INFO8004-1 – Advanced Machine Learning

Louis Wehenkel

Lecture 2 - Introduction to Statistical Learning Theory

February 13, 2020

Table of contents

Lecture 1 - Objectives

- Understand the general need for machine learning theories
- Get some insight about the nature of the main theorems of statistical learning theory, and a glimpse at the mathematical complexity behind these ideas
- Understand the empirical risk minimization principle and the related (VC-)theory
- Open eyes beyond ERM and VC-theory

Lecture 1 - Format

- **Before Lecture 2:** A syllabus was distributed (Statistical Learning Theory: A primer, 27 pages), and students were asked to read through this document once before the lecture takes place.
- **During Lecture 2:** LW presents the logic behind the theory and discusses the meaning of its results
- **End of Lecture 2:** LW introduces further reading material beyond the Prof's lecture, the goal of the Student's lecture to be done on these topics, the Quiz to prepare the oral exam.
- **After Lecture 2:** Students solve the Quiz about the lecture to prepare the oral exam, based on lectures, reading material and syllabus.

Motivations for Machine Learning Theories

- Reliance on ML requires confidence that it works sufficiently well; gaining such confidence requires 'understanding why it works' beyond pure empirical evidence and gossip.
- Improving ML methods requires understanding the intrinsic weaknesses of currently available methods, beyond trial and error, black-magic, or cooking recipes.
- Theory is a way to embody scientific knowledge in mathematical language, which is the only accurate way to describe conjectures, hypotheses, facts, and their stable logical consequences.
- Machine learning theory has been and will continue to be a main driver of progress in Artificial Intelligence.

Different families of Machine Learning Theories

Many of these ML theories essentially focus on characterizing the ability of Learning algorithms to exploit observational data-sets so as to make accurate "future predictions":

- So-called **"Statistical"** approaches **(this Lecture)**
- So-called "Bayesian" approaches
- So-called "Computational" approaches (for further Reading)
- So-called "Information Theoretic" approaches (other course)

NB: there are also other relevant questions, not covered by these theories, such as

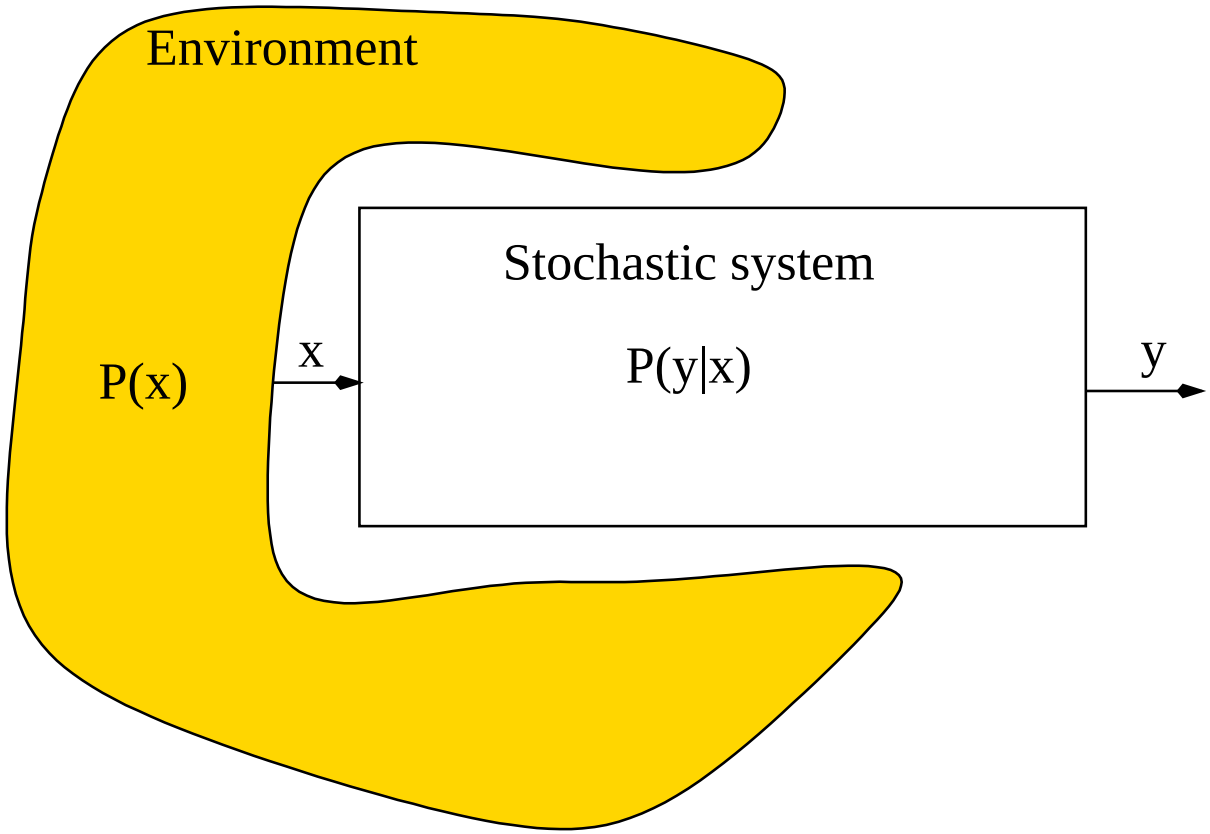
- Inference over cause-effect relationships (reasoning and learning)
- Learning to probe the environment (active learning)
- Learning while controlling a system (reinforcement learning)

Main ingredients of the Statistical Learning model

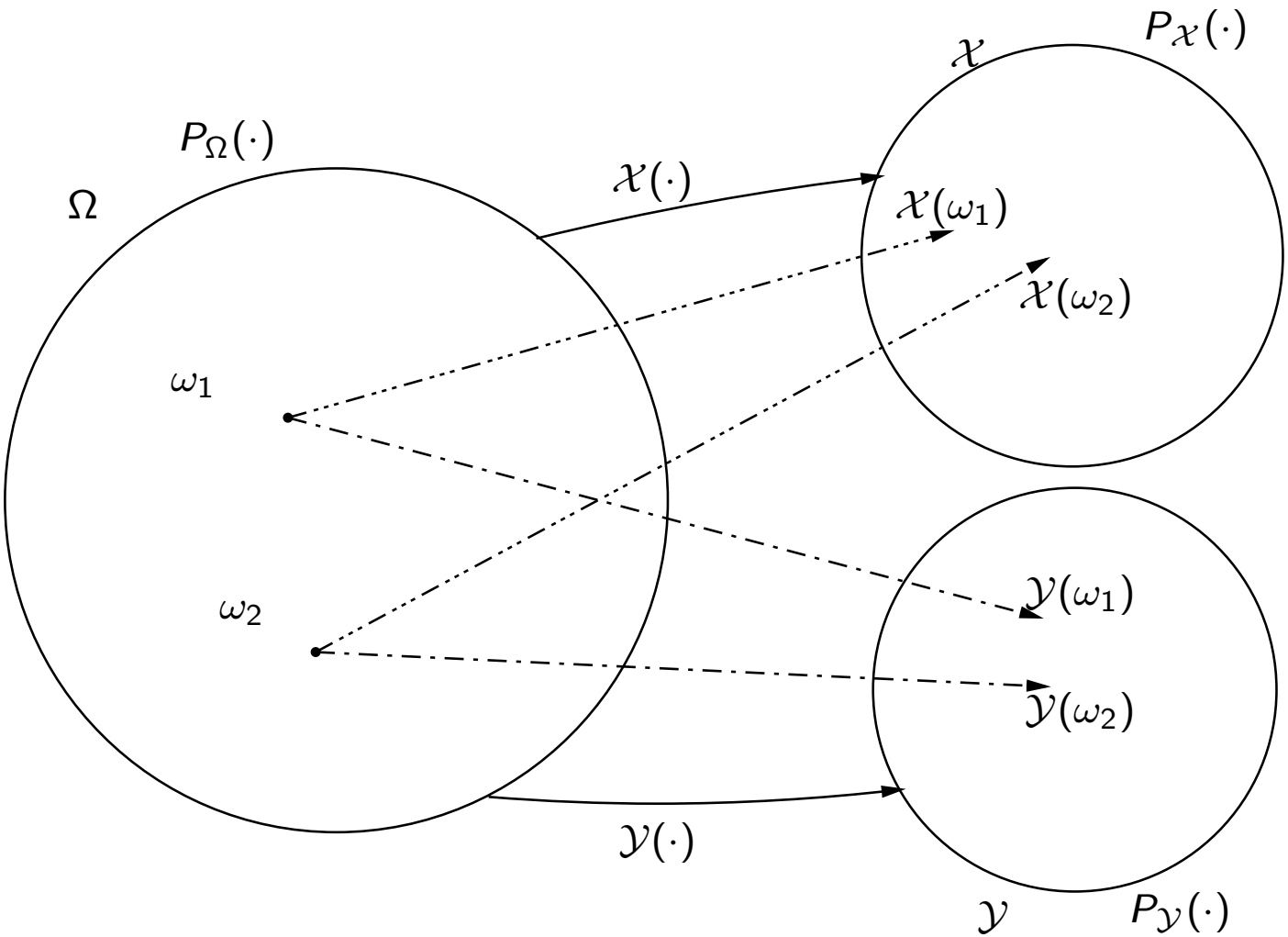
NB: we focus on supervised learning; extensions to non-supervised learning exist as well.

- Stochastic system, probability space and random variables
- Data generating model
- Supervised learning algorithm and hypothesis space
- Loss function and notions of Risk

The viewpoint of a 'Stochastic system'



The viewpoint of 'Probability space and random variables'



The Standard Data Generating Model

- A learning sample, namely an N-tuple of pairs

$$LS = ((x_1, y_1), \ldots, (x_N, y_N)) \in (\mathcal{X} \times \mathcal{Y})^N,$$

  is generated *i.i.d.* according to some $P_{\mathcal{X},\mathcal{Y}}$.
- We however don't have any prior information at all about the sampling distribution $P_{\mathcal{X},\mathcal{Y}}$
- The learning sample can be of any finite size N.
- We aim at studying how the size N of the learning sample influences the performance of the models produced by (supervised) learning algorithms, irrespectively of $P_{\mathcal{X},\mathcal{Y}}$.

Supervised learning Algo and Hypothesis space

- A learning algorithm chooses, according to some procedure, a function f in $\mathcal{Y}^{\mathcal{X}}$, and to do so it uses as input only the learning sample LS.
- We denote by $\mathcal{F}$ the set of all possible functions that can actually be produced by the considered learning algorithm (we call $\mathcal{F}$ the *Hypothesis space* of the learning algorithm).
- A learning algorithm is thus defined by a sequence of rules (i.e. one rule for each sample size)

$$\text{Algo}_N(\cdot) \ : (\mathcal{X} \times \mathcal{Y})^N \rightarrow \mathcal{F} \tag{1}$$

for selecting a function $f \in \mathcal{F}$ given a sample of size N.

Loss function, expected and empirical Risk

- We start by defining a loss-function ℓ in order to measure how close two values in $\mathcal{Y}$ are to each other:

$$\ell(y, y') \geq 0, \text{with } \ell(y, y) = 0.$$

(We say that the loss-function is *bounded*, if $\ell(y, y') \leq \bar{\ell} \in \mathbb{R}$).

- The **expected** risk of f w.r.t. $P_{\mathcal{X},\mathcal{Y}}$ is then defined by

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} \ell(f(x), y) dP_{\mathcal{X},\mathcal{Y}}.$$

- The **empirical** risk of f w.r.t. a learning sample LS is defined by

$$R_e(f, LS) = \frac{1}{N} \sum_{i=1}^{N} \ell(f(x_i), y_i).$$

Learning according to the ERM principle

- Ideally, we would like our learning algorithm to produce the best possible function in $\mathcal{Y}^{\mathcal{X}}$, for any data-generating distribution $P_{\mathcal{X},\mathcal{Y}}$, loss function ℓ, and learning sample LS:

$$f_B = \arg \min_{f \in \mathcal{Y}^{\mathcal{X}}} R(f).$$

  Function f_B is the so-called *Bayes model*; it depends on ℓ and on $P_{\mathcal{X},\mathcal{Y}}$, but it is clearly independent of the LS.
- Therefore, no such 'ideal' learning algorithm exists. Instead, the **Empirical Risk Minimization principle** (ERM) considers learning algorithms that compute a function along

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

  for some chosen hypothesis space $\mathcal{F}$.

The goals of VC-theory

Goals of the VC-Theory:

- Study learning algorithms based on the ERM principle.
- When do such algorithms generalize well ?
- When are they near optimal, i.e. are consistent ?

Remarks about the VC-Theory:

- it provides conditions which are agnostic w.r.t. the data generating distribution (non-parametric, worst-case results)
- it covers many relevant types of loss functions that one wants to use in practice (classification and regression);
- it fully characterizes both the "large sample" and the "small sample" regimes.

Good generalization of learning algos (1)

- Notice that since LS is a random sample, the output f of any reasonable learning algorithm is a random function.
- For a given data-generating distribution, a given sample size N, a given hypothesis space $\mathcal{F}$, we denote by $f^N \in \mathcal{F}$ the random function learned by applying some learning algo, and by $R_e^N(f^N)$ its empirical risk, and by $R(f^N)$ its actual risk.
- When N increases, this generates two sequences of random variables $R(f^N)$ and $R_e^N(f^N)$.
- We say that the learning algo **generalizes well** for some data generating mechanism if (for this mechanism)

$$\lim_{N\to\infty} |R(f^N) - R_e^N(f^N)| = 0 \quad \text{('in Probability').}$$

Discussion of generalization of learning algos (2)

- The notion of "good generalization" can (rather obviously) be applied to ERM based learning algorithms.
- It means that eventually, for sufficiently large samples, the empirical risk converges to the expected risk.
- It doesn't mean that the expected risk is small or near optimal.
- However, later on we will see that when ERM based algos produce near optimal models they must also generalize well.
- In any case, since we don't know the data-generating mechanism, we want such properties to hold whatever the data-generating mechanism.

Universal consistency of learning algos (1)

- For a given data generating distribution, and for a given hypothesis space, let us denote the optimal model by

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

Notice that depending on the data generating distribution and hypothesis space, this model may be near optimal or far away w.r.t. to the Bayes model f_B.

- We say that a learning algorithm is **universally consistent**, if **for every** data generating distribution, we have

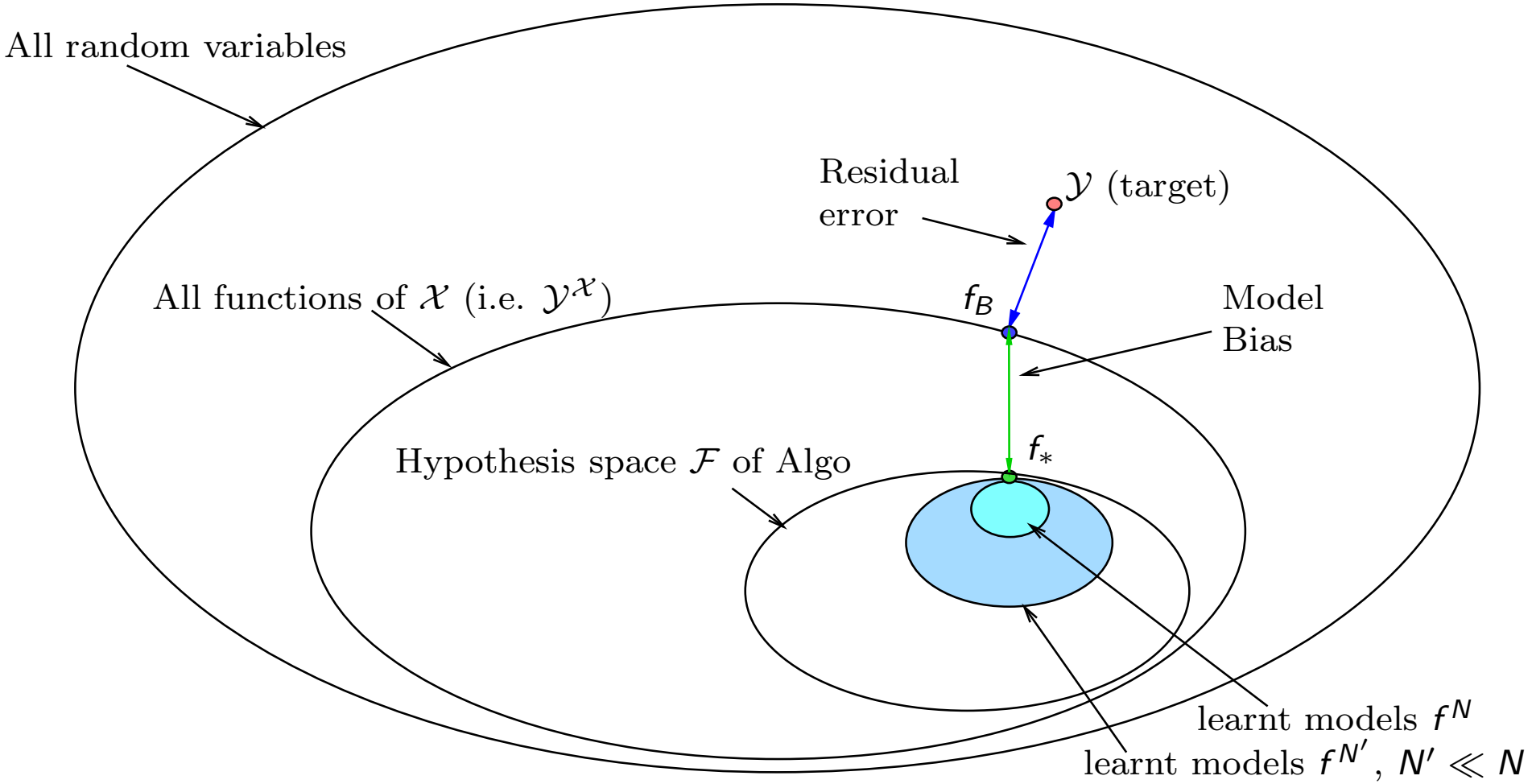
$$\lim_{N \to \infty} |R(f^N) - R(f_*)| = 0 \text{ in Probability.}$$

Discussion of universal consistency of learning algos (1)

- Universal consistency means that the learning algorithm will eventually produce a model close in performance to the best possible model in its hypothesis space $\mathcal{F}$.
- If $\mathcal{F}$ also contains the Bayes model for the data-generating mechanism, this means that the algorithm will eventually produce near-optimal models for large enough N.
- The consistency property may (obviously) be used to study learning algorithms implementing the ERM principle over some hypothesis space $\mathcal{F}$.
- NB: when both consistency and generalization hold true, then we necessarily have also that

$$\lim_{N\to\infty} |R_e^N(f^N) - R(f_*)| = 0 \text{ in Probability.}$$

A picture to summarize what we have discussed



Scope of the VC Statistical Learning Theory

The VC- theory that we will discuss, focuses on learning algorithms that implement the **Empirical Risk Minimization** principle, i.e. learning algorithms that compute the function

$$f_*^{LS} = \arg\min_{f \in \mathcal{F}} R_e(f, LS)$$

for some chosen hypothesis space $\mathcal{F}$.

It aims at understanding, in terms of the properties of the used hypothesis space $\mathcal{F}$, under which conditions such algorithms are (universally) consistent, generalize well, and it also derives large sample and finite sample bounds on convergence and generalization.

The main steps to establish VC-theory

- First, focus on a fixed Data-Generation Mechanism (DGM) $P_{\mathcal{X},\mathcal{Y}}$.
  - Definition of a suitable notion of consistency for ERM-based learning
  - Definition of suitable notions of uniform convergence over $\mathcal{F}$, and show that uniform convergence is equivalent to consistency.
  - Show that uniform convergence is equivalent to slow enough growth of a suitable DGM-dependent measure of complexity of the hypothesis space, called the VC-entropy.
- Second, focus on convergence rate, 0-1 problems, and distribution independent statements.
  - Definition of fast convergence rate
  - Definition of a distribution independent version of VC-entropy (growth function) for 0-1 problems and the resulting notion of VC-dimension.
  - Show that finite VC-dimension is equivalent to distribution independent consistency and also to fast convergence of the ERM principle.
  - Construct bounds for 'finite sample' generalization, based on VC-dimension
- Third, generalize to regression and joint density learning problems

Definition of "Consistency of the ERM principle"

For a given data-generating mechanism $P_{\mathcal{X},\mathcal{Y}}$ and loss function ℓ, we say that the ERM principle is consistent if the following two properties hold :

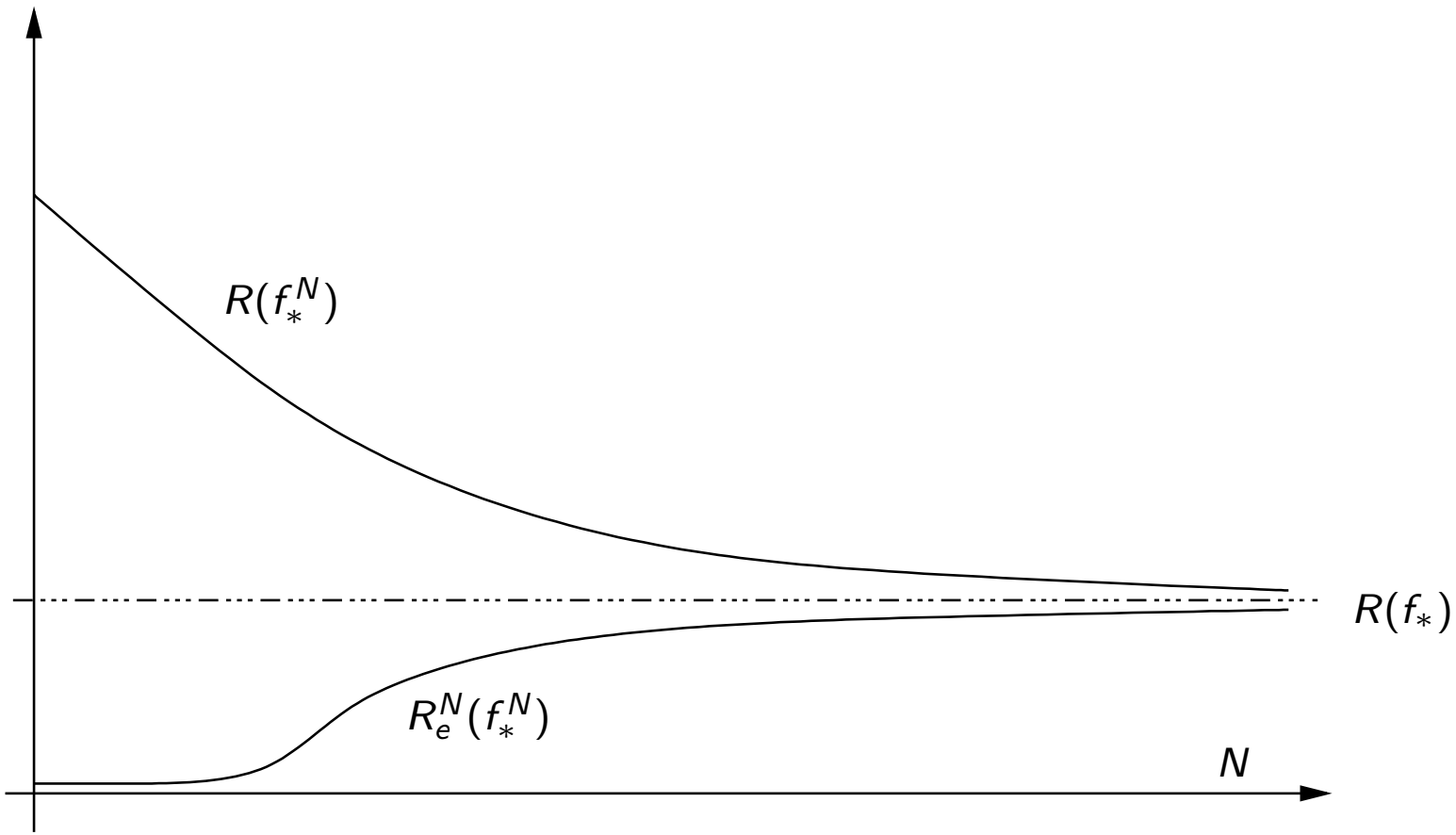
$$\forall \epsilon, \eta > 0, \exists N_1 \ : N \geq N_1 \Rightarrow Pr\left\{|R(f_*^N) - R(f_*)| > \epsilon\right\} < \eta,$$

and

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

Here, $Pr\{\text{Condition}(N)\}$ denotes the probability to observe an i.i.d. sample of size N along $P_{\mathcal{X},\mathcal{Y}}$ such that the condition holds true.

Graphical view of a typical consistent behavior



Implies also $R_e^N(f_*^N) \to R(f_*^N)$, i.e. good generalization.

Uniform *one-sided* convergence and consistency

We define uniform *one-sided* convergence over $\mathcal{F}$ of the empirical risk to the actual risk as follows

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

Theorem (Necessary and sufficient condition of consistency of the ERM principle)

Uniform one-sided convergence over $\mathcal{F}$ is a necessary and sufficient condition of consistency of the ERM principle applied to $\mathcal{F}$.

Uniform *two-sided* convergence and consistency

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

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

Theorem (Sufficient condition of consistency of the ERM principle)

Uniform two-sided convergence over $\mathcal{F}$ implies uniform one-sided convergence over $\mathcal{F}$ and is only a sufficient condition of consistency of the ERM principle applied to $\mathcal{F}$.

Example and counter-example of consistent behavior

Example: For finite hypothesis spaces, it is 'trivial' to show that uniform *two-sided* convergence always holds true (law of large numbers), hence that the ERM principle always yields consistent learning algorithms in this case. Notice that when both $\mathcal{X}$ and $\mathcal{Y}$ are finite, then also $\mathcal{Y}^{\mathcal{X}}$ (and any meaningful $\mathcal{F}$) is finite.

Counter-example: $\mathcal{Y} = \{0, 1\}$; continuous $\mathcal{X}$ uniform over $[0, 1]$, $P(y = 0 \mid x) = 0.75$, independent of x. Consider the 0-1 loss function and the space $\mathcal{F}$ of all functions from $\mathcal{X}$ to $\mathcal{Y}$. We have $f_*(x) = 0, \forall x \in [0, 1]$, hence $R(f_*) = 0.25$, thus $R(f) \geq 0.25$, $\forall f \in \mathcal{F}$. On the other hand, we have that $R_e^N(f_*^N) = 0$ for almost every sample of size N. Hence ERM does not generalize well, hence it can also not be consistent.

VC-Entropy of $\mathcal{F}$ w.r.t. $P_{\mathcal{X},\mathcal{Y}}$ (and ℓ)

For a given LS of size N, function $f \in \mathcal{F}$, and loss ℓ we define

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

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

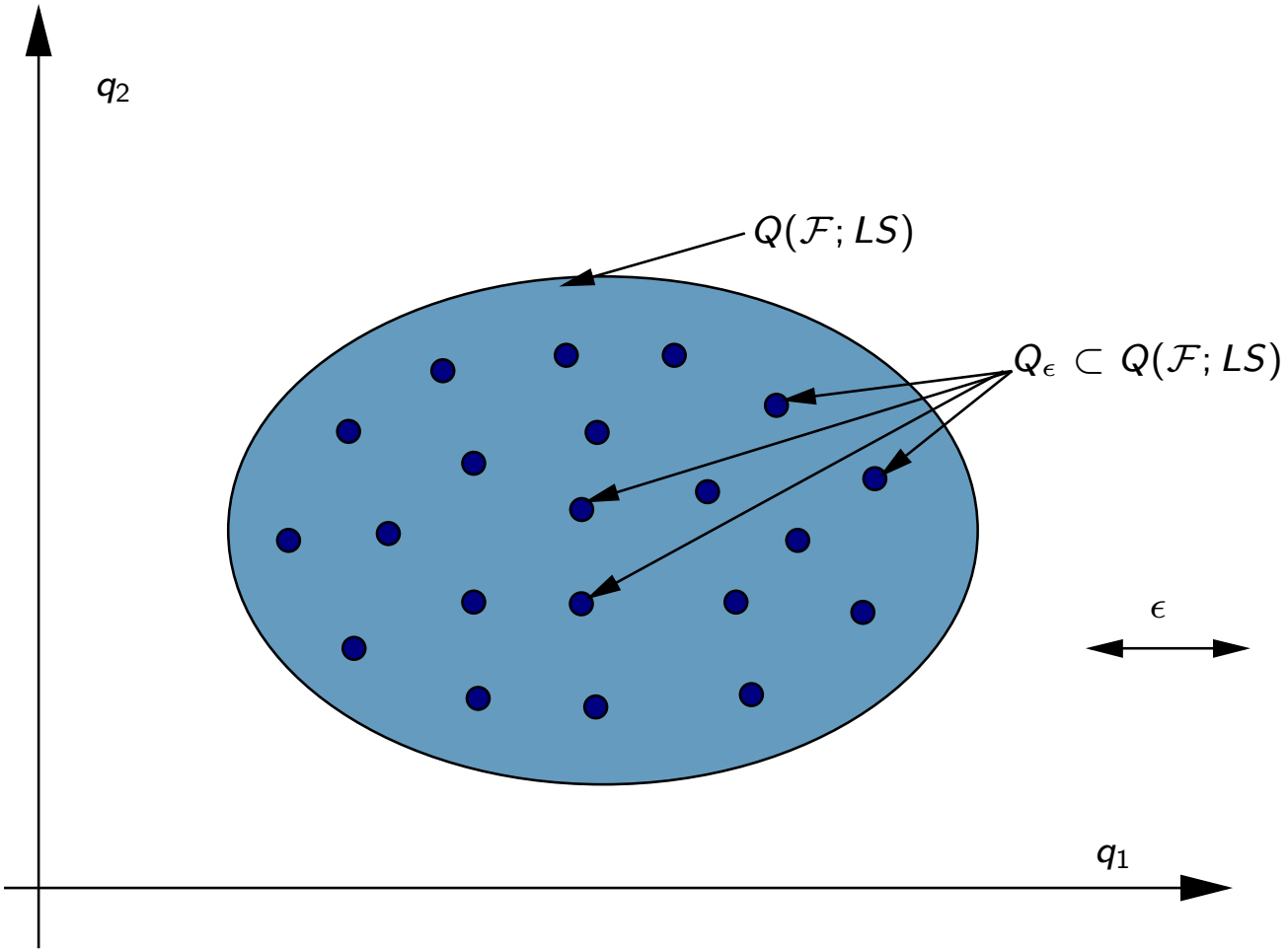
$$Q(\mathcal{F}; LS) = \{q(f; LS) \ : f \in \mathcal{F}\}.$$

The set $Q(\mathcal{F}; LS)$ is potentially of infinite cardinality (if $\mathcal{F}$ is itself of infinite cardinality).

Thus we define (see figure on next slide), $\forall \epsilon > 0$, 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, such that

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

Graphical view of ϵ-net (for $N = 2$)



Theorem 1: distribution dependent uniform convergence

If $Q(\mathcal{F}; LS)$ is bounded then, $\forall \epsilon > 0$, there exists always a finite ϵ-net. Hence denote the number of vectors of the smallest ϵ-net of $Q(\mathcal{F}; LS)$ by $N^{\mathcal{F}}(\epsilon; LS)$.

We then define the VC-entropy of $\mathcal{F}$ by

$$H^{\mathcal{F}}(\epsilon; N) = E_{LS}\{\ln N^{\mathcal{F}}(\epsilon; LS)\}, \tag{2}$$

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 $P_{\mathcal{X},\mathcal{Y}}$.

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

$$\lim_{N\to\infty} \frac{H^{\mathcal{F}}(\epsilon; N)}{N} = 0, \forall \epsilon > 0. \tag{3}$$

Application to binary classification

NB: Here $y, f(x) \in \{0,1\}$, $\ell(y,y') = \delta_{y,y'} \in \{0,1\}$ (R is the error rate). Hence

$$q(f; LS) = (\ell(f(x_1), y_1), \ldots, \ell(f(x_N), y_N)) \in \{0,1\}^N.$$

In this case $\#Q(\mathcal{F}; LS)$ is actually equal to the number of ways we can classify the LS by choosing functions from $\mathcal{F}$. Further, for $\epsilon < 1$, we have $N^{\mathcal{F}}(\epsilon; LS) = \#Q(\mathcal{F}; LS)$.

Thus, if we suppose that $\mathcal{F}$ is sufficiently rich to classify LS in all possible ways, we have $N^{\mathcal{F}}(\epsilon; LS) = 2^N$.

If $\forall N$ this happens with probability 1, then $H^{\mathcal{F}}(\epsilon; N) = N \ln 2$ for $\epsilon < 0$ and hence ERM can not be consistent.

The notion of sample "shattering" by $\mathcal{F}$

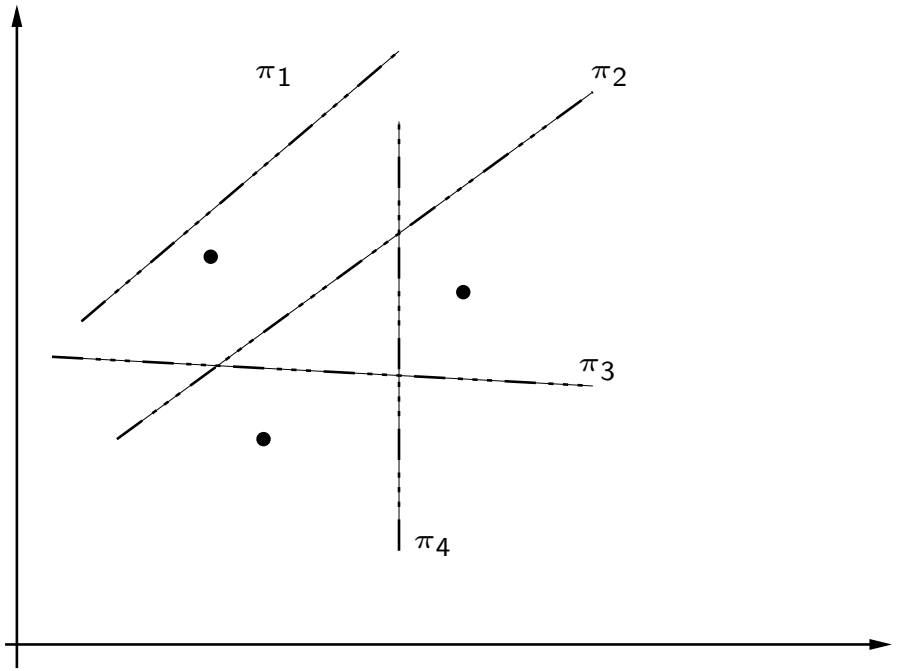
For a given sample LS, let us denote by $N^{\mathcal{F}}(LS)$ the number of different ways we can dichomotize LS by choosing functions from $\mathcal{F}$. Notice that this number only depends on the values of the inputs $x_1, \ldots, x_N$ and on $\mathcal{F}$, but not on $y_1, \ldots, y_N$.

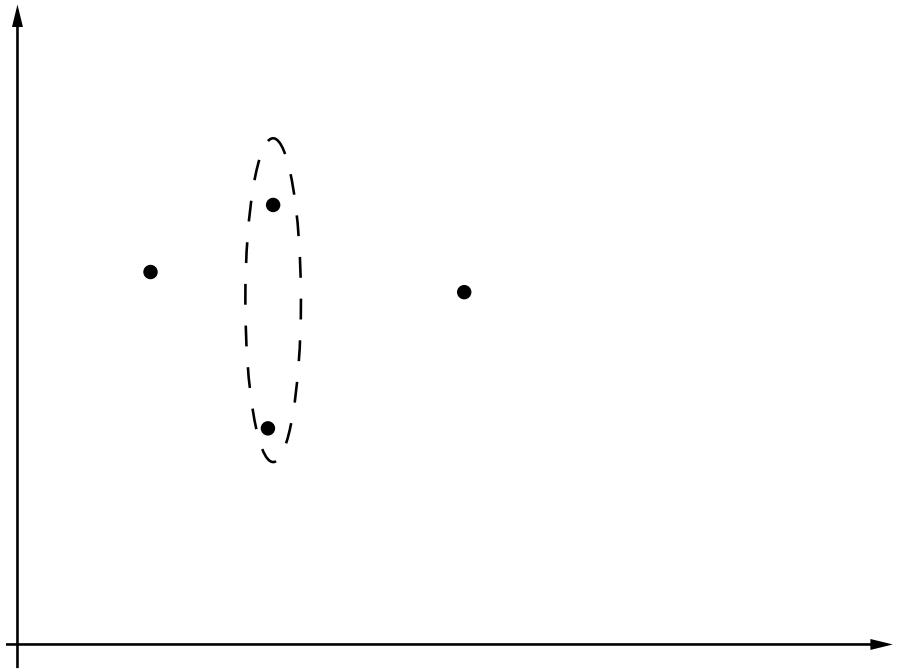
Clearly, if $N^{\mathcal{F}}(LS) = 2^N$ (for some LS of size N) then there exists a function in $\mathcal{F}$ with $R_e^N = 0$, whatever the values of $y_1, \ldots, y_N$.

In such a situation, we say that the LS is *shattered by* $\mathcal{F}$.

Obviously, if this is the case then ERM will compute a function that perfectly classifies LS, whatever the values of $y_1, \ldots, y_N$.

Shattering of points in $\mathbb{R}^{d=2}$ by linear classifiers





Left Figure: the set of all linear classifiers in two dimensions can shatter most samples of size $N = 3$. For a sample in general position, such as the one shown, we indeed have $N^{\mathcal{F}}(LS) = 2^N$.

Right Figure: the set of all linear classifiers in two dimensions can't shatter any sample of size $N = 4$. $N^{\mathcal{F}}(LS) < 2^N$ forall LS of size 4. This remains the case $\forall N > 4$.

Asymptotic rate of convergence for binary classification

We say that the asymptotic rate of convergence of the sequence of classifiers selected by the ERM principle is fast if

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

We further define the *Growth-function* by

$$G^{\mathcal{F}}(N) = \ln \left\{ \sup_{LS \text{ of size } N} N^{\mathcal{F}}(LS) \right\}, \tag{4}$$

which is a distribution-independent measure of the capacity of a set of indicator functions. Clearly, $H^{\mathcal{F}}(\epsilon, N) \leq G^{\mathcal{F}}(N)$. Therefore the condition

$$\lim_{N \to \infty} \frac{G^{\mathcal{F}}(N)}{N} = 0, \tag{5}$$

certainly provides a sufficient condition for consistency of the ERM principle on $\mathcal{F}$ with respect to any possible probability measure.

Theorem 2: distribution-independent fast convergence rate

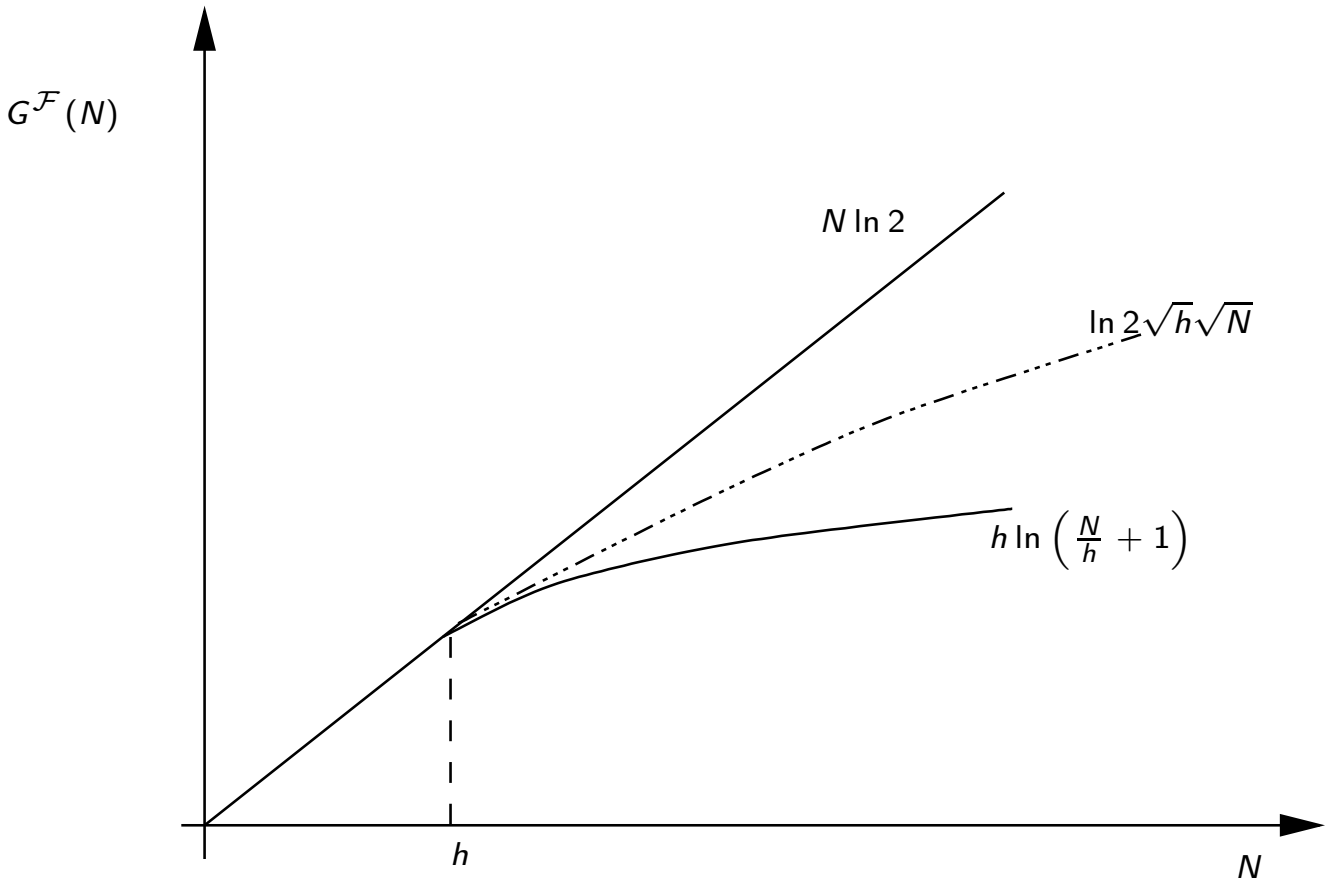
Actually, it turns out that this condition is much stronger.

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

$$\lim_{N\to\infty} \frac{G^{\mathcal{F}}(N)}{N} = 0, \tag{6}$$

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.

Possible shapes of the growth-function $G^{\mathcal{F}}(N)$



For $N \leq h$, $\mathcal{F}$ shatters at least one LS of size N; beyond h it is not the case anymore.

VC-dimension of a set $\mathcal{F}$ of binary classifiers

When $G^{\mathcal{F}}(N) = N \ln 2$, $\mathcal{F}$ shatters some (most) samples of size N (those in general position).

If $G^{\mathcal{F}}(N) = N \ln 2$ remains true for $N \to \infty$, we say that the VC-dimension of $\mathcal{F}$ is *infinite*.

Otherwise, the VC-dimension is equal to the largest sample size h suchthat $G^{\mathcal{F}}(N) = N \ln 2, \forall N \leq h$.

In this latter case, $G^{\mathcal{F}}(N) = h \ln \left(\frac{N}{h} + 1\right), \forall N \geq h$.

Finite VC-dimension of a set $\mathcal{F}$ of binary classifiers

Theorem (Finite VC-dimension)

Finite VC-dimension of $\mathcal{F}$ is a necessary and sufficient condition both for distribution independent consistency and distribution independent fast asymptotic convergence of ERM based learning applied to binary classification problems.

It is therefore also a sufficient condition for distribution independant good generalization of ERM based learning applied to binary classification problems.

The main interest of this result 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).

See syllabus for examples, and extensions to regression problems.

Finite sample bounds of ERM based learning

Questions:

- Bounds on the actual risk of $R(f_*^N)$?
- Bounds on the suboptimality w.r.t. $R(f_*)$?

Answers in terms of finite VC-dimension h of $\mathcal{F}$

- For any function $f \in \mathcal{F}$ (hence for f_*^N), we have with probability $1-\alpha$

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

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

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

Synthesis of scope of VC-theory

- The theory characterizes in many insightful respects ERM algos, and thus helps understanding learning machines.
- Its scope covers multi-class classification, mutli-dimensional regression, and density learning
- It provides necessary and sufficient conditions, for generalization, consistency and fast convergence rates, as well as finite sample bounds.
- Most of its useful results are distribution independent, therefore widely usable while at the same time often very conservative.
- The way of reasoning has paved the way for other learning theories, and for the design of algos (e.g. SVM, Boosting).

Limitations of VC-theory

- Conservative nature of the distribution independent finite sample bounds makes them of little practical use.
- Distribution dependent part of theory is of limited use for practical problems.
- The theory leaves several questions open:
  - How to take into account prior knowledge about the likely data-generation mechanisms ?
  - What about finite (but very large) hypothesis spaces, and hence for finite input-output spaces ?
  - What about computational complexity ?
  - What about algos that do not follow the ERM principle ?

Quiz to check understanding of Lecture 2 (part 1)

NB: All questions refer to binary classification problems with a 0-1 loss function ℓ, and with $x \in \mathbb{R}^d$ and continuous.

- Basics (finite $\mathcal{F}$)
  - Given an oracle to sample from $P(x)$ and another oracle to sample from $P(y \mid x)$, explain how you would generate T different LSs of size N.
  - Given a finite hypothesis space $\mathcal{F}$, explain how you would then proceed to compute $f_*^N(LS)$ and $R_e^N(f_*^N(LS))$, and then $R(f_*^N(LS))$.
  - How would you proceed, using the computer, in order to estimate $R(f_*)$ and then $Prob\{|R_e^N(f_*^N(LS)) - R(f_*)| > \epsilon\}$.
- Consistency vs generalization (finite vs infinite $\mathcal{F}$)
  - Do we have consistency of the ERM principle in the above case ?
  - Do we also have good generalization of ERM based learning ?
  - In the above case, can you imagine a learning algorithm using a finite hypothesis space that would not be consistent and/or not generalize well ?
  - In the above case, can you imagine a learning algorithm using an infinite hypothesis space that would not generalize well and/or not be consistent ?

Quiz to check understanding of Lecture 2 (part 2)

- Consistency versus generalization for general $\mathcal{F}$.
  - Explain why consistency implies good generalization for ERM based algorithms, and why good generalization of an arbitrary learning algorithm does in general not imply its consistency.
  - How would you use a sample to estimate $R(f_*)$, by exploiting some consistent ERM based algorithm and an oracle to generate i.i.d. samples from some $P_{\mathcal{X},\mathcal{Y}}$
- VC-dimension vs VC-entropy
  - Explain why finite VC-dimension of $\mathcal{F}$ implies slow growth of VC-entropy for any data generating mechanism $P_{\mathcal{X},\mathcal{Y}}$.
  - Explain why the VC-dimension of a finite $\mathcal{F}$ must be finite.
  - Provide an example of a hypothesis space of infinite VC-dimension, and for this space provide a data-generating mechanism for which the ERM based learning would not be consistent.
- Beyond VC-theory
  - Can you think of a scenario where non-ERM based algorithms would generalize well ?
  - Can you think of a scenario where non-ERM based algorithms would outperform asymptotically any consistent ERM-based algorithm ?