No Free Lunch Theorem

Recap from last time:

- $\mathcal{X}$ = domain set of "feature vectors" ; $\mathcal{Y}$ = "label" set
- $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, $z = (x, y) \leftarrow$ sample.
- $S = \{ (x_1, y_1), \dots, (x_m, y_m) \}$ "Training set"
- $\mathcal{D}(x,y) = \mathcal{D}(y \mid x)\, \mathcal{D}(x)$ unknown distr generating iid samples.
- $h : \mathcal{X} \to \mathcal{Y}$ a rule or an hypothesis.
- $\mathcal{H} \ni h$ a class of rules or "hypothesis class"
- $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$ loss fct
  $(h, z) \mapsto \ell(h, z)$
  - 0-1 loss is $\ell(h, z) = \mathbb{1}(h(x) \neq y)$
- True loss, true error, gen error
  $$L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}} [\ell(h, z)]$$
  - square loss is $\ell(h, z) = (y - h(x))^2$
- Empirical loss
  $$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

- ERM rule : $h_S = ERM(S) = \operatorname{argmin} L_S(h)$

Definition : **Agnostic PAC learning**

$\mathcal{H}$ is agnostic PAC learnable w.r.t Z, ℓ if

$\exists\ m_{\mathcal{H}} : [0,1]^2 \to \mathbb{N}\ ;\ (\epsilon,\delta) \mapsto m_{\mathcal{H}}(\epsilon,\delta)$ and a

learning rule A such that :

$\forall (\epsilon,\delta)$ and $\forall \mathcal{D}$ (on Z), when running

A on $m > m_{\mathcal{H}}(\epsilon,\delta)$ iid samples $S \sim \mathcal{D}^m$

$$\mathbb{P}rob_S \left\{ L_{\mathcal{D}}(A(S)) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon \right\} \geq 1-\delta$$

Terminology :

- PAC = Probably Approximately Correct
- Realizable case is when $\exists h \in \mathcal{H}$ s.t $L_{\mathcal{D}}(h) = 0$ so $\min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) = 0$. We say that $\mathcal{H}$ is PAC learnable if $\exists\ m_{\mathcal{H}}$ and A s.t $\forall(\epsilon,\delta)$, $\forall\mathcal{D}$ realizable (i.e $\exists h\in\mathcal{H}$ with $L_{\mathcal{D}}(h)=0$) for $m > m_{\mathcal{H}}(\epsilon,\delta)$ we have $\mathbb{P}rob_S \{ L_{\mathcal{D}}(A(S)) \leq \epsilon \} \geq 1-\delta$.
- Agnostic PAC learnable is more general and considers any $\mathcal{D}$.

Last time we saw that all finite hypothesis classes $|\mathcal{H}| < +\infty$ are <u>agnostic PAC learnable</u> with the ERM rule for $m > m_{\mathcal{H}}^{UC}(\epsilon,\delta) = \left\lceil \frac{1}{2\epsilon^2} \log \frac{2|\mathcal{H}|}{\delta} \right\rceil$

Of course this implies that finite hypothesis classes are also <u>PAC learnable</u>.

In exercises on the other hand you saw that infinite classes $|\mathcal{H}| = +\infty$ may be PAC learnable. For example $\mathcal{H}$ = Threshold functions (and the loss fct is the 0-1 classifier).

This brings the questions :

* Are all infinite hyp classes PAC learnable ?
* We may also ask a more naive and basis question: Do we need to restrict ourselves to rules belonging to some hypothesis class ? In other words could we find a universal Algo that learns for any $\mathcal{D}$ irrespective of have the restriction $h \in \mathcal{H}$?

The answer to first question will be given in next chapters and turns out to be subtle. Some infinite classes are PAC learnable when they are in some sense of effective finite size as captured by the notion of VC (Vapnik-Chervonenkis) dimension.

In this chapter we answer the second question:

From now on we assume that we are in the realizable case. So basically we consider $\mathcal{H}$ and $\mathcal{D}$ such that $\exists h \in \mathcal{H} : L_{\mathcal{D}}(h) = 0$.

Moreover we address the question only for $\mathcal{Y} = \{0,1\}$ classification problems and the loss function $\ell(h; x, y) = \mathbb{1}(h(x) \neq y)$. In particular

$$L_{\mathcal{D}}(h) = \mathbb{E}_{\mathcal{D}}(\mathbb{1}(h(x) \neq y)) = \mathbb{P}(h(x) \neq y).$$

ε $\quad 1/8$

Proposition. (e.g. $\mathcal{X} = \mathbb{R}^N$)

Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ the set of all functions $h : \mathcal{X} \to \{0,1\}$. Then $\mathcal{H}$ is not PAC learnable.

Discussion: This proposition basically says that some form of prior knowledge is necessary for PAC learnability. The hypothesis class should be suitably restricted (in particular the set of all functions $h : \mathcal{X} \to \{0,1\}$ is much too big when $\mathcal{X}$ is infinite).

Proof ← Read after NFL theorem

Since $\mathcal{X}$ is infinite we can take $C \subset \mathcal{X}$ with $|C| = 2m$ whatever integer m size of S is. By the construction of No Free Lunch Theorem (see later) $\exists\, \mathcal{D}$ which is realizable (for some f) and at the same time $\mathbb{P}(L_{\mathcal{D}}(f) \geq \frac{1}{8}) \geq \frac{1}{7}$.

Take any $\varepsilon < \frac{1}{8}$, $\delta < \frac{1}{7}$. We have:

$$\delta < \frac{1}{7} \leq \mathbb{P}\left(L_{\mathcal{D}}(f) \geq \frac{1}{8}\right) \leq \mathbb{P}\left(L_{\mathcal{D}}(f) \geq \varepsilon\right) \Rightarrow \mathbb{P}\left(L_{\mathcal{D}}(f) \leq \varepsilon\right) \leq 1-\delta$$

which contradicts PAC learnability.

Theorem: No Free Lunch.

Let A be any learning rule for the task of binary classification over a finite domain $\mathcal{X}$. (0-1 loss assumed here). Let $m < \frac{|\mathcal{X}|}{2}$ be the size of the training set $S = (x_1 \dots x_m)$.

Then there exist a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\} = \mathcal{Z}$ such that:

a) There exist $f: \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$

b) $\mathbb{P}\left(L_{\mathcal{D}}(A(S)) \geq \frac{1}{8} \right) \geq \frac{1}{7}$.

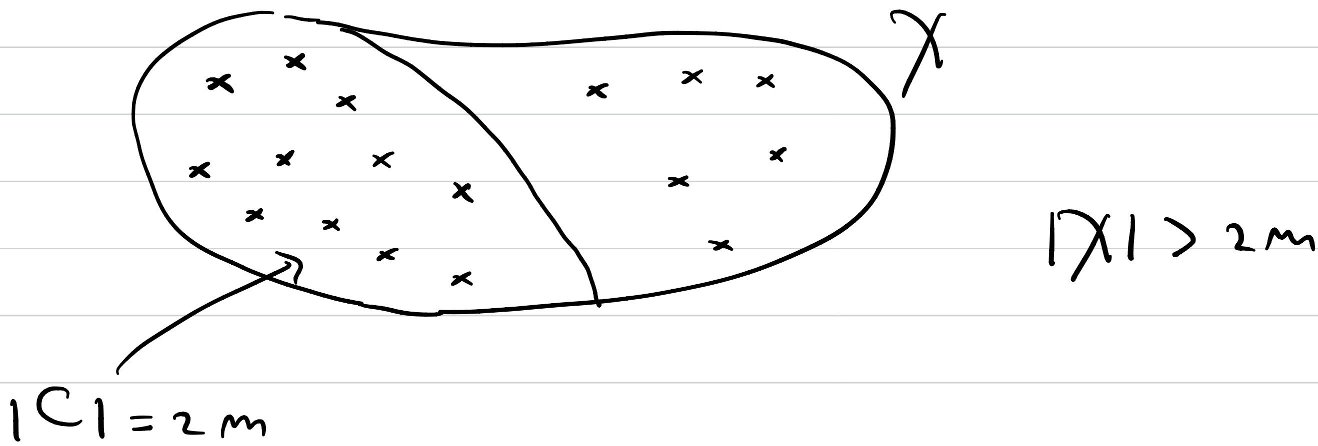
\# for same $\mathcal{D}$.

The theorem says that a) there exist a learner which succeeds since we are in realizable case it suffices to take ERM and "minimize" it over the set $\{f\}$ trivialy.
b) at the same time the learner A will fail on $\mathcal{D}$.
So there is No universaly succesful A and "No free lunch".

Proof.

a) We first show $\exists\, \mathcal{D}$ s.t. $L_{\mathcal{D}}(f) = 0$.

Let $C \subset \mathcal{X}$ with $|C| = 2m$



→ Set of all functions $f_i : C \to \{0, 1\}$

has cardinality $T = 2^{2m}$

(for each input $x_1 \dots x_{2m}$ we have $2 \cdot 2 \cdot \dots 2 = 2^{2m}$ possible values that define the fct).

→ Define $\mathcal{D}_i(x, y) = \begin{cases} \frac{1}{|C|} & \text{if } y = f_i(x) \\ 0 & \text{if } y \neq f_i(x). \end{cases}$

prob to choose (x, y) is $\frac{1}{|C|}$ if label is correct according to f_i.

$$\mathbb{P}_{\mathcal{D}_i}\left(y \neq f_i(x) \right) = 0$$

so equivalently $\mathbb{E}_{\mathcal{D}_i}\left(\mathbb{1}\left(y \neq f_i(x)\right)\right)$

$$= L_{\mathcal{D}_i}(f_i) = 0.$$

We see that all $\mathcal{D}_i$'s $i = 1 \dots 2^{2m}$

are realizable.

This achieves to prove part a).

<u>Now we turn to b).</u>

We will show that

$$* \quad \max_{i=1 \dots 2^{2m}} \mathbb{E}_{S \sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S_i)) \right] \geq \frac{1}{4}$$

So there is some i_0 for which

$$\mathbb{E}_{S \sim \mathcal{D}_i^m}\left(L_{\mathcal{D}_i}(A(S_i)) \right) \geq \frac{1}{4}$$

We will prove later (<u>see Lemma</u>) that this implies

$$\mathbb{P}_{S \sim \mathcal{D}_i^m}\left(L_{\mathcal{D}_i}(A(S)) \geq \frac{1}{8} \right) \geq \frac{1}{7}$$

+

To show * we proceed as follows:

Consider all possible sequences $\underbrace{\text{with samples}}_{}$ $S=(x_1 \cdots x_m)$ in C. There are $K=|C|^m$ of them and recall $|C|=2m$ so $K=(2m)^m$.

Denote them $S_1, S_2, \ldots, S_K$.

For $S_j=(x_1 \ldots x_m)$ we denote by $S_j^i=((x_1\ f_i(x_1)) \ldots (x_m\ f_i(x_m)))$ the training sets with the labels corresponding to f_i.

If samples are drawn $\sim \mathcal{D}_i$ then the algorithm A receives the possible training sets $S_1^i\ S_2^i \ldots S_K^i$ with uniform prob (since the samples are drawn in C with unif prob)

$$\Rightarrow \mathbb{E}_{S\sim\mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right] = \frac{1}{K}\sum_{j=1}^{K} L_{\mathcal{D}_i}(A(S_j^i))$$

$$\Rightarrow \max_{i=1\dots 2^{2m}} \mathbb{E}_{S\sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right]$$

$$\geq \frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} \mathbb{E}_{S\sim \mathcal{D}_i^m}\left[L_{\mathcal{D}_i}(A(S))\right]$$

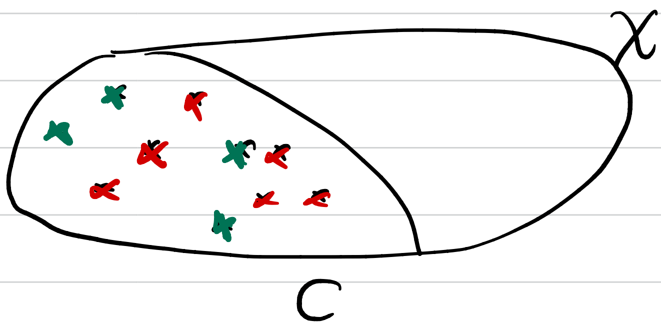
$$= \frac{1}{K}\sum_{j=1}^{K} \frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} L_{\mathcal{D}_i}(A(S_j^i))$$

$$\geq \min_{j=1\dots K} \frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} L_{\mathcal{D}_i}(A(S_j^i)).$$

Now fix some sequence $j \in \{1 \dots K\}$ (recall $K = |C|^m$)

$S_j = (x_1 \dots x_m) \subset C$. Let $v_1 \dots v_p$ the $x \in C$

that are not in S_j.



$\begin{cases} p \geq m \text{ since } |C| = 2m \\ \text{and there can be repetitions in } S_j. \end{cases}$

for any rule h by definition:

$$L_{\mathcal{D}_i}(h) = \frac{1}{|C|} \sum_{x \in C} \mathbb{1}\left(f_i(x) \neq h(x) \right)$$

since $\mathcal{D}_i$ has weight $\frac{1}{|C|}$ on $(x, y = f_i(x))$.

$$\Rightarrow L_{\mathcal{D}_i}(h) \geq \frac{1}{|C|} \sum_{r=1}^{p} \mathbb{1}\left(f_i(v_r) \neq h(v_r) \right)$$

↑ take only v_r that do not appear in S_j^i

$$\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\left(f_i(v_r) \neq h(v_r) \right)$$

since $|C| = 2m \leq 2p$.

$$\Rightarrow L_{\mathcal{D}_i}\left(A(S_j^i) \right) \geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}\left(f_i(v_r) \neq A(S_j^i)(v_r) \right)$$

$$\Rightarrow \frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} L_{\mathcal{D}_i}\left(A(S_j^i) \right) \geq \frac{1}{2p} \sum_{r=1}^{p} \underbrace{\frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} \mathbb{1}\left(f_i(v_r) \neq A(S_j^i)(v_r) \right)}_{\geq \frac{1}{2}}$$

we show below that $\geq \frac{1}{2}$

Thus $\frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{4} \quad \forall j$.

and combining with inequ above on page 9 ;

$$\max_{i=1\dots 2^{2m}} \mathbb{E}_{S\sim\mathcal{D}_i^m} \left[L_{\mathcal{D}_i}(A(S)) \right]$$

$$\geq \min_{j=1\dots K} \frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{4}.$$

$*$

Now it remains to show :

$$\frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} \mathbb{1}\left(f_i(v_r) \neq A(S_j^i)(v_r) \right) \geq \frac{1}{2}.$$

This sum is over set of functions $f_1\, f_2 \cdots f_{2^{2m}}$

There are $\frac{2^{2m}}{2}$ disjoint pairs $(f_i, f_{i'})$

such that ;

$\forall x \in C \quad f_i(x) \neq f_{i'}(x)$ iff $x = v_r$.

Recall $v_1 \dots v_p$ are the $x \in C$ not seen in S_j. The pairs are as follows:

	x	$f_i(x)$	$f_{i'}(x)$	
S_j	x_1	1	1	← all possible equal sequences.
	x_2	0	0	
	$\vdots$	$\vdots$	$\vdots$	
	x_m	1	1	
Not in S_j	v_1	1	0	← complementary sequences.
	v_2	0	1	
	$\vdots$	1	0	
	v_p	1	0	

C

$$\frac{1}{2^{2m}} \sum_{i=1}^{2^{2m}} \underbrace{\mathbb{1}\left(f_i(v_r) \neq A(S_j^i)(v_r) \right)}_{\substack{= 1 \text{ for half of } f_i \\ = 0 \text{ for other half of } f_{i'}}} = \frac{1}{2^{2m}} \frac{2^{2m}}{2} = \frac{1}{2}$$

$\boxed{\text{This ends proof of No Free Lunch.}}$

Lemma (used previously).

$$\mathbb{E}\left(L_{\mathcal{D}_i}(A(S_i))\right) \geq \frac{1}{4} \Rightarrow \text{Prob}\left(L_{\mathcal{D}_i}(A(S_i)) \geq \frac{1}{8}\right) \geq \frac{1}{7}$$

Proof.

It suffices to show that for a r.v $Z \in [0,1]$ we must have:

$$\mathbb{P}(Z \geq a) \geq \frac{\mathbb{E}(Z) - a}{1-a}.$$

Choose $Z = L_{\mathcal{D}_i}(A(S_i))$ & $a = \frac{1}{8}$.

The proof is an application of Markov's inequ for a positive r.v $X \geq 0$:

$$\mathbb{P}(X \geq b) \leq \frac{\mathbb{E}(X)}{b}.$$

Now let $X = 1 - Z$. This is ≥ 0. Thus

$$\mathbb{P}(1 - Z \geq b) \leq \frac{1 - \mathbb{E}(Z)}{b}$$

$$\Rightarrow \mathbb{P}(1 - Z \leq b) \geq 1 - \frac{1 - \mathbb{E}(Z)}{b}$$

$$= \frac{b - 1 + \mathbb{E}(Z)}{b}$$

Take $b = 1 - a$. We get

$$\mathbb{P}(\underbrace{1 - Z \leq 1 - a}_{Z \geq a.}) \geq \frac{\mathbb{E}(Z) - a}{1 - a}$$

$\blacksquare$

Remarks on Bias-Complexity Tradeoff.

Let $h_S = \mathrm{ERM}(S) = \underset{h}{\mathrm{argmin}}\ L_S(h)$

the hypothesis class given by the ERM rule.

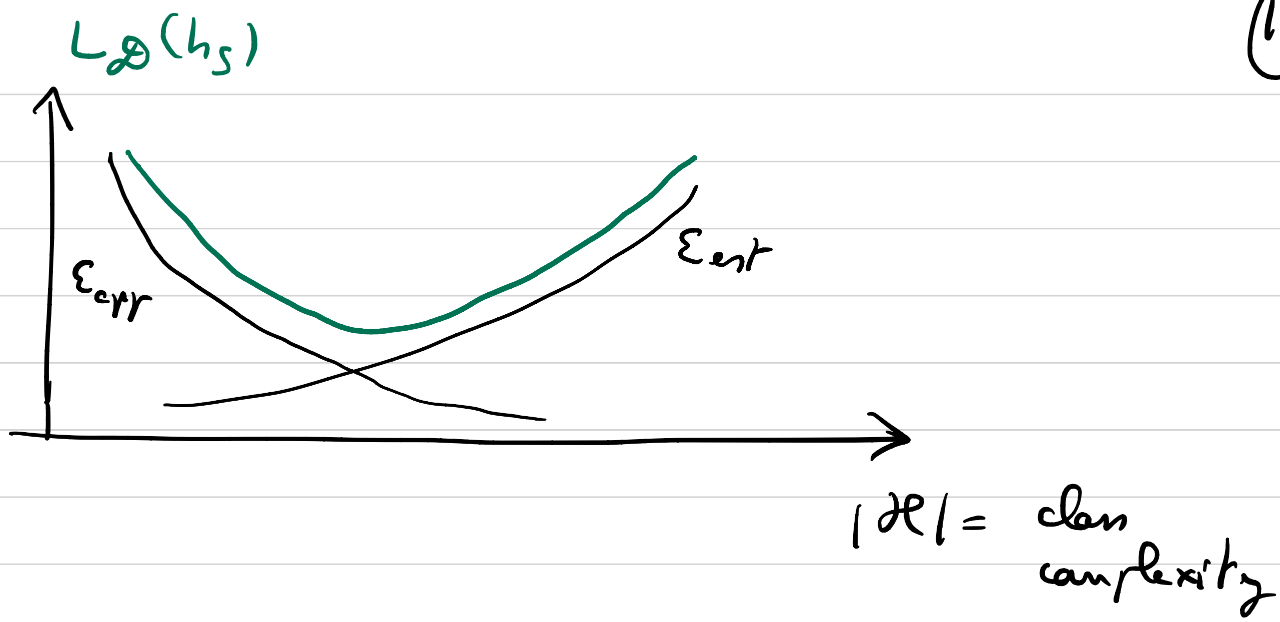
The test error of an ERM predictor is $L_{\mathcal{D}}(h_S)$.

We can decompose it as:

$$L_{\mathcal{D}}(h_S) = \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\substack{\text{approx error} \\ \text{or "bias" of} \\ \text{hyp class } = \varepsilon_{app}}} + \underbrace{\left(L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)\right)}_{\substack{\text{estimation error} \\ \text{of ERM.} = \varepsilon_{est}.}}$$

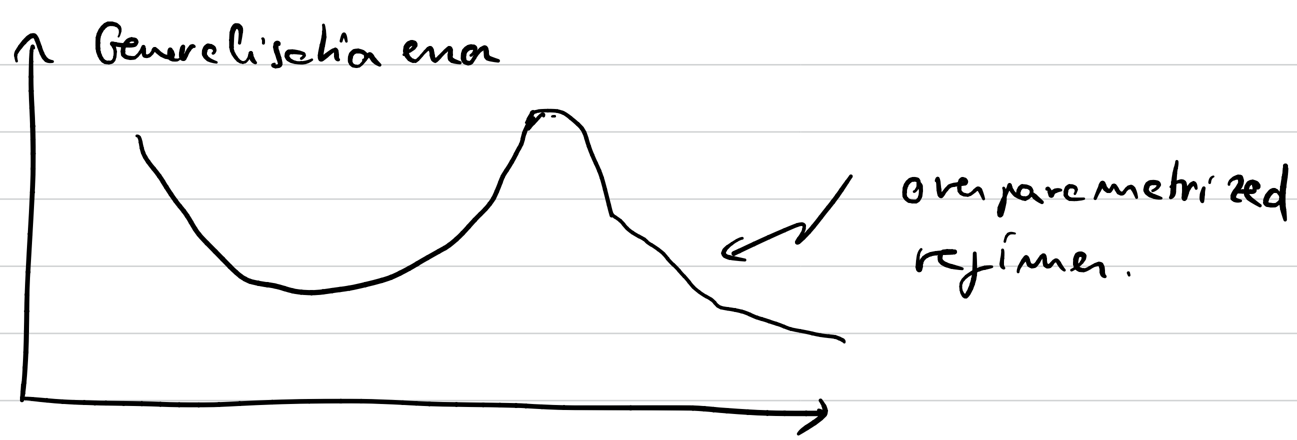
- We have obviously that $\varepsilon_{app} \searrow$ as $|\mathcal{H}| \nearrow$
- We also expect usually $\varepsilon_{est} \nearrow$ as $|\mathcal{H}| \nearrow$.

For example the generalisation bound shown for finite classes suggests $\varepsilon_{est} \sim \sqrt{\frac{1}{2m} \log \frac{2|\mathcal{H}|}{\delta}} \nearrow$ as $|\mathcal{H}| \nearrow$



This is the traditional bias-complexity tradeoff picture. Increasing $\mathcal{H}$ decreasing our bias but at some point may lead to overfitting.

This picture has been challenged in recent years. Indeed there are situations displaying a "double descent" phenomenon



and it is recognized that the picture is much richer.