : the empirical risk

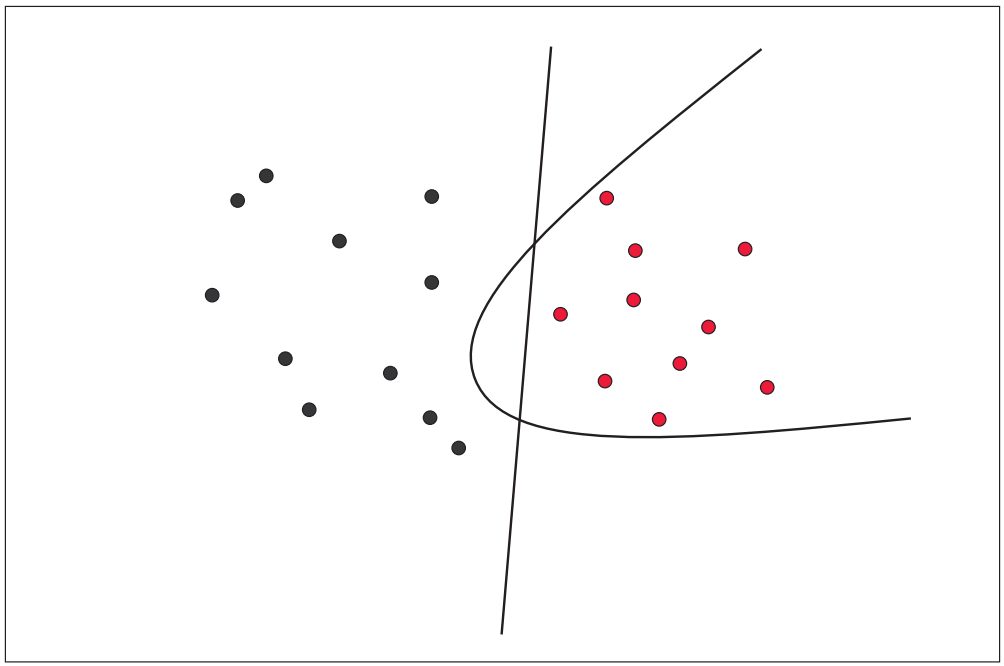
$$J_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} Q(d(\boldsymbol{x}_k;\theta), y_k),$$

$J_{emp}(d)$ can be estimated using the data of the training set $\mathcal{A}_n$.

Empirical Probability of error :
The empirical risk that corresponds to the probability of error depends on the number of classification errors made by $d(\boldsymbol{x};\theta)$ on the training data set $\mathcal{A}_n$

$$P_{emp}(d) = \frac{1}{n} \sum_{k=1}^{n} \mathbb{1}_{d(\boldsymbol{x}_k;\theta) \neq y_k}.$$

Problem : Two gaussien classes ω_0 et ω_1 in $\mathbb{R}^2$, with different mean and covariance - the training data set is composed of 10 samples for each class.



Which border should we choose ?
Which conclusions can we drawn from the fact that $\hat{P}_e(\text{linéaire}) = 5\%$ while $\hat{P}_e(\text{quadratique}) = 9\%$?

Let define $d^* = \arg\min J(d)$ the minimum risk decision rule.
Let denote $d_n^* = \arg\min_{d\in\mathcal{D}} J_{emp}(d)$ the decision rule obtained by minimizing the empirical risk on the functional class $\mathcal{D}$ based on the data set $\mathcal{A}_n$.

Definition (Estimation error)

It is the difference in performance between the best rule in $\mathcal{D}$ and the one obtained at the end of learning process :

$$J_{estim} = J(d_n^*) - \inf_{d\in\mathcal{D}} J(d)$$

▷ *relevance of empirical criteria and performance of the algorithm*

Definition (Approximation error)

It is the difference in performance between the optimal decision rule d^* and the best in $\mathcal{D}$:

$$J_{approx} = \inf_{d\in\mathcal{D}} J(d) - J(d^*)$$

▷ *Choice of class $\mathcal{D}$*

Learning

The objective of learning method is to minimize the modeling error, defined by :

$$J_{mod}(d_n^*) = J(d_n^*) - J(d^*).$$

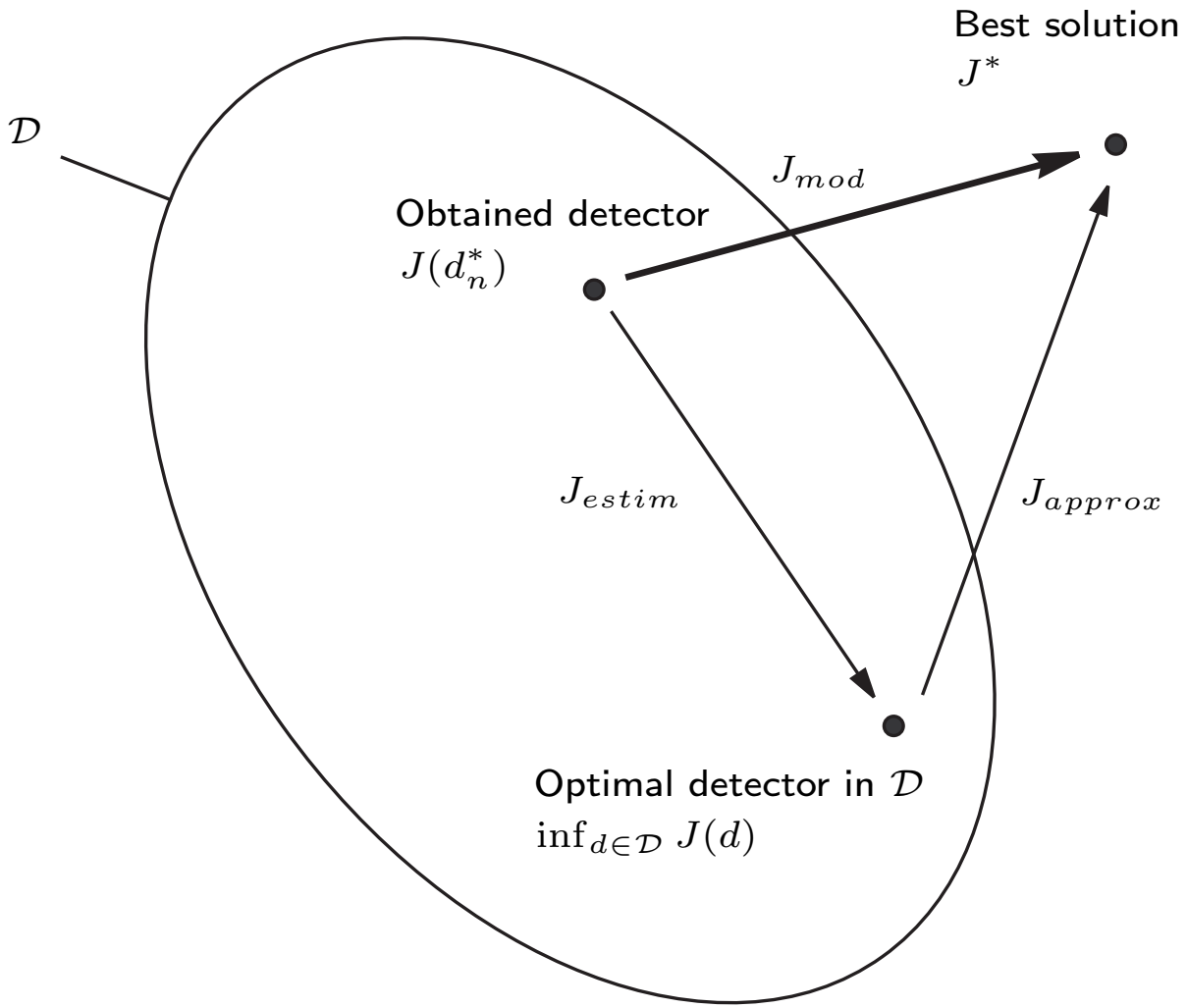
There are two different types of contributions in this error :

$$J_{mod}(d_n^*) = \underbrace{\left(J(d_n^*) - \inf_{d \in \mathcal{D}} J(d) \right)}_{J_{estim}} + \underbrace{\left(\inf_{d \in \mathcal{D}} J(d) - J(d^*) \right)}_{J_{approx}}.$$

The minimization of J_{mod} is based on the search for a compromise between these two opposing terms :

- increasing the number of tests in $\mathcal{D}$ leads to increase J_{estim}
- increasing the number of tests in $\mathcal{D}$ leads to decrease J_{approx}

and vice versa.


Best solution
J^*
$\mathcal{D}$
J_{mod}
Obtained detector
$J(d_n^*)$
J_{estim}
J_{approx}
Optimal detector in $\mathcal{D}$
$\inf_{d\in\mathcal{D}} J(d)$

1. *Is the goal achievable ?*

$\rightarrow$ *Consistency of the decision rule*

$\rightarrow$ *Consistency of the induction principle*

$\rightarrow$ *Convergence speed*

2. *: If Yes - How to do this ?*

Within the considered class of functional $\mathcal{D}$, one can expect that there exist a sequence of optimal detectors $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ according to the chosen criteria such that $J(d_n^*)$ can be made arbitrarily close to $\inf_{d\in\mathcal{D}} J(d)$ when n tends to infinity.

Définition (Consistency and strong consistency)

Given a data base $\mathcal{A}_n$, a sequence of optimal detectors $\{d_n^*(\boldsymbol{X};\theta)\}_{n>0}$ according to the chosen criteria is said to be consistent for a probability law $p(\boldsymbol{x},y)$ if :

$$\lim_{n\to\infty} \mathrm{E}\{J(d_n^*;\mathcal{A}_n)\} = \inf_{d\in\mathcal{D}} J(d).$$

It is said that the sequence is strongly consistent if, with probability equal to 1 :

$$\lim_{n\to\infty} J(d_n^*;\mathcal{A}_n) = \inf_{d\in\mathcal{D}} J(d).$$

Two cases : (Strong) Consistency can be satisfied :

- a single density law $p(\boldsymbol{x}, y)$,
- for any probability law.

Définition (Universel consistency)

A sequence $\{d_n^*(\boldsymbol{X}; \theta)\}_{n>0}$ is said to be (strongly) universally consistent if it is (strongly) consistent for any probability law $p(\boldsymbol{x}, y)$.

This property was first observed in 1977 by Stone in the method of *k-nearest neighbors*, provided that the parameter k grows slower than n the size of the learning set. Since then, it has been shown that other decision rules met this property :

– *regular kernel functions,*

– *some generalized linear detectors,*

– *Adaboost*

– *(. . .)*

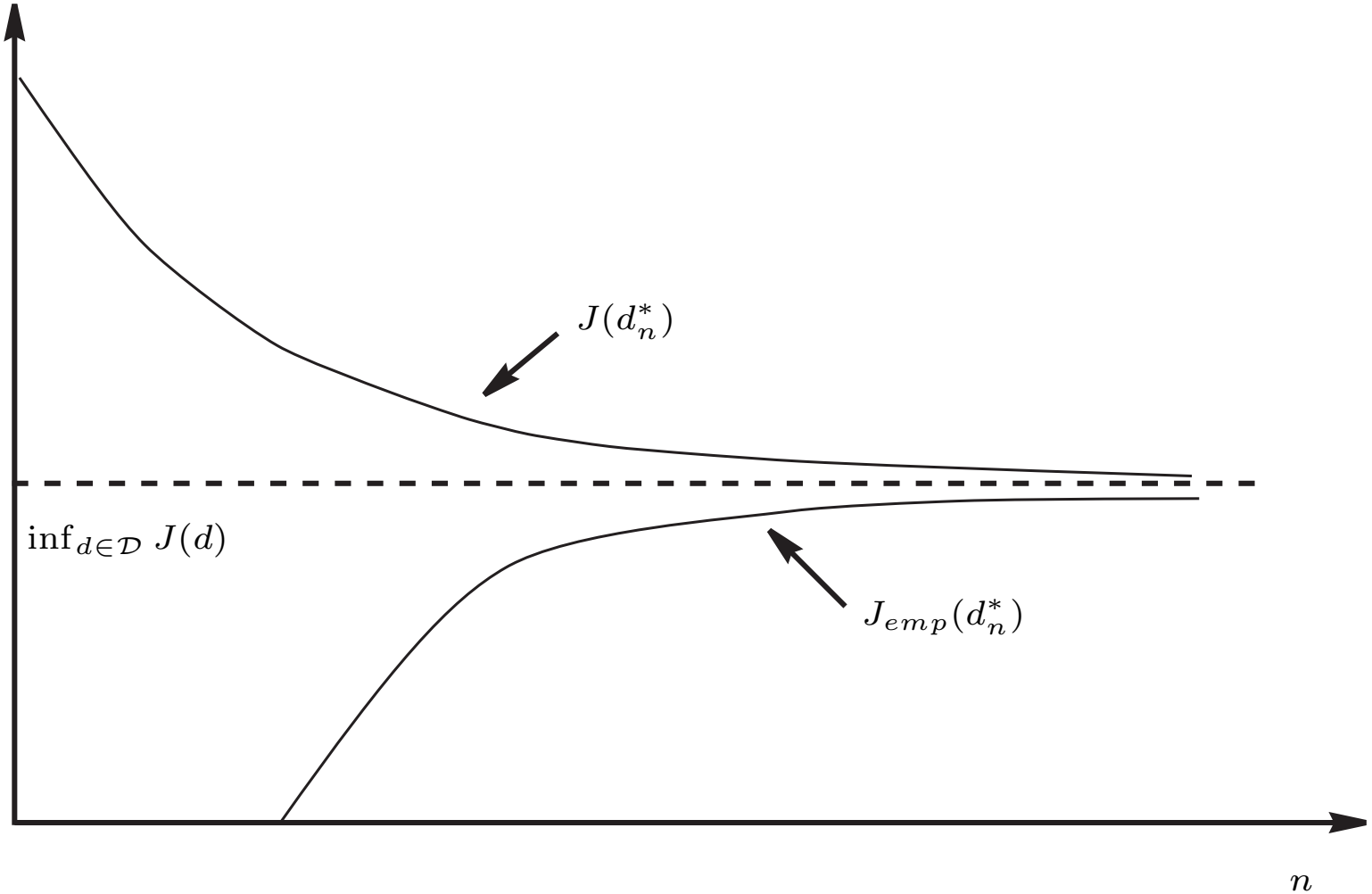
The minimization of empirical risk principle is consistent for the chosen risk and a given problem if the learner does its best when the sample size tends to infinity

Consistency of the minimization of empirical risk principle

The MRE principle is consistent for a cost Q, a class of function $\mathcal{D} = \{d(\boldsymbol{x}; \theta) : \theta \in \Theta\}$ and a probability density function $p(\boldsymbol{x}, y)$ if applied at each sample set $\mathcal{A}_n$, it generate a sequence $\{d_n^*(\boldsymbol{x}; \theta) : \theta \in \Theta\}_{n>0}$ that satisfies :

$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d).$$



$$J(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

$$J_{emp}(d_n^*) \xrightarrow[n\to\infty]{p} \inf_{d\in\mathcal{D}} J(d)$$

For the sake of clarity, up to the end of this section we will consider that the cost function Q is an indicator function. Thus :

$$Q(d(\boldsymbol{x};\theta);y) = \mathbb{1}_{d(\boldsymbol{x};\theta)\neq y} \triangleq \begin{cases} 0 & \text{si} \quad y = d(\boldsymbol{x};\theta) \\ 1 & \text{si} \quad y \neq d(\boldsymbol{x};\theta), \end{cases}$$

The VC dimension (for Vapnik-Chervonenkis dimension) is a measure of the capacity of a statistical classification algorithm. Informally, the capacity of a classification model is related to how complicated it can be

Definition (VC-dimension)

The Vapnik-Chervonenkis dimension of a given class $\mathcal{D}$ of detectors is defined as the largest number of samples $\boldsymbol{x}_k$ from the representation space $\mathcal{X}$ which can be split into any two subset partition using detectors from $\mathcal{D}$.

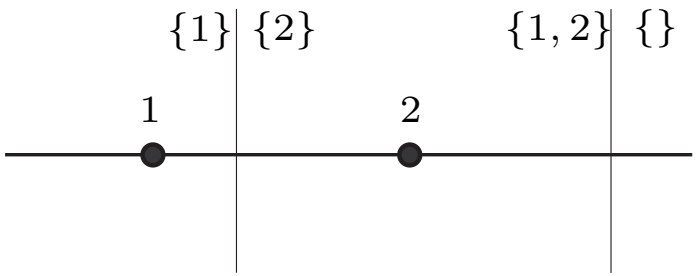
Example 1. Let consider the class $\mathcal{D}$ of linear detectors in $\mathbb{R}^l$ defined by $d(\boldsymbol{x};\theta) = \text{sign}(\sum_{k=1}^{l} \theta_k\, x(k) + \theta_0)$, the parameters θ_k are reals and sign$(\cdot)$ is the "sign" function.
We can show that :

$$h_{\mathcal{D}} = l + 1$$

Example 2. Let consider the class $\mathcal{D}$ of detectors such that $\{d(x;\theta) = \text{sign}(\sin(\theta x)) \ : \ \theta \in \mathbb{R}\}$ defined for $x \in \mathbb{R}$.
It is easy to show that :

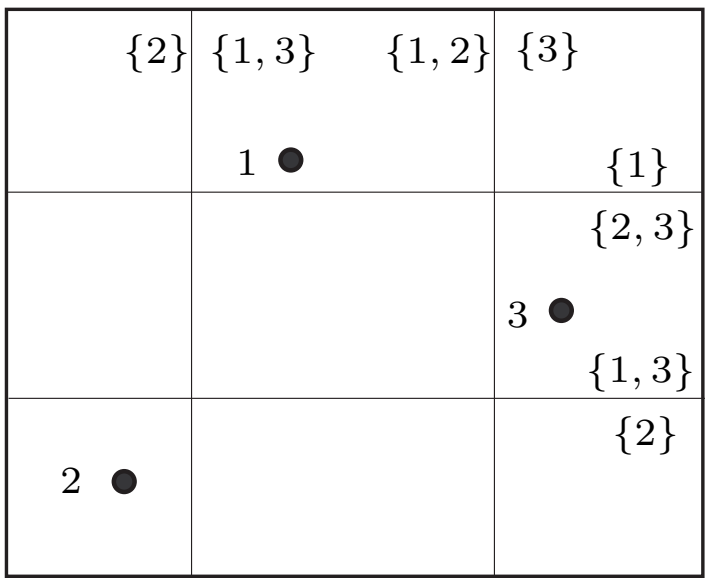
$$h_{\mathcal{D}} = +\infty$$

in $\mathbb{R}$

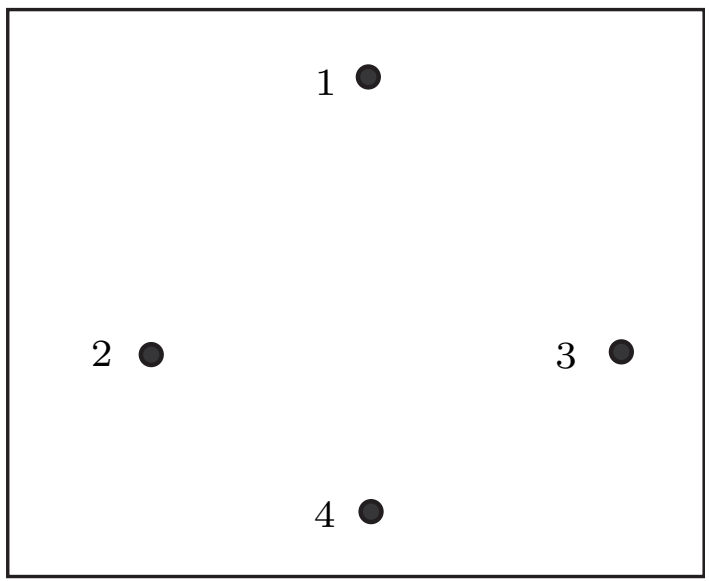


in $\mathbb{R}$

in $\mathbb{R}^2$



in $\mathbb{R}^2$



Theorem

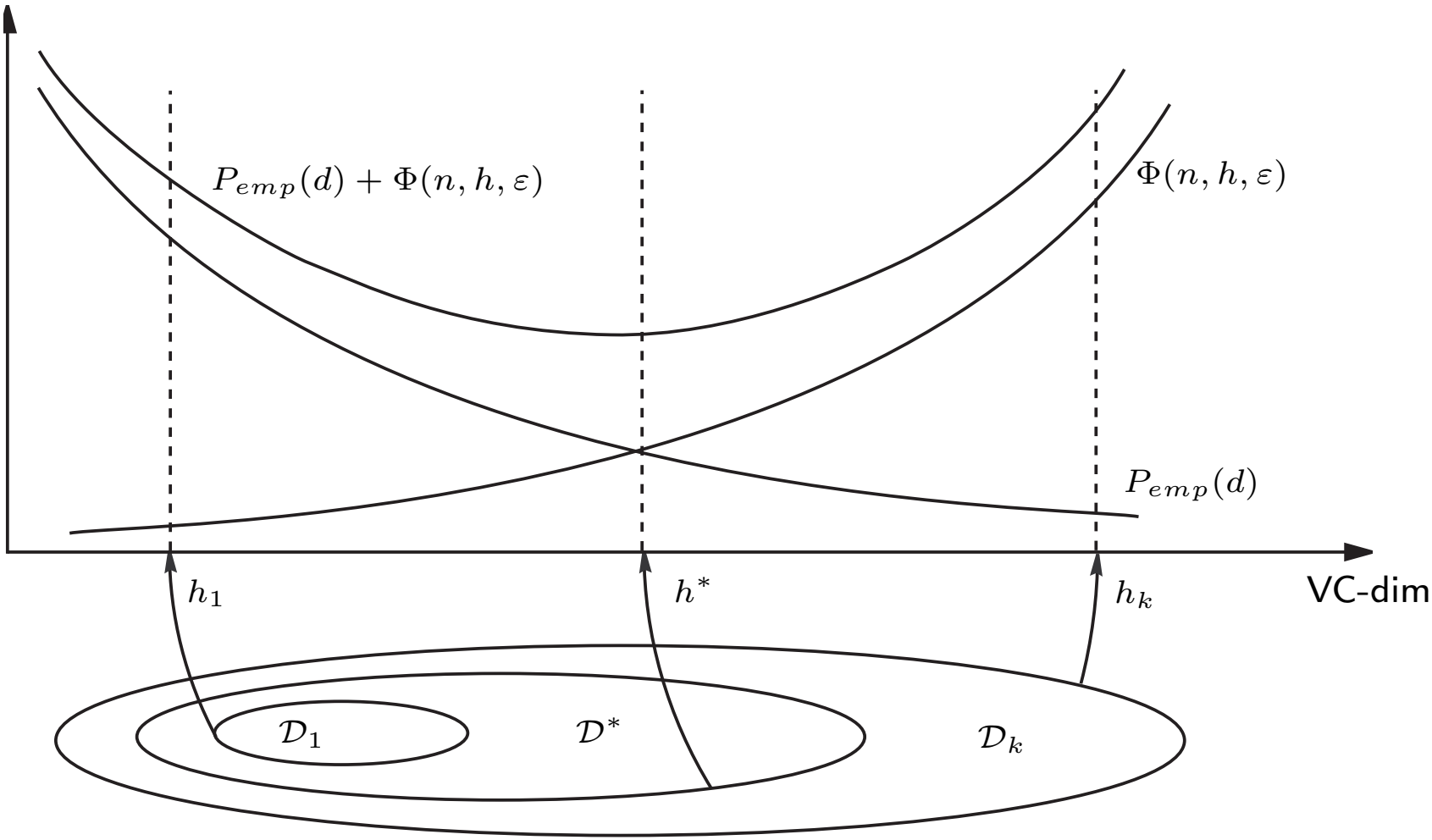
In order, for the minimization of empirical risk principle, to be consistent for any probability distribution it is sufficient for the VC-dimension h of the detector class $\mathcal{D}$ to be finite.

The pioneering work of Vapnik and Chervonenkis (1971) have also made quantitative findings about the convergence rate of P_{emp} to P_e.

Inégalité de Vapnik-Chervonenkis.
With a probability larger or equal to $1-\varepsilon$, we have :

$$P_e(d_n) \leq P_{emp}(d_n) + \sqrt{\frac{h\left(\ln\left(\frac{2n}{h}\right)+1\right) - \ln\frac{\varepsilon}{4}}{n}}.$$

Warning ! Often rough upper bound... but independent of any probability distribution $p(\boldsymbol{x}, y)$.


$P_{emp}(d) + \Phi(n, h, \varepsilon)$
$\Phi(n, h, \varepsilon)$
$P_{emp}(d)$
h_1
h^*
h_k
VC-dim
$\mathcal{D}_1$
$\mathcal{D}^*$
$\mathcal{D}_k$

The minimization of empirical risk principle
Structural risk minimization principle advocated by Vapnik involves the construction, within the class $\mathcal{D}$, of a sequence of nested subsets $\mathcal{D}_k$

$$\mathcal{D}_1 \subset \ldots \subset \mathcal{D}_k \subset \ldots \subset \mathcal{D}.$$

Once this structure established, the learning phase is conducted in two-steps :

1. Research the detector that minimizes the empirical error within each subset $\mathcal{D}_k$:

$$d^*_{n,k} = \arg \min_{d \in \mathcal{D}_k} P_{emp}(d).$$

2. Select the detector with the best guaranteed error $P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)$:

$$d^*_n = \arg \min_{k \geq 1} \{P_{emp}(d^*_{n,k}) + \Phi(n, h_k, \varepsilon)\}.$$