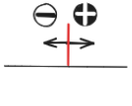
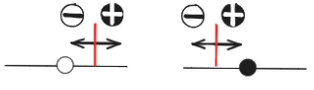
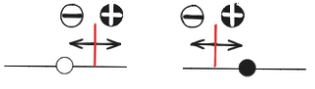
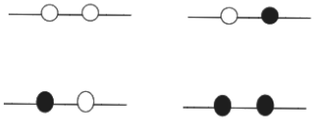
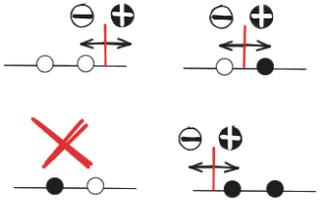
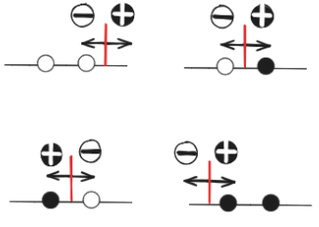
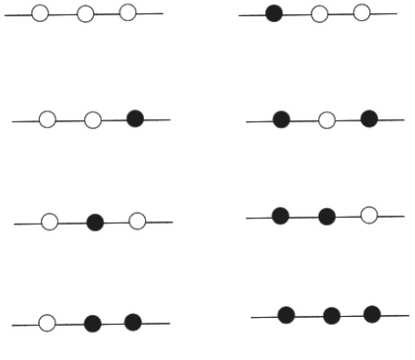
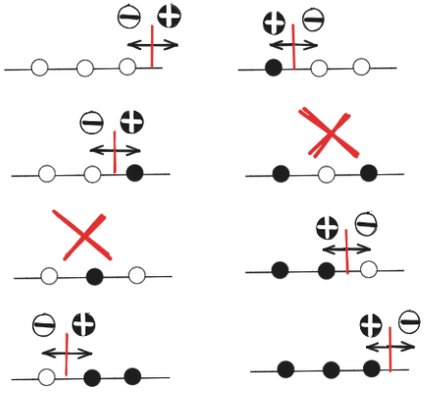
Statistical Machine Learning (BE4M33SSU)
VC dimension

Czech Technical University in Prague
V. Franc

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

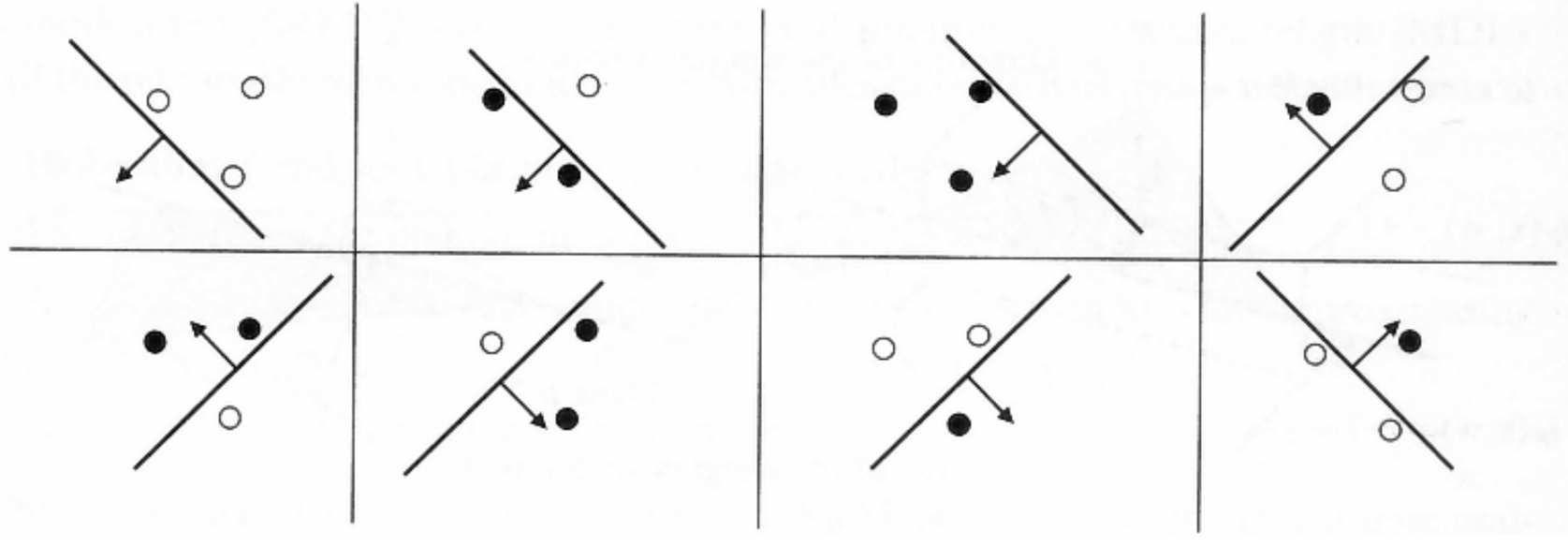
$$h(x_i) = y_i, \quad \forall\, i \in \{1, \ldots, m\}.$$

Definition (VC dimension): Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$. The Vapnik–Chervonenkis dimension of $\mathcal{H}$, denoted $d_{\mathrm{VC}}(\mathcal{H})$, 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 $\mathcal{H} = \{h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b) \mid (\boldsymbol{w}, b) \in \mathbb{R}^{d+1})\}$ is $d_{\mathrm{VC}}(\mathcal{H}) = d + 1$.

Example for $n = 2$-dimensional feature space



Quiz 1: 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 $\mathcal{H}$?

Quiz 2: Let $\mathcal{H} = \{h_1, \ldots, h_K\}$ be a finite hypothesis class. What is the VC dimension of $\mathcal{H}$?

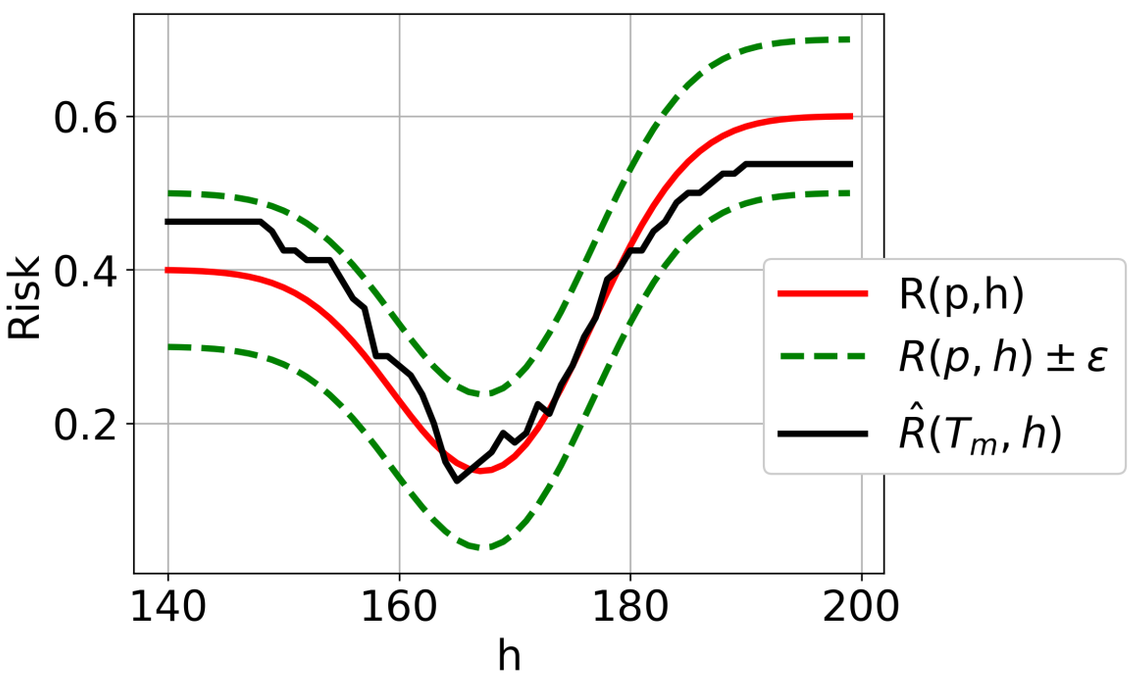
VC Dimension Generalization Bound

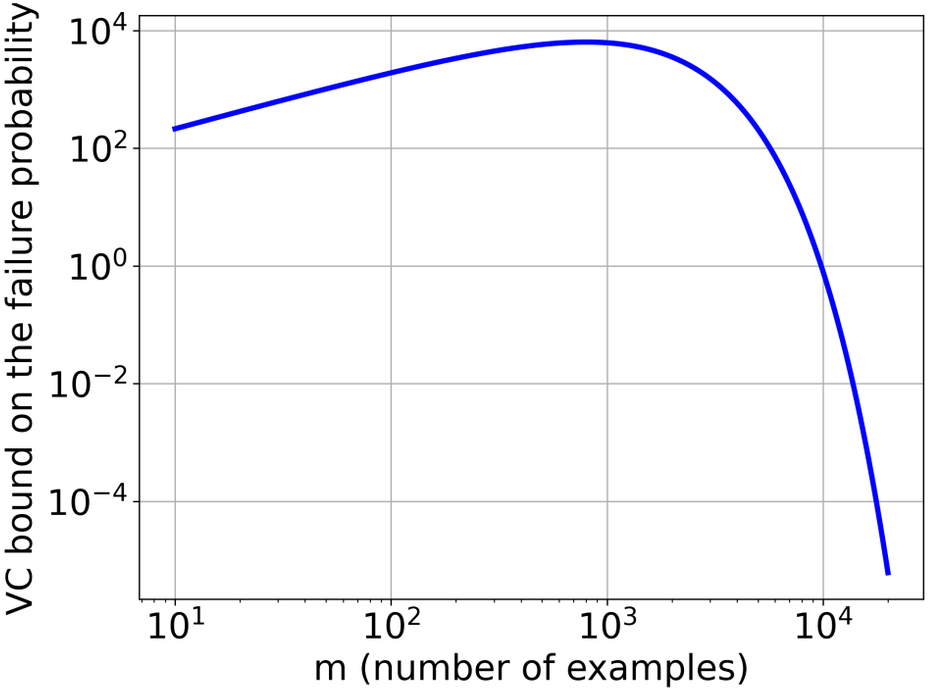
Theorem: Let $\mathcal{H} \subset \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d_{\mathrm{VC}}(\mathcal{H}) < \infty$ and $T_m = \{(x_1, y_1), \ldots, (x_m, y_m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ a training set i.i.d. drawn from a distribution $p(x, y)$. Then for any $\varepsilon > 0$ it holds

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

where $R(p, h) = \mathbb{E}_{(x,y)\sim p}\left[\left[\!\left[y \neq h(x) \right]\!\right]\right]$ is the true error and $\hat{R}(T_m, h) = \frac{1}{m} \sum_{i=1}^{m} \left[\!\left[y_i \neq h(x_i) \right]\!\right]$ is the empirical error.

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





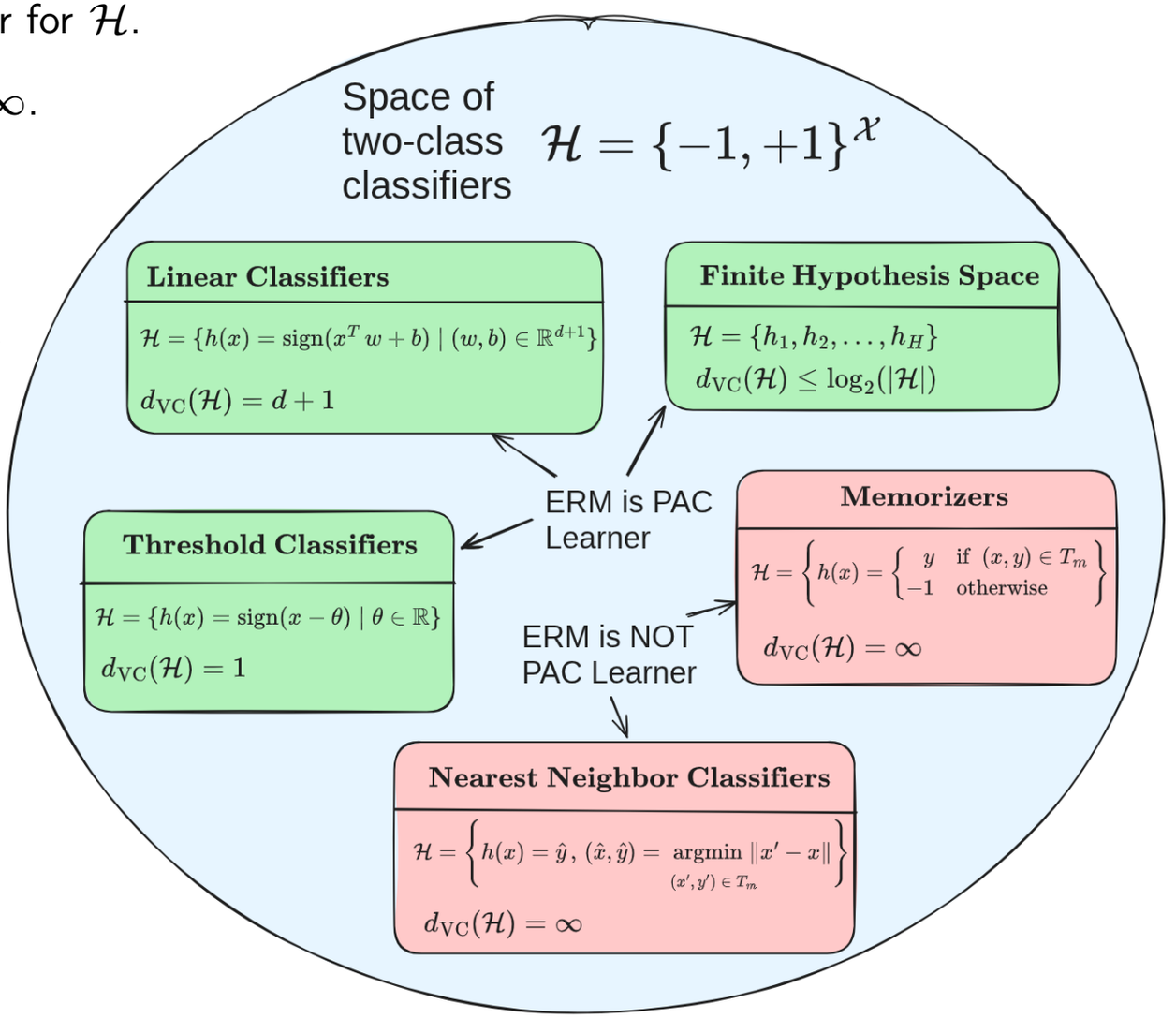
Fundamental Theorem of PAC Learning

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 $\mathcal{H}$.
- ERM algorithm is a successful PAC learner for $\mathcal{H}$.
- $\mathcal{H}$ has finite VC dimension, $d_{\mathrm{VC}}(\mathcal{H}) < \infty$.

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

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



VC Dimension Generalization Bound

Theorem. Let $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ be a hypothesis class of binary classifiers, and let $\ell(y, y') = [\![y \neq y']\!]$ denote the 0–1 loss. Assume that $\mathcal{H}$ has a finite VC dimension, $d_{\mathrm{VC}}(\mathcal{H}) < \infty$. Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over an i.i.d. training sample $T_m = ((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m)$, the following inequality holds for all $h \in \mathcal{H}$ simultaneously:

$$R(h, p) \;\leq\; \underbrace{\widehat{R}(T_m, h)}_{\text{empirical risk}} \;+\; \underbrace{4\sqrt{\frac{d_{\mathrm{VC}}(\mathcal{H})\,\log\!\left(\frac{2\,e\,m}{d_{\mathrm{VC}}(\mathcal{H})}\right) \;+\; \log\!\left(\frac{4}{\delta}\right)}{m}}}_{\text{complexity term}}$$

Practical implications of the VC bound:

- Minimize the empirical risk $\widehat{R}(T_m, h)$ (fit the data well).
- Control the complexity term:
  - Use as many training examples m as possible.
  - Incorporate prior knowledge to restrict the complexity of $\mathcal{H}$ (simpler models $\Rightarrow$ smaller VC dimension).

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:

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

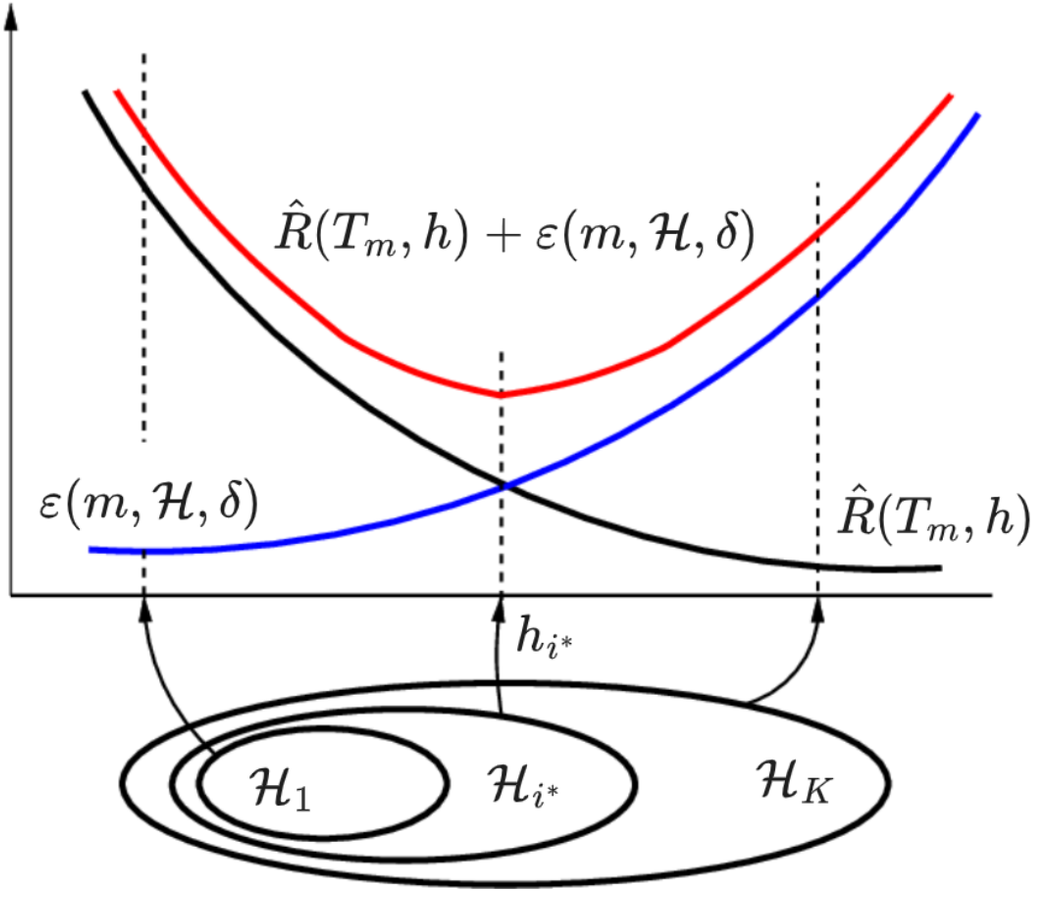
3. Select the best model using the VC generalization bound:

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

where

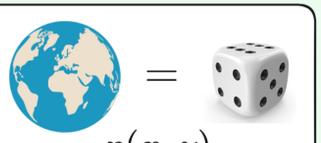
$$\varepsilon(m, \mathcal{H}_i, \delta) = 4\sqrt{\frac{d_{\mathrm{VC}}(\mathcal{H}) \, \log\left(\frac{2\,e\,m}{d_{\mathrm{VC}}(\mathcal{H})}\right) \, + \, \log\left(\frac{4}{\delta}\right)}{m}}$$

4. Output h_{i^*}.



Summary of Key Concepts

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

"Too complex" hypothesis space

E.g. Memorizer

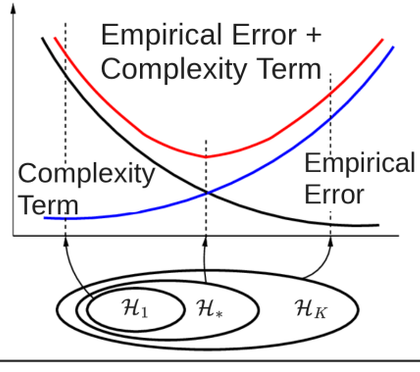
ERM is not PAC learner

Uniform Law of Large Numbers

ULLN holds for $\mathcal{H}$ $\iff$

ERM is PAC learner.

Structured Risk Minimization



PAC learning

Successful PAC learner: finds approximately correct predictor with high probability.

$m \geq m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$:

$\mathbb{P}[\text{estimation error} \leq \varepsilon] \geq 1 - \delta$

Finite hypothesis space

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

ERM is PAC learner

$m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta) = \frac{2}{\varepsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$

VC dimension

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

Fundamental Theorem:

$\text{VCdim}(\mathcal{H}) < \infty$

$\iff$ ULLN holds for $\mathcal{H}$

$\iff$ ERM is PAC learner

Summary of Key Concepts

- **Vapnik–Chervonenkis (VC) dimension:** measures the complexity of a hypothesis class containing two-class classifiers $\mathcal{H} \subset \{-1, +1\}^{\mathcal{X}}$.
  - Defined as the largest number of points that can be *shattered* by $\mathcal{H}$.
- **Fundamental Theorem of PAC Learning:** A finite VC dimension implies that the Uniform Law of Large Numbers (ULLN) holds, and that Empirical Risk Minimization (ERM) is a PAC learner.
- **Structural Risk Minimization (SRM):** Minimize the empirical risk while simultaneously controlling the complexity of the hypothesis class.