DEFINITION 2.1 (The Realizability Assumption) There exists $h^* \in \mathcal{H}$ s.t. $L_{(\mathcal{D},f)}(h^*) = 0$. Note that this assumption implies that with probability 1 over random samples, S, where the instances of S are sampled according to $\mathcal{D}$ and are labeled by f, we have $L_S(h^*) = 0$.

More generally, we now define *Probably Approximately Correct* (PAC) learning.

DEFINITION 3.1 (PAC learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$ and $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

Formally, given a domain subset, $A \subset \mathcal{X}$, the probability distribution, $\mathcal{D}$, assigns a number, $\mathcal{D}(A)$, which determines how likely it is to observe a point $x \in A$. In many cases, we refer to A as an event and express it using a function $\pi : \mathcal{X} \to \{0,1\}$, namely, $A = \{x \in \mathcal{X} : \pi(x) = 1\}$. In that case, we also use the notation $\mathbb{P}_{x\sim\mathcal{D}}[\pi(x)]$ to express $\mathcal{D}(A)$.

We define the error of a prediction rule, $h : \mathcal{X} \to \mathcal{Y}$, to be

$$L_{(\mathcal{D},f)}(h) \stackrel{\text{def}}{=} \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq f(x)] \stackrel{\text{def}}{=} \mathcal{D}(\{x : h(x) \neq f(x)\}). \quad (2.1)$$

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exists a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: for every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

THEOREM 6.7 (The Fundamental Theorem of Statistical Learning) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the 0 − 1 loss. Then, the following are equivalent:*

1. *$\mathcal{H}$ has the uniform convergence property.*
2. *Any ERM rule is a successful agnostic PAC learner for $\mathcal{H}$.*
3. *$\mathcal{H}$ is agnostic PAC learnable.*
4. *$\mathcal{H}$ is PAC learnable.*
5. *Any ERM rule is a successful PAC learner for $\mathcal{H}$.*
6. *$\mathcal{H}$ has a finite VC-dimension.*

DEFINITION 4.1 (ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon.$$

DEFINITION 4.3 (Uniform Convergence) We say that a hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ examples drawn i.i.d. according to $\mathcal{D}$, then, with probability of at least $1-\delta$, S is ϵ-representative.

We have already seen that $1 \to 2$ in Chapter 4. The implications $2 \to 3$ and $3 \to 4$ are trivial and so is $2 \to 5$. The implications $4 \to 6$ and $5 \to 6$ follow from the No-Free-Lunch theorem. The difficult part is to show that $6 \to 1$. The proof is based on two main claims:

- If $\text{VCdim}(\mathcal{H}) = d$, then even though $\mathcal{H}$ might be infinite, when restricting it to a finite set $C \subset \mathcal{X}$, its "effective" size, $|\mathcal{H}_C|$, is only $O(|C|^d)$. That is, the size of $\mathcal{H}_C$ grows polynomially rather than exponentially with $|C|$. This claim is often referred to as *Sauer's lemma*, but it has also been stated and proved independently by Shelah and by Perles. The formal statement is given in Section 6.5.1 later.
- In Section 4 we have shown that finite hypothesis classes enjoy the uniform convergence property. In Section 6.5.2 later we generalize this result and show that uniform convergence holds whenever the hypothesis class has a "small effective size." By "small effective size" we mean classes for which $|\mathcal{H}_C|$ grows polynomially with $|C|$.

① → ② Assume if $\mathbb{P}_{S\sim\mathcal{D}}(|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon, \forall h \in \mathcal{H}) \geq 1-\delta$ for $|S| \geq m_{\mathcal{H}}(\epsilon,\delta)$ h_S is the ERM rule

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \epsilon \leq L_S(h) + \epsilon \leq L_{\mathcal{D}}(h) + 2\epsilon.$$

$$\Rightarrow L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{D}} L_{\mathcal{D}}(h) + 2\epsilon$$

② → ③ Obvious

③ → ④ Realizability obvious

② → ⑤ Obvious

④ → ⑥

⑤ → ①

THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*