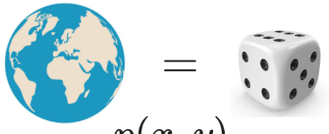
Machine Learning Fundamentals - LS2026
VC dimension

Czech Technical University in Prague
V. Franc

Big Picture of ERM based learning

Empirical Risk Minimization



Hypothesis class

$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h \colon \mathcal{X} \to \mathcal{Y}\}$

Loss function

$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$

Training set

i.i.d. data

$T_m = ((x_1, y_1), \ldots, (x_m, y_m)) \sim p^m$

ERM Learning Algorithm

Learning algorithm $A \colon (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$ finding a predictor in $\mathcal{H}$ which minimizes the empirical error:

$\hat{R}(T_m, h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y_i, h(x_i; \theta))$

Predictor

$\hat{y} = h_m(x)$

Error decomposition

Predictor Error = Estimation Error + Approximation Error + Bayes Error

Uniform Law of Large Numbers

Empirical risk of all member in $\mathcal{H}$ is a good proxy of the true risk.

ULLN holds for $\mathcal{H} \iff \mathcal{H}$ is PAC learnable

$\iff$ ERM is succesful PAC learner

VC dimension

0/1-loss + binary classifier

$\mathrm{VCdim} \colon \{-1, +1\}^{\mathcal{X}} \to \mathbb{N}$

$\mathrm{VCdim}(\mathcal{H}) < \infty \iff$ ULLN holds for $\mathcal{H}$

PAC learning

A hypothesis class $\mathcal{H}$ is PAC learnable if there exists an lagorithm that, with high probability, can make the estimation error arbitrarily small given sufficiently many examples

Hypothesis class represents our prior knowledge

"Too complex" hypothesis class

E.g. Memorizer

$\mathcal{H}$ is not PAC learnable

$\ldots$

Finite hypothesis class

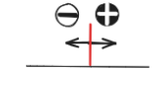
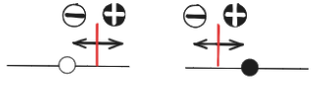
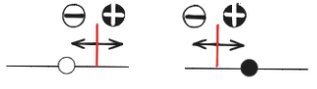
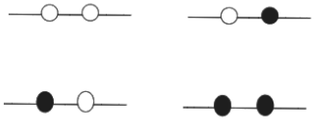
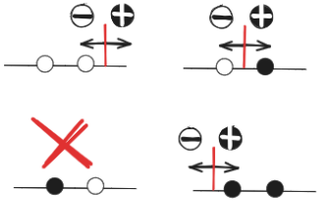
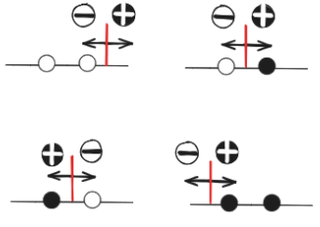
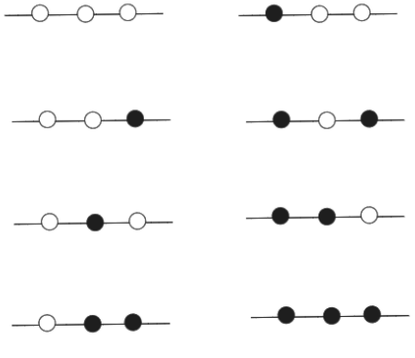
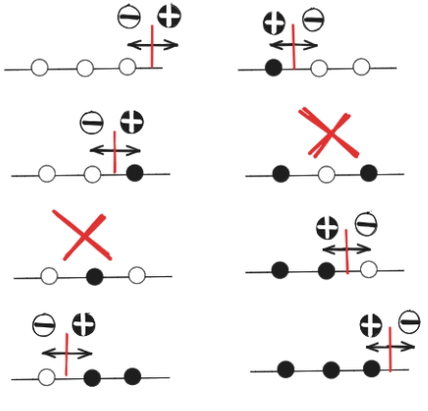
$\mathcal{H} = \{h_1, h_2, \ldots, h_{\mathcal{H}}\}$

$\mathcal{H}$ is PAC learnable

Vapnik-Chervonenkis Dimension

Threshold classifiers: $\mathcal{H}_1 = \{h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$

Oriented threshold classifiers: $\mathcal{H}_2 = \{\{h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\} \cup \{h(x) = \text{sign}(\theta - x) \mid \theta \in \mathbb{R}\}\}$

#inputs m	possible label configurations 2^m	$\mathcal{H}_1 = \{$  $\}$	$\mathcal{H}_2 = \{$  $\}$
1		 VCdim = 1	
2			 VCdim = 2
3			

Vapnik-Chervonenkis Dimension

- The VC dimension quantifies the complexity (capacity) of a hypothesis class $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$.

Definition (Shattering): Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ and let $\{x_1, \ldots, x_m\} \subset \mathcal{X}^m$ be a set of m input points. The set $\{x_1, \ldots, x_m\}$ is shattered by $\mathcal{H}$ if, for every labeling $y \in \{-1, +1\}^m$, there exists a hypothesis $h \in \mathcal{H}$ such that

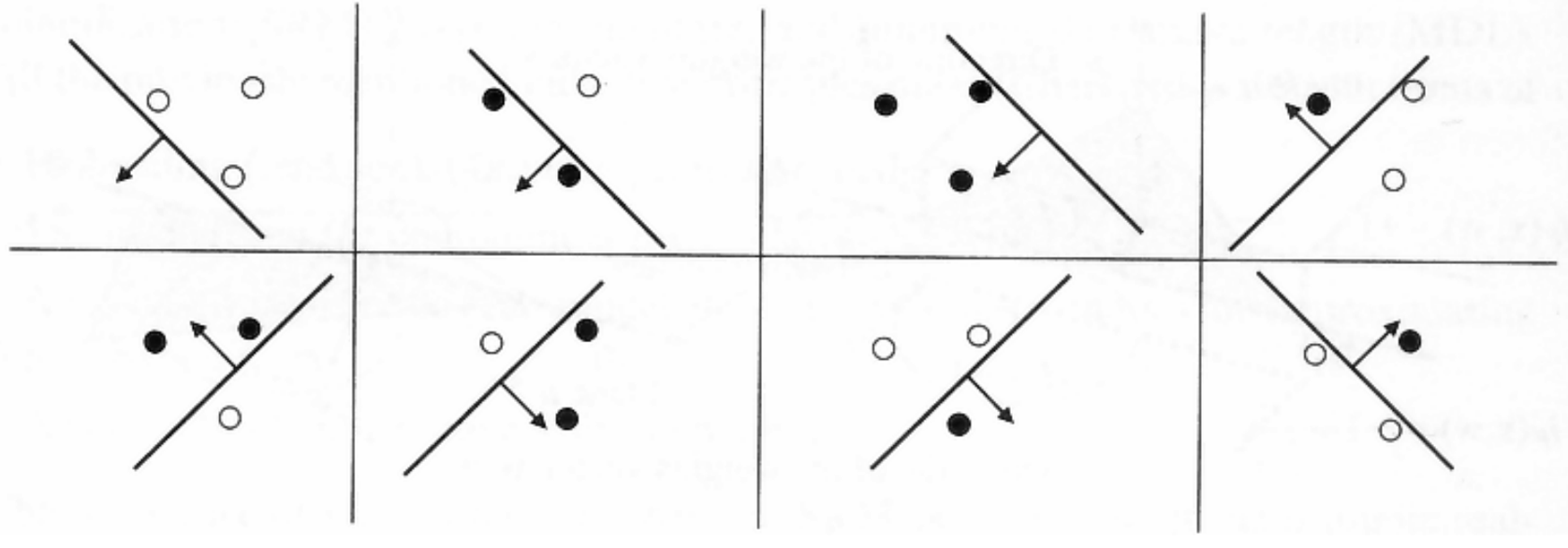
$$h(x_i) = y_i, \quad \forall\, i \in \{1, \ldots, m\}.$$

Definition (VC dimension): Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$. The Vapnik–Chervonenkis dimension of $\mathcal{H}$, denoted $\mathrm{VCdim}(\mathcal{H})$, is the cardinality of the largest set of points from $\mathcal{X}$ that can be shattered by $\mathcal{H}$.

Vapnik-Chervonenkis Dimension

Theorem: The VC-dimension of the hypothesis class of all two-class linear classifiers in d-dimensional feature space $\mathcal{H} = \{h(x; \boldsymbol{w}, b) = \text{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b) \mid (\boldsymbol{w}, b) \in \mathbb{R}^{d+1})\}$ is $\text{VCdim}(\mathcal{H}) = d + 1$.

Example for $n = 2$-dimensional feature space



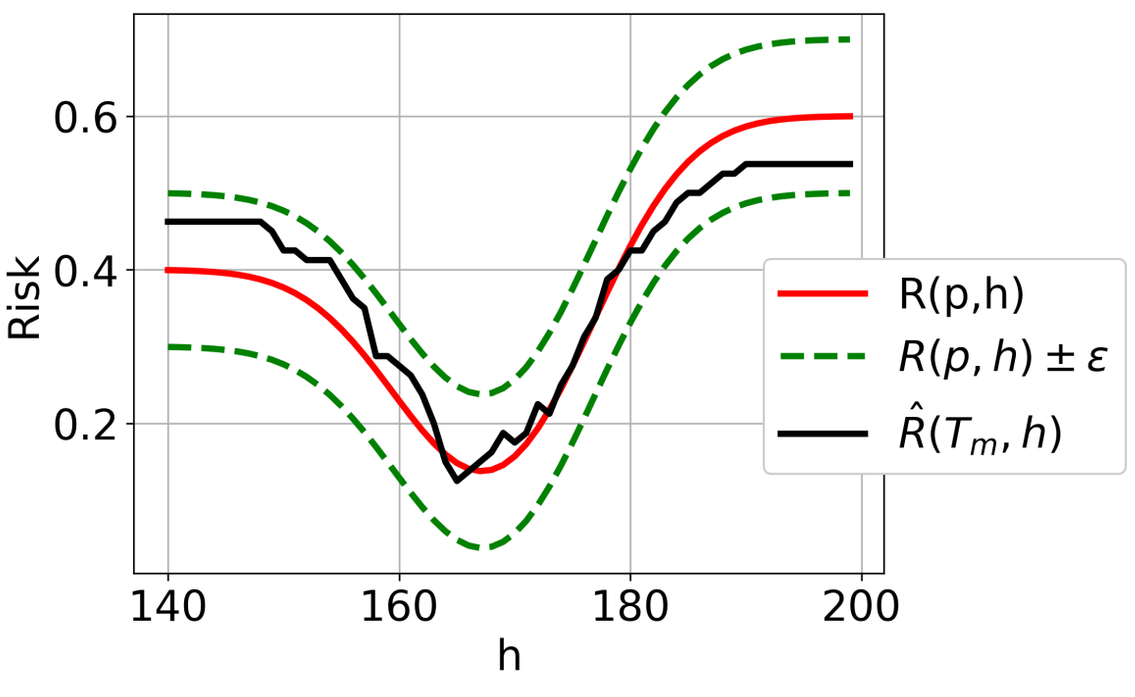
Quiz: Let $\mathcal{H} = \left\{h(x) = \hat{y}\ ,\ (\hat{x}, \hat{y}) = \arg\min_{(x',y') \in T_m} \|x' - x\|\right\}$ be a space of Nearest-Neighbor classifiers. What is the VC dimension of $\mathcal{H}$?

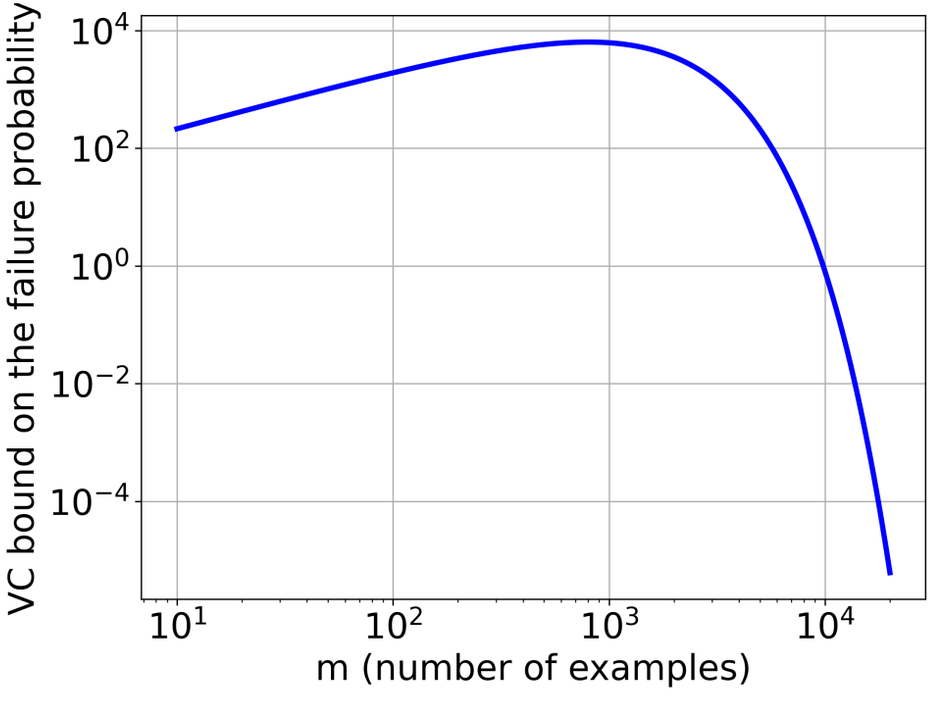
Finite VC Dimension implies Uniform Law of Large Numbers

Theorem: Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ be a hypothesis class of binary classifiers, and let $\ell(y, y') = [\![y \neq y']\!]$ be the $0/1$ loss. Let $T_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be a sample i.i.d. drawn from a distribution $p(x, y)$. Assume that $\mathcal{H}$ has a finite VC dimension, $\mathrm{VCdim}(\mathcal{H}) < \infty$. Then, Uniform Law of Large Numbers applies for $\mathcal{H}$: for every $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} \left| R(p, h) - \hat{R}(T_m, h) \right| \geq \varepsilon \right) \leq 4 \left(\frac{2\, e\, m}{\mathrm{VCdim}(\mathcal{H})} \right)^{\mathrm{VCdim}(\mathcal{H})} e^{-\frac{m\, \varepsilon^2}{8}}$$

Example: $\mathcal{H} = \{h(x; \theta) = \mathrm{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$, $\mathrm{VCdim}(\mathcal{H}) = 1$, $\varepsilon = 0.1$





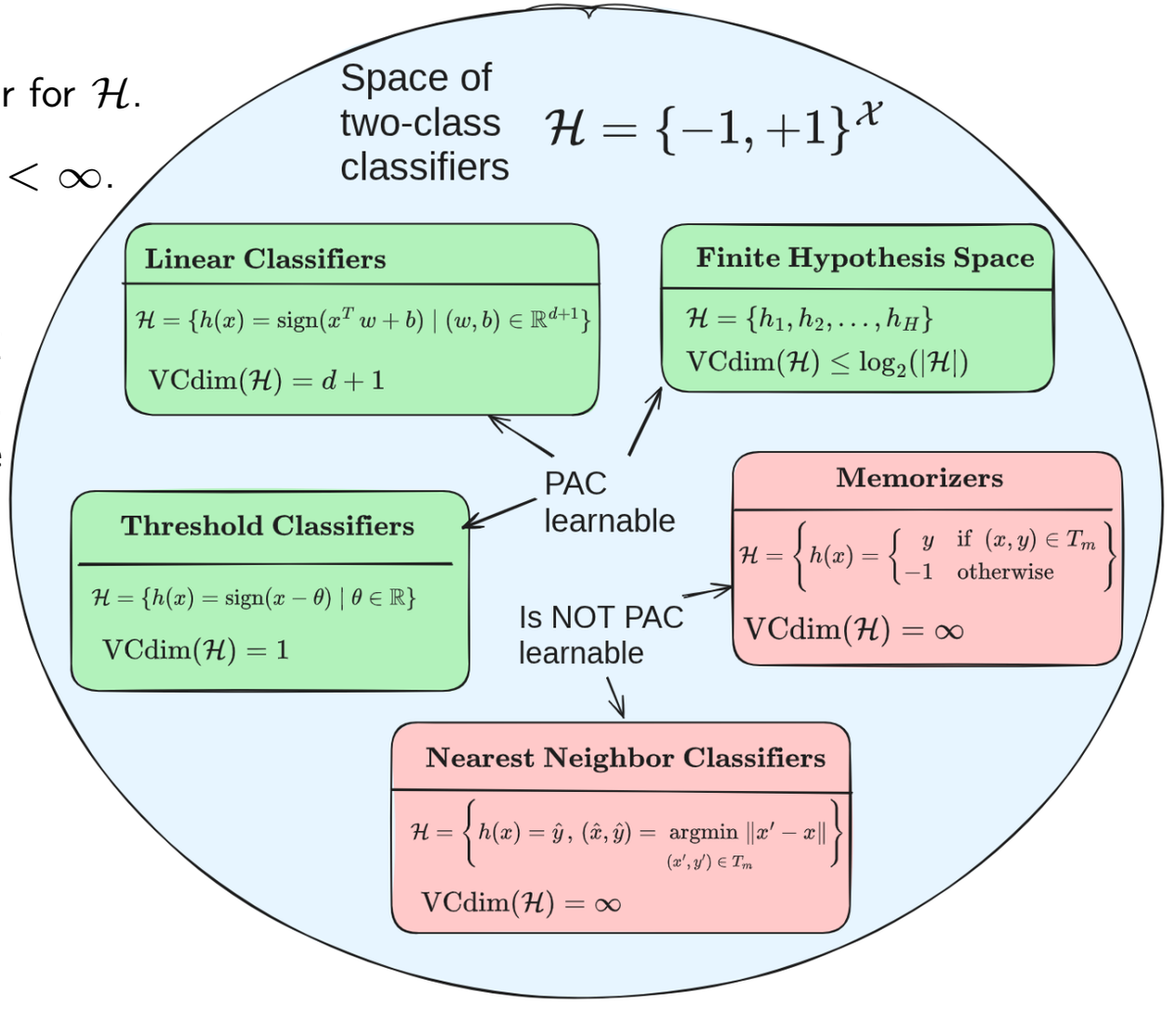
Fundamental Theorem of PAC Learning

Theorem: Let $\mathcal{H} \subset \{-1,+1\}^{\mathcal{X}}$ be a hypothesis class of functions from $\mathcal{X}$ to $\{-1,+1\}$ and let $\ell(y,y') = [\![y \neq y']\!]$ be the 0/1-loss function. Then, the following statements are equivalent:

- Uniform Law of Large numbers holds for $\mathcal{H}$.
- $\mathcal{H}$ is PAC learnable.
- ERM algorithm is a successful PAC learner for $\mathcal{H}$.
- $\mathcal{H}$ has finite VC dimension, $\mathrm{VCdim}(\mathcal{H}) < \infty$.

Sample complexity: Assume the VC dimension of $\mathcal{H}$ is finite, $\mathrm{VCdim}(\mathcal{H}) < \infty$. Then, there is a constant C such that the sample complexity is

$$m^{\mathcal{H}}_{\mathrm{pac}}(\varepsilon,\delta) \leq C \frac{\mathrm{VCdim}(\mathcal{H}) + \log(\frac{1}{\delta})}{\varepsilon^2}$$



VC Dimension Generalization Bound

Theorem. Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ be a hypothesis class of binary classifiers, and let $\ell(y, y') = [\![y \neq y']\!]$ be the $0/1$ loss. Let $T_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be a sample i.i.d. drawn from a distribution $p(x, y)$. Assume that $\mathcal{H}$ has a finite VC dimension, $\mathrm{VCdim}(\mathcal{H}) < \infty$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ the inequality

$$\underbrace{R(h, p)}_{\text{true error}} \leq \underbrace{\widehat{R}(T_m, h)}_{\text{training error}} + \underbrace{4\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) \ln\left(\frac{2\,e\,m}{\mathrm{VCdim}(\mathcal{H})}\right) + \ln\left(\frac{4}{\delta}\right)}{m}}}_{\text{complexity term}}$$

holds for all $h \in \mathcal{H}$ simultaneously.

- Compare the VC generalization bound with the bound for a finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_H\}$:

$$R(p, h) \leq \hat{R}(T_m, h) + \sqrt{\frac{\ln 2|\mathcal{H}| + \ln \frac{1}{\delta}}{2m}}$$

- Recommendations for learning:
  1. Use as much training examples m as you can.
  2. Minimize the empirical risk.
  3. Limit the complexity of the hypothesis space $\mathcal{H}$.

Structural Risk Minimization

Algorithm:

1. Construct a nested sequence of hypothesis classes:

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_K$$

2. For each $i \in \{1, \ldots, K\}$, apply ERM:

$$h_i = \arg\min_{h \in \mathcal{H}_i} \hat{R}(T_m, h)$$

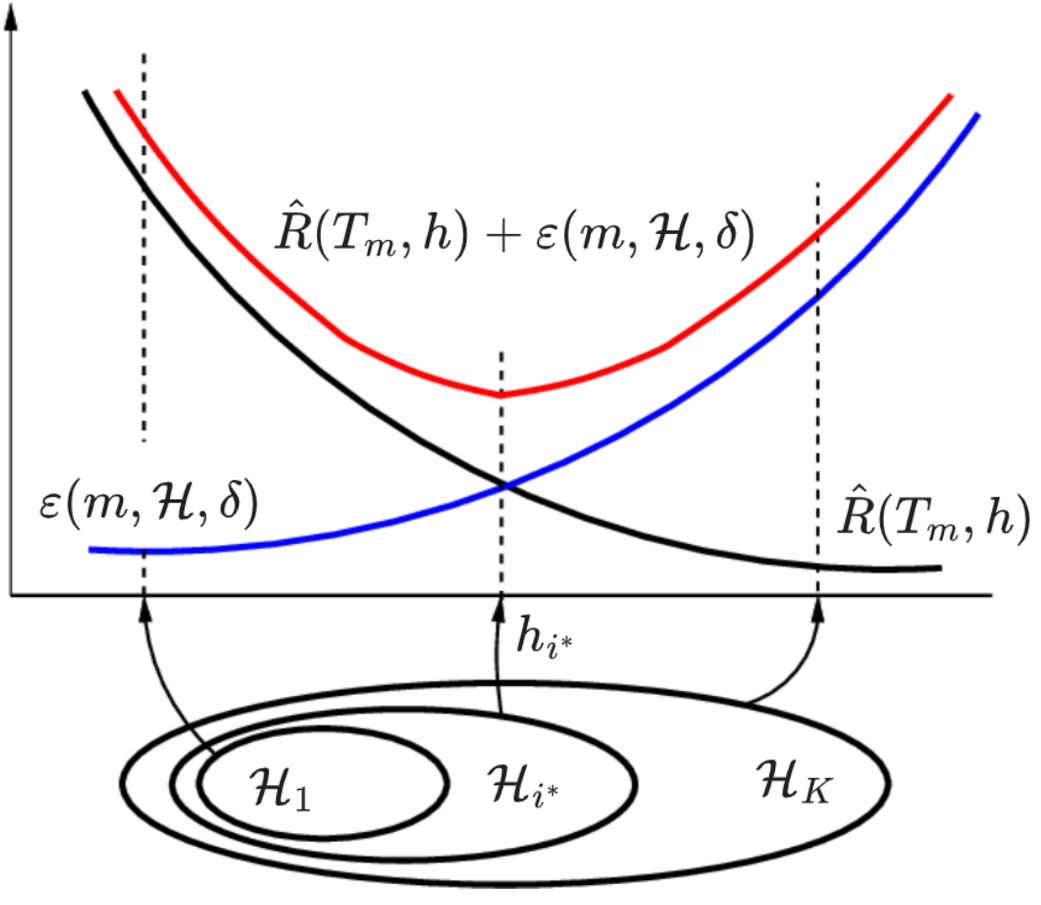
3. Select the best model using the VC generalization bound:

$$i^* = \arg\min_{i=1,\ldots,K} \Big(\hat{R}(T_m, h_i) + \varepsilon(m, \mathcal{H}_i, \delta) \Big)$$

where

$$\varepsilon(m, \mathcal{H}_i, \delta) = 4\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) \, \log\left(\frac{2\,e\,m}{\mathrm{VCdim}(\mathcal{H})}\right) \, + \, \log\left(\frac{4}{\delta}\right)}{m}}$$

4. Output h_{i^*}.



Summary

- **Vapnik–Chervonenkis (VC) dimension**
  - Measures the complexity of a hypothesis class of binary classifiers, $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$.
  - Defined as the largest number of points that can be *shattered* by $\mathcal{H}$.
- **Fundamental Theorem of PAC Learning**
  - Finite VC dimension implies that the Uniform Law of Large Numbers (ULLN) holds.
  - Consequently, Empirical Risk Minimization (ERM) is a PAC learner.
  - *Remark:* This statement is for binary hypothesis classes $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ under the $0/1$ loss.
- **Structural Risk Minimization (SRM)**
  - Extends the ERM principle.
  - Minimizes empirical risk while controlling the complexity of the hypothesis class.