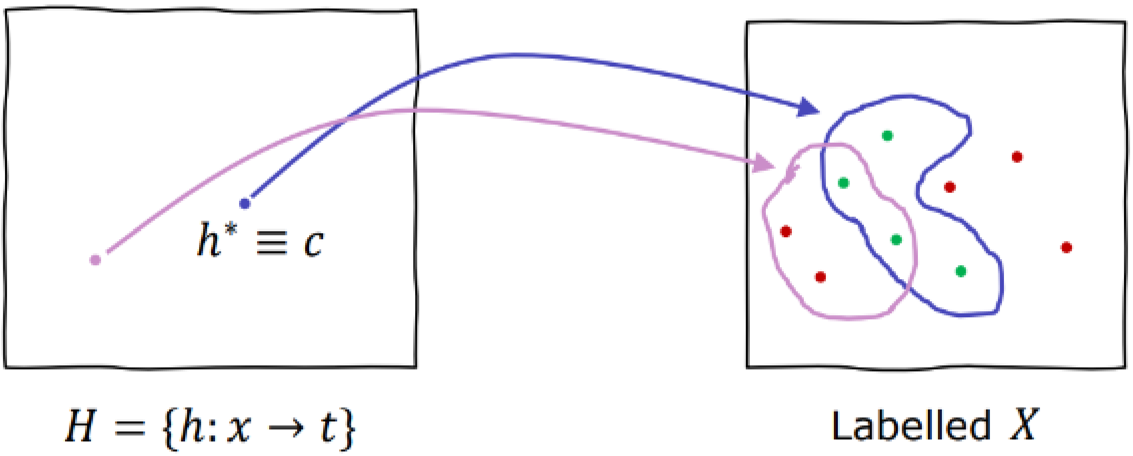
# Computational Learning Theory

It aims at studying the general laws of inductive learning, by modeling:

- Complexity of hypothesis space
- Bound on training samples
- Bound on accuracy
- Probability of succesfull learning
- ...
  This allows to answer to questions like:
- How many training samples do a learner need to converge (with some probability) to a successful (with some minimum accuracy) hypothesis?
- How many training samples will be misclassified by the learner before converging to a succesful hypothesis?
  We expect something that confirm what we obtain from machine learning library results
  The problem is to define the function to map a given input in a target and find the function and the hypotesis space that correspond to the examples.



A learner ($L$) wants to learn a concept(c) that maps the data in the input space ($X$) to a target (t)
Let assume that L found an hypothesis $h*$ with no errors on the training data
How many training samples of $X$ are necessary to be sure that $L$ actually learned the true concept i.e. $h^* \equiv c$

## No Free Lunch Theorems

On average every ML model will behave like a random guess. We are focusing on classification and binary classification.
Let $ACC_G(L)$ be the **generalization accuracy** of learner $L$ i.e., the accuracy of L on samples that are not in the training set
Let $\mathcal{F}$ be the set of all the possible concepts
For any learner L and any possible training set:

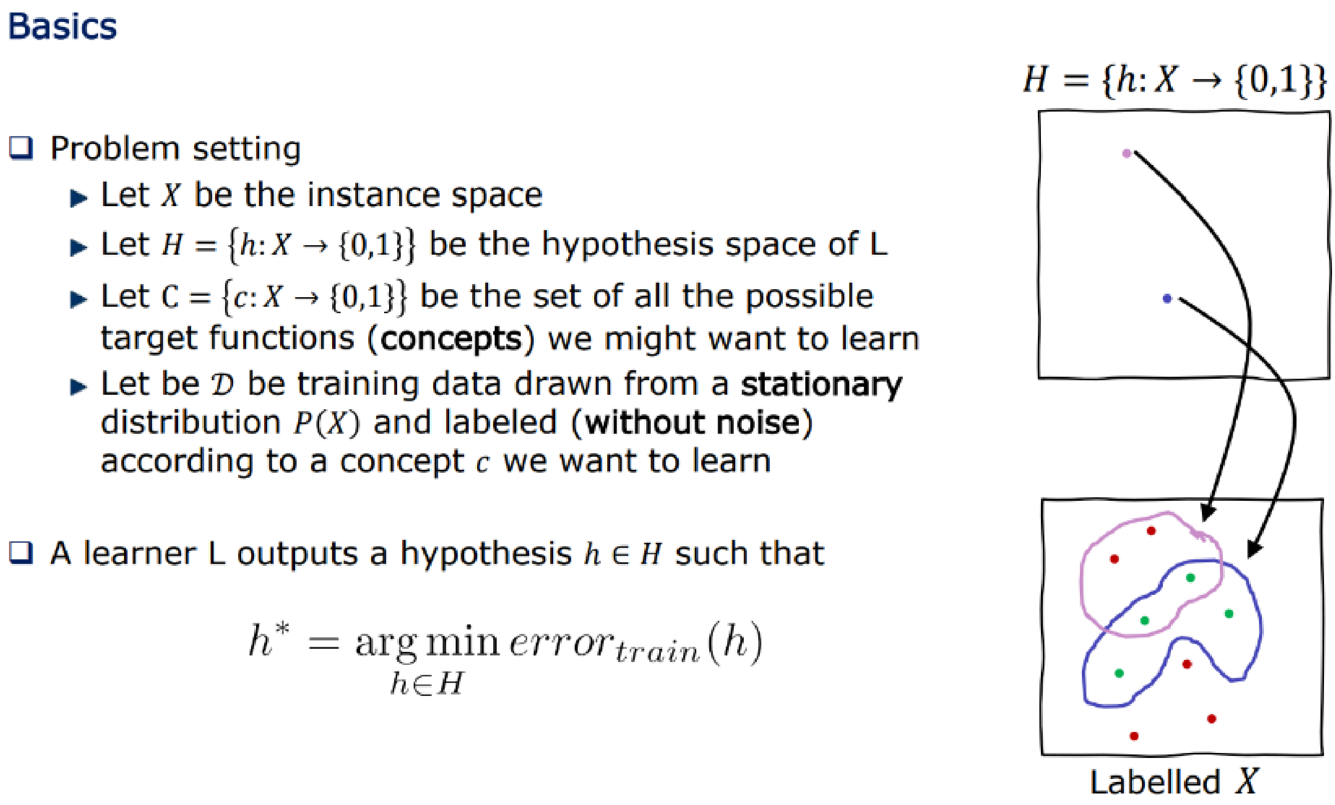
$$\frac{1}{\mathcal{F}} = \sum_{\mathcal{F}} ACC_G(L) = \frac{1}{2}$$

- Proof Sketch: for every concept $f$ where $ACC_G(L) = 0.5 + \delta$ , exists a concept $f$' where $ACC_G(L) = 0.5 + \delta : \forall \mathbf{x} \in \mathcal{D}, f'(\mathbf{x}); \forall x \notin \mathcal{D}, f'(\mathbf{x}) \neq f(\mathbf{x})$
- Corollary: for any two learners $L_1$ and $L_2$, if $\exists f$ where $ACC_G(L_1) > ACC_G(L_2)$ then $\exists f'$ where $ACC_G(L_2) > ACC_G(L_1)$

This accomplish two factors:

- There is not a ML algorithm that is superior to the other(The solution can be derived after the definition of the problem)
- We always operate under assumptions. We assume that the hypotesis space contains a good approsimation of the function to derive. We use some techniques to reduce the complexity of the model(We want a model not too complex, no overfitting of the data)

## Probably Learning an Approximately Correct Hypothesis



### Basics

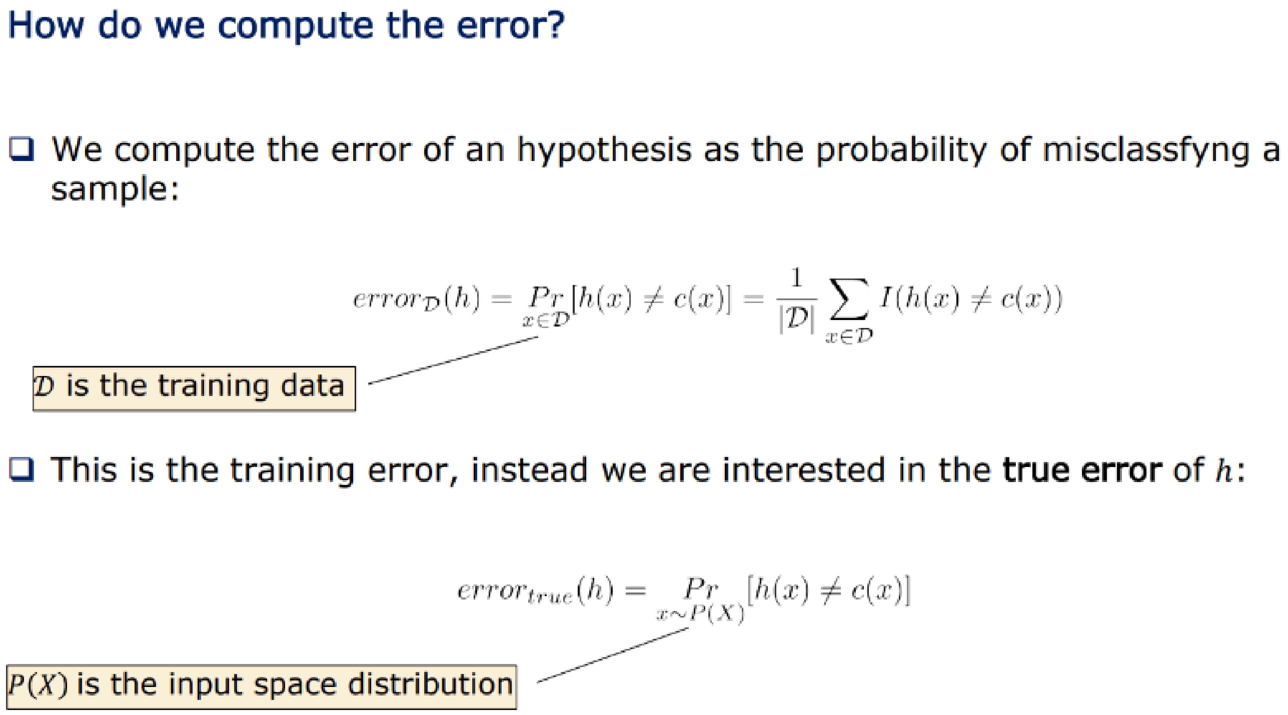
❑ Problem setting

- Let $X$ be the instance space
- Let $H = \{h: X \rightarrow \{0,1\}\}$ be the hypothesis space of L
- Let C = $\{c: X \rightarrow \{0,1\}\}$ be the set of all the possible target functions (**concepts**) we might want to learn
- Let be $\mathcal{D}$ be training data drawn from a **stationary** distribution $P(X)$ and labeled (**without noise**) according to a concept $c$ we want to learn

❑ A learner L outputs a hypothesis $h \in H$ such that

$$h^* = \arg\min_{h \in H} error_{train}(h)$$

Binary classification is the more convenient setting to compute this analysis.



How do we compute the error?

❑ We compute the error of an hypothesis as the probability of misclassfyng a sample:

$$error_{\mathcal{D}}(h) = \Pr_{x \in \mathcal{D}}[h(x) \neq c(x)] = \frac{1}{|\mathcal{D}|}\sum_{x \in \mathcal{D}} I(h(x) \neq c(x))$$

$\mathcal{D}$ is the training data

❑ This is the training error, instead we are interested in the **true error** of $h$:

$$error_{true}(h) = \Pr_{x \sim P(X)}[h(x) \neq c(x)]$$

$P(X)$ is the input space distribution

We want to compute on all possible data and so we calculate the true error
The error is also defined by the region covered only y the defined space or the concept(This definition doesn't comprehend the fact that the true error cannot be zero, this lead to not having idea of the sign of my distribution, I can make mistake on a region not much populated or density populated)
We want to find a limit to the true error:

- $error_{true}$ is the probability of making a mistake on a sample
- we can compute $error_{\mathcal{D}}$ that is the average error probability on $\mathcal{D}$
- assuming a Bernoulli distribution for the error probability, the 95% CI is:

$$error_{true}(h) = error_{\mathcal{D}}(h) \pm 1.96\sqrt{\frac{error_{\mathcal{D}}(h)(1 - error_{\mathcal{D}}(h))}{n}}$$

  This isn't correct because this is not a true confident interval but not work for us as our training error are completely independent from the actual data used. No reason to suppose our model will perform the same on unseen data and training data. But we trained our model specifically to behave correctly on unseen data.

## Version Space

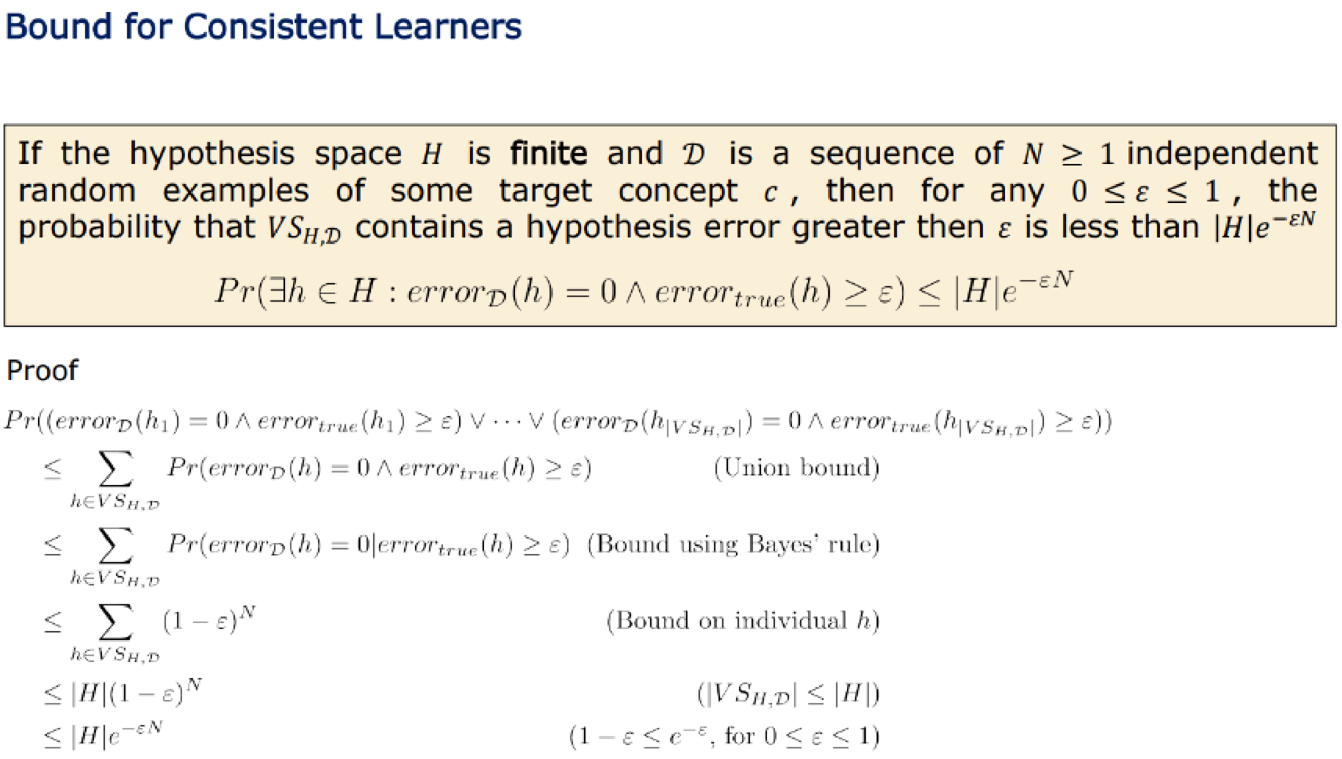
A hypothesis h is consistent with a training dataset $\mathcal{D}$ of the concept c if and only if h(x)=c(x) for each training sample in $\mathcal{D}$(We have zero training error)

$$Consistent(h, \mathcal{D}) \overset{def}{=} \forall \langle x, c(x) \rangle \in \mathcal{D}, h(x) = c(x)$$

The version space $VS_{H,D}$ with respect to hypothesis space H and labelled dataset $\mathcal{D}$, is the subset of hypotheses in H consistent with $\mathcal{D}$

$$VS_{h,\mathcal{D}} \overset{def}{=} \{h \in H | Consistent(h, \mathcal{D})\}$$

From now on, we consider only **consistent learners**, that always output a **consistent hypothesis**, i.e., an hypothesis in $VS_{H,\mathcal{D}}$ assuming is not empty



## Bound for Consistent Learners

If the hypothesis space $H$ is **finite** and $\mathcal{D}$ is a sequence of $N \geq 1$ independent random examples of some target concept $c$ , then for any $0 \leq \varepsilon \leq 1$ , the probability that $VS_{H,\mathcal{D}}$ contains a hypothesis error greater then $\varepsilon$ is less than $|H|e^{-\varepsilon N}$

$$Pr(\exists h \in H : error_{\mathcal{D}}(h) = 0 \wedge error_{true}(h) \geq \varepsilon) \leq |H|e^{-\varepsilon N}$$

Proof

$$\begin{aligned}
&Pr((error_{\mathcal{D}}(h_1) = 0 \wedge error_{true}(h_1) \geq \varepsilon) \vee \cdots \vee (error_{\mathcal{D}}(h_{|VS_{H,\mathcal{D}}|}) = 0 \wedge error_{true}(h_{|VS_{H,\mathcal{D}}|}) \geq \varepsilon)) \\
&\leq \sum_{h \in VS_{H,\mathcal{D}}} Pr(error_{\mathcal{D}}(h) = 0 \wedge error_{true}(h) \geq \varepsilon) \qquad \text{(Union bound)} \\
&\leq \sum_{h \in VS_{H,\mathcal{D}}} Pr(error_{\mathcal{D}}(h) = 0 | error_{true}(h) \geq \varepsilon) \quad \text{(Bound using Bayes' rule)} \\
&\leq \sum_{h \in VS_{H,\mathcal{D}}} (1-\varepsilon)^N \qquad \text{(Bound on individual } h\text{)} \\
&\leq |H|(1-\varepsilon)^N \qquad (|VS_{H,\mathcal{D}}| \leq |H|) \\
&\leq |H|e^{-\varepsilon N} \qquad (1-\varepsilon \leq e^{-\varepsilon}, \text{ for } 0 \leq \varepsilon \leq 1)
\end{aligned}$$

In practice:

- Let say that $\delta$ is the probability to have $error_{true} > \varepsilon$ for a consistent hypothesis:

$$|H|e^{-\varepsilon N} \leq \delta$$

- We can bound N after setting $\varepsilon$ and $\delta$:

$$N \geq \frac{1}{\varepsilon}(ln|H| + ln(\frac{1}{\delta}))$$

- We can bound $\varepsilon$ after setting N and $\delta$:

$$\varepsilon \geq \frac{1}{N}(ln|H| + ln(\frac{1}{\delta}))$$

## Probably Learning an Approximately Correct Hypothesis

Considering a class $C$ of possible target concepts defined over an instance space $X$ with an encoding lenght $M$, and a learner $L$ using an hypotesis space $H$ we define:

$C$ is PAC-learnable by $L$ using $H$ if for all $c \in C$, for any distribution $P(X)$, $\varepsilon$ (such that $0 < \varepsilon < 1/2$), and $\delta$ (such that $0 < \delta < 1/2$), learner $L$ will with a probability at least $(1 - \delta)$ output a hypotehsis $h \in H$ such that $error_{true}(h) \leq \varepsilon$, in time that is polynomial in $1/\varepsilon$, $1/\delta$, $M$, and $size(c)$.

So, PAC-learnability is only about computational complexity? What about the complexity with respect to the number of training samples $N$?

**A sufficient condition to prove PAC-learnability is proving that a learner L requires only a polynomial number of training examples, and processing per example is polynomial**

- So far, we assumed that $c \in H$, or at least that $VS_{H,\mathcal{D}}$ is not empty, and the learner L will always output a hypothesis $h$ such that $error_{\mathcal{D}}(h) = 0$
- But in general (**agnostic**) leaner will output a hypothesis $h$ such that $error_{\mathcal{D}}(h) > 0$
- Can we bound $error_{true}(h)$ given $error_{\mathcal{D}}(h)$ ?

If the hypothesis space $H$ is **finite** and $\mathcal{D}$ is a sequence of $N \geq 1$ i.i.d. examples of some target concept $c$, then for any $0 \leq \varepsilon \leq 1$, and for any learned hypothesis h, the probability that $error_{true}(h) - error_{\mathcal{D}}(h) > \varepsilon$ is less than $|H|e^{-2N\varepsilon^2}$

$$Pr(\exists h \in H : error_{true}(h) > error_{\mathcal{D}}(h) + \varepsilon) \leq |H|e^{-2N\varepsilon^2}$$

Proof

- **Additive Hoeffding Bound**: let $\hat{\theta}$ be the empirical mean of $N$ i.i.d. Bernoulli random variables with mean $\theta$:

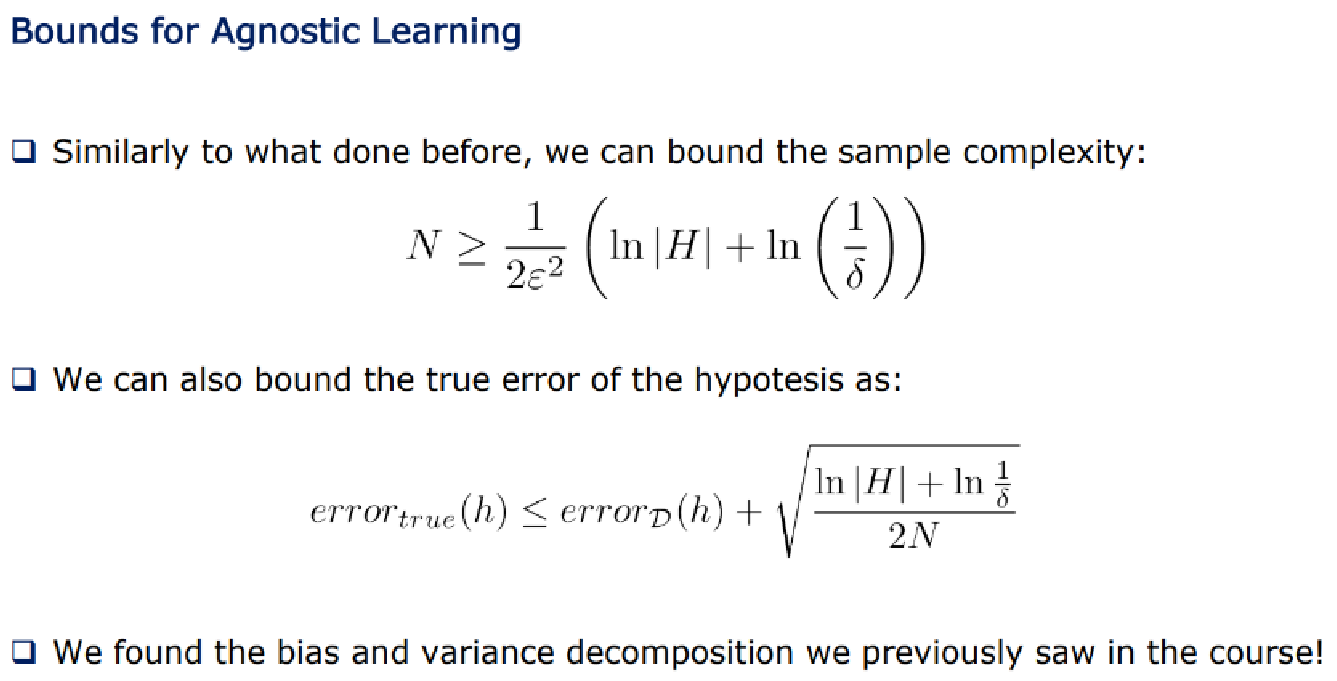
$$Pr(\theta > \hat{\theta} + \varepsilon) \leq e^{-2N\varepsilon^2}$$

- So for any **single** hypothesis h:

$$Pr(error_{true}(h) > error_{\mathcal{D}}(h) + \varepsilon) \leq e^{-2N\varepsilon^2}$$

- As we want this to be true for all the hypothesis in $H$:

$$Pr(\exists h \in H : error_{true}(h) > error_{\mathcal{D}}(h) + \varepsilon) \leq |H|e^{-2N\varepsilon^2}$$



## Bounds for Agnostic Learning

- Similarly to what done before, we can bound the sample complexity:

$$N \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$$

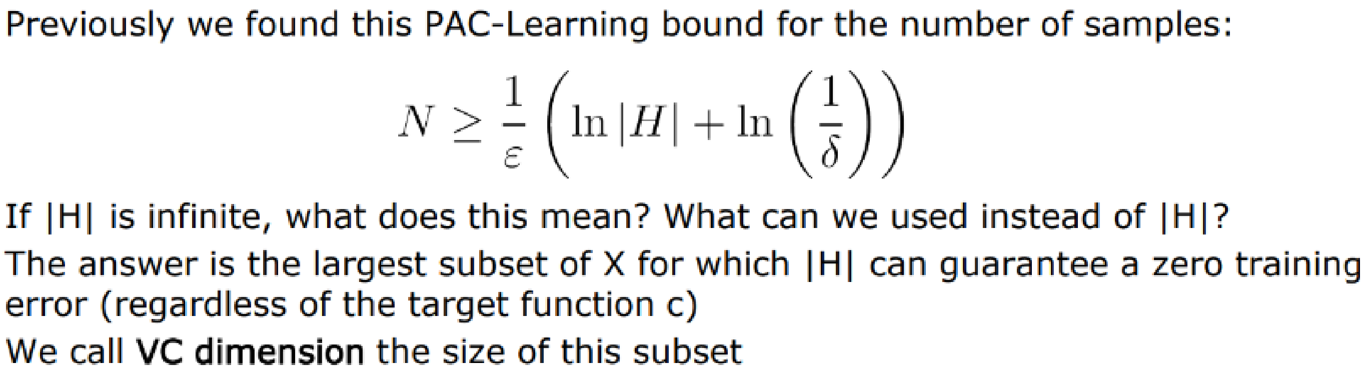
- We can also bound the true error of the hypotesis as:

$$error_{true}(h) \leq error_{\mathcal{D}}(h) + \sqrt{\frac{\ln|H| + \ln\frac{1}{\delta}}{2N}}$$

- We found the bias and variance decomposition we previously saw in the course!

error = noise+$bias^2$+ variance (The bias influence the most the error)

## PAC-Learning with Infinite Hypotheses Spaces



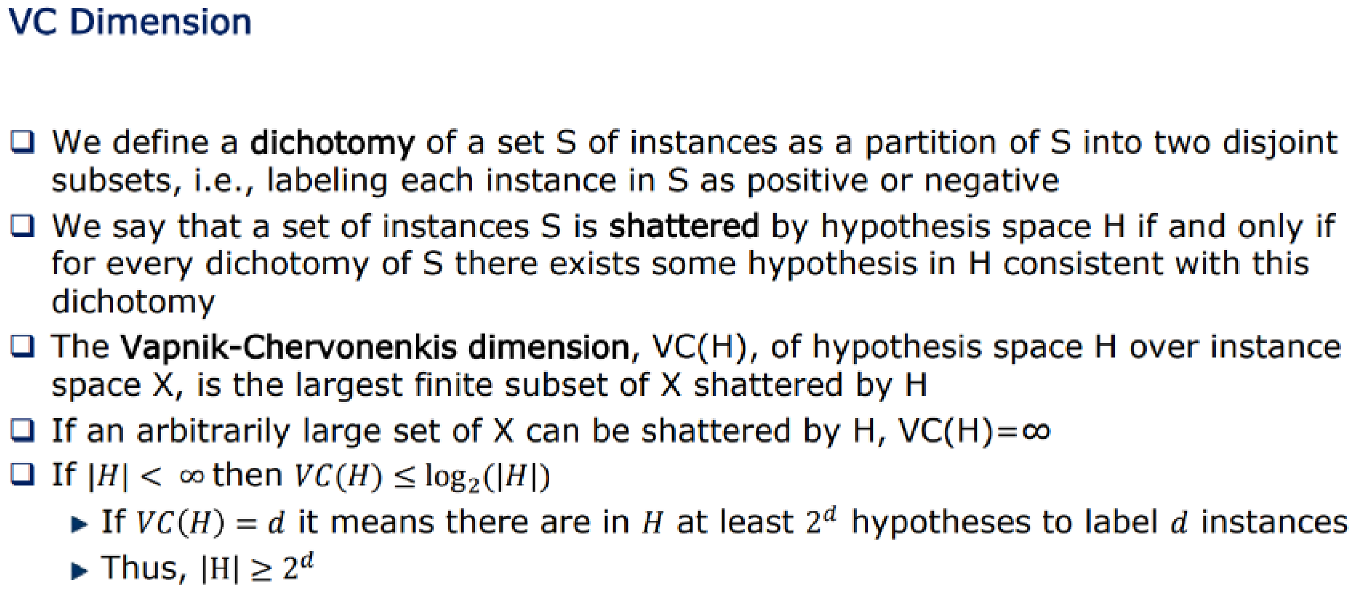
Previously we found this PAC-Learning bound for the number of samples:

$$N \geq \frac{1}{\varepsilon}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right)$$

If |H| is infinite, what does this mean? What can we used instead of |H|?

The answer is the largest subset of X for which |H| can guarantee a zero training error (regardless of the target function c)

We call **VC dimension** the size of this subset



## VC Dimension

- We define a **dichotomy** of a set S of instances as a partition of S into two disjoint subsets, i.e., labeling each instance in S as positive or negative
- We say that a set of instances S is **shattered** by hypothesis space H if and only if for every dichotomy of S there exists some hypothesis in H consistent with this dichotomy
- The **Vapnik-Chervonenkis dimension**, VC(H), of hypothesis space H over instance space X, is the largest finite subset of X shattered by H
- If an arbitrarily large set of X can be shattered by H, VC(H)=∞
- If $|H| < \infty$ then $VC(H) \leq \log_2(|H|)$
  - If $VC(H) = d$ it means there are in $H$ at least $2^d$ hypotheses to label $d$ instances
  - Thus, $|\mathrm{H}| \geq 2^d$