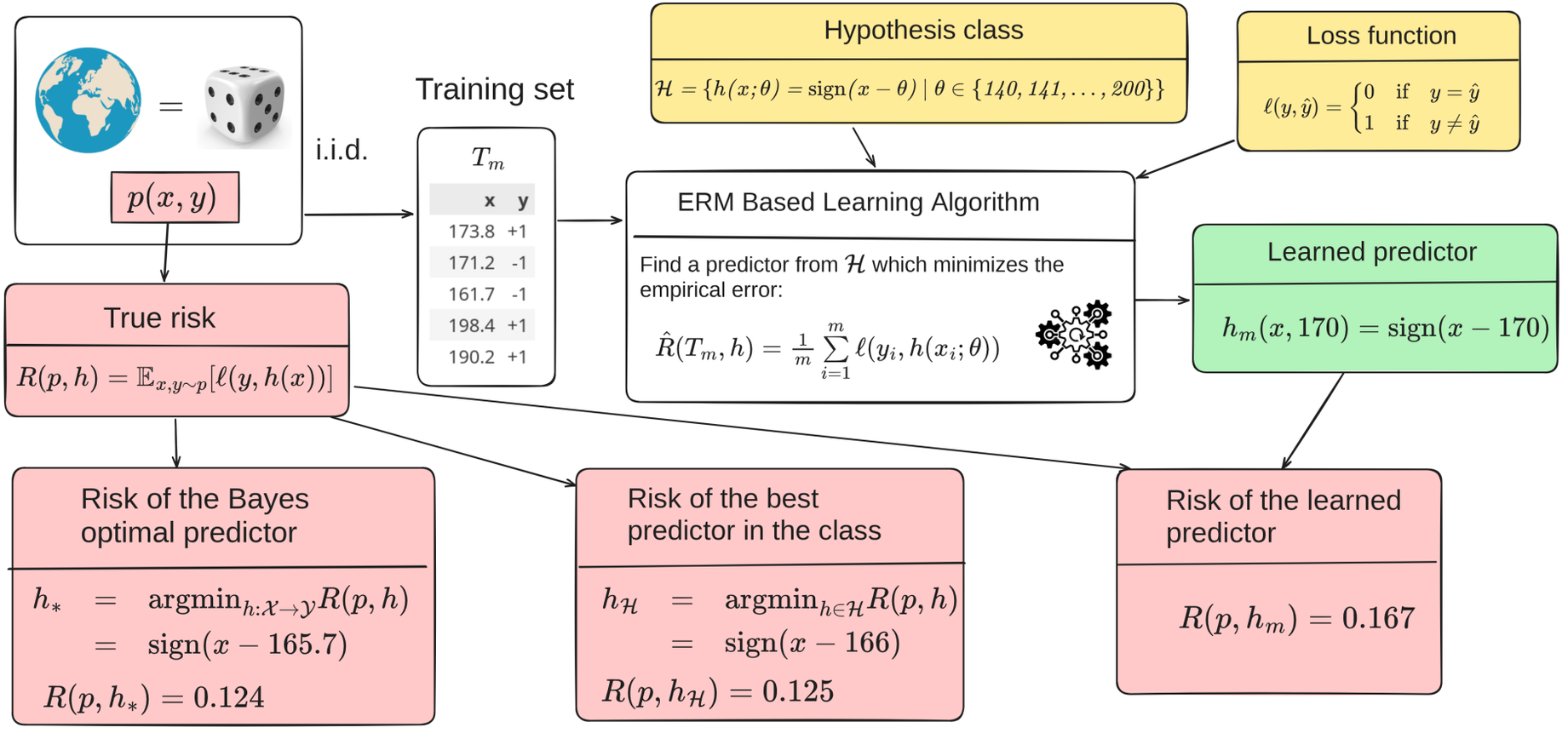
Machine Learning Fundamentals - LS2026
Probably Approximately Correct Learning

Czech Technical University in Prague
V. Franc

Empirical Risk Minimization



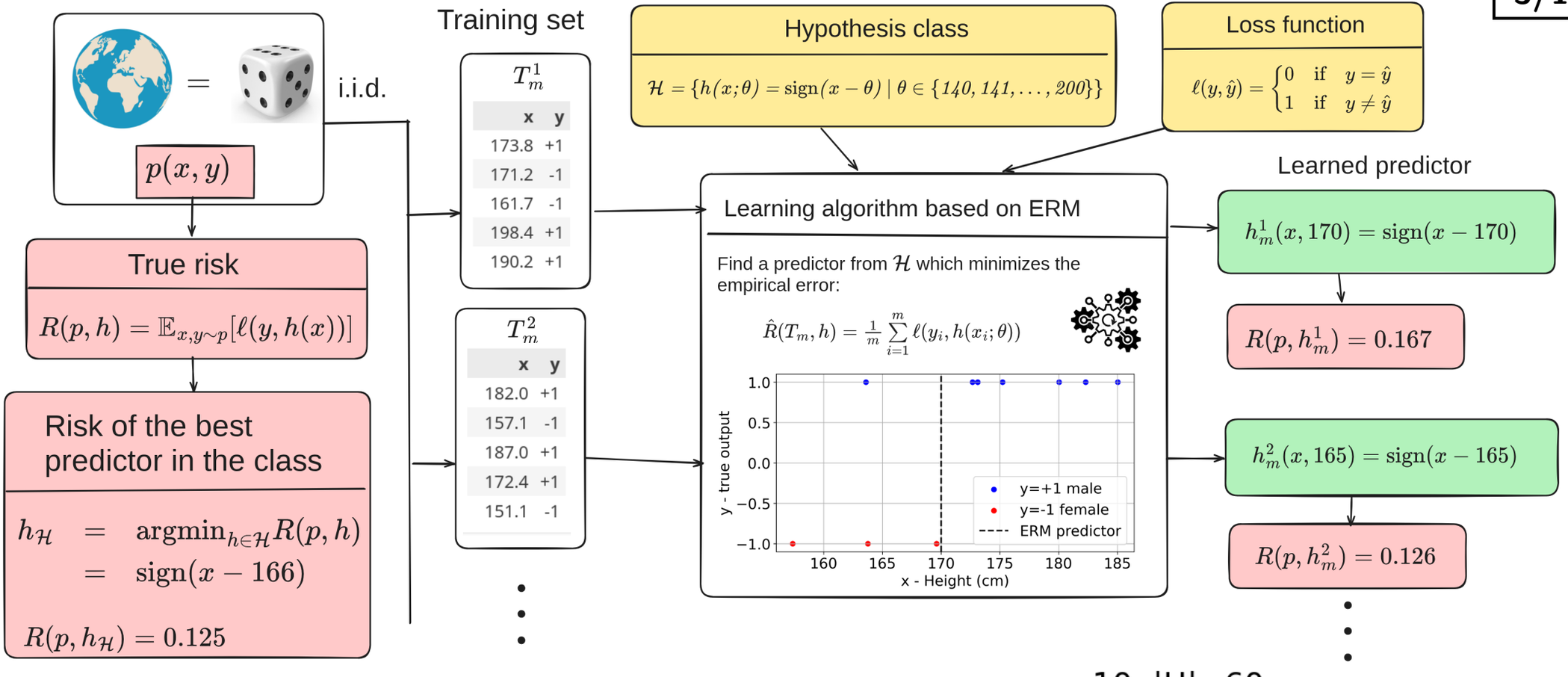
Error decomposition:

$$\underbrace{R(p, h_m)}_{\text{learned predictor risk}} = \underbrace{\Big(R(p, h_m) - R(p, h_{\mathcal{H}})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(p, h_{\mathcal{H}}) - R(p, h_*)\Big)}_{\text{approximation error}} + \underbrace{R(p, h_*)}_{\text{Bayes risk}}$$

Questions to answer in this lecture:

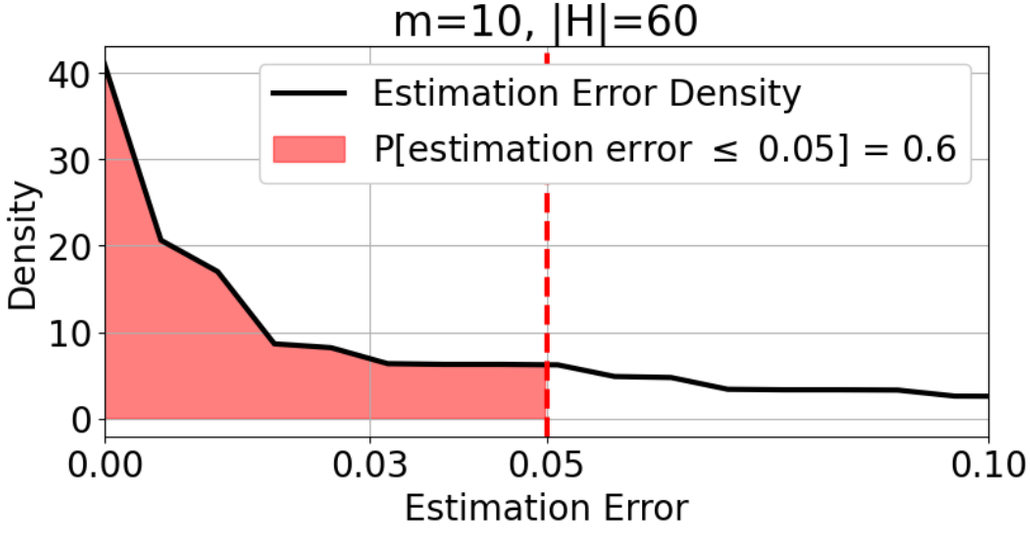
- For what choises of $\mathcal{H}$ we can find a good approximation of the best predictor in the class?
- How many examples are needed to guarnatee that such an approximation is likely to be found?

Probably Approximately Correct



Given $\varepsilon > 0$, learned predictor h_m is approximately correct provided:

$$\underbrace{\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}})}_{\text{estimation error}} \leq \varepsilon}_{\text{approximately correct}}$$



Probably Approximately Correct

- **Definition:** A hypothesis class $\mathcal{H}$ is PAC learnable if there exists a function $m_{\text{pac}}^{\mathcal{H}}\colon (0,1) \times (0,1) \to \mathbb{N}$ and an algorithm $A\colon \cup_{i=1}^{m} (\mathcal{X} \times \mathcal{Y})^i \to \mathcal{H}$ such that for every $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and every distribution $p(x,y)$ whenever the algorithm is run on $m \geq m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples $T_m = ((x_i, y_i) \in \mathcal{X} \times \mathcal{Y} \mid i = 1, \ldots, m)$ drawn from $p(x,y)$, it returns $h_m = A(T_m)$ satisfying, with probability at least $1 - \delta$,

$$\boxed{R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon}$$

- **Key Concepts:**
  - **PAC learnable:** There exists a learning algorithm (a *successful PAC learner*) that, given enough examples, finds with high probability a predictor close to the best one in $\mathcal{H}$.
  - **Approximately correct:** The learned predictor has risk at most ε larger than the risk of the best predictor in $\mathcal{H}$.
  - **Probably:** The probability that the algorithm fails to achieve this accuracy is at most δ.
  - **Sample complexity** $m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$**:** The number of examples sufficient to guarantee accuracy ε with confidence $1 - \delta$.
  - **Distribution independence:** The guarantee holds for every data distribution $p(x,y)$.

Uniform Law of Large Numbers

- If **ULLN** applies for a hypothesis space $\mathcal{H}$ then:

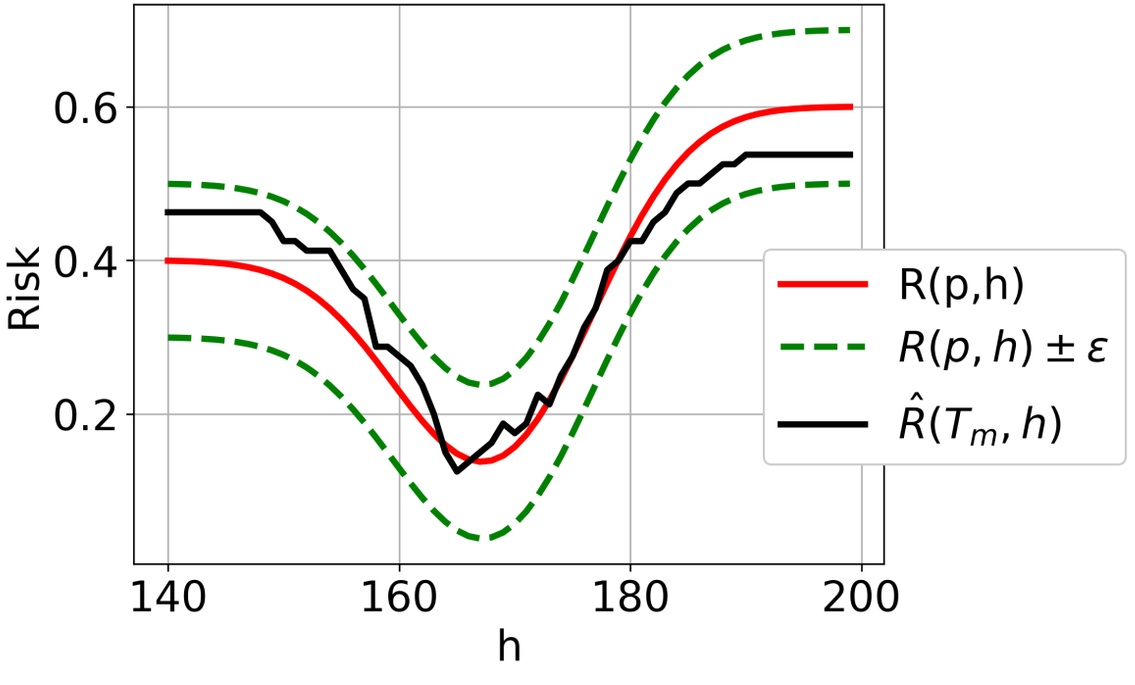
$$\forall \varepsilon > 0: \qquad \lim_{m \to \infty} \mathbb{P}\Big(\underbrace{\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big|}_{\text{maximal generalization gap}} \geq \varepsilon \Big) = 0$$

emprical risk fails to approximate true risk

ULLN applies only for some $\mathcal{H}$.

- **ULLN applies for the finite hypothesis space** $\mathcal{H} = \{h_1, h_2, \ldots, h_H\}$:

$$\forall \varepsilon > 0: \qquad \mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big| \geq \varepsilon \Big) \leq 2\,|\mathcal{H}|\, e^{-2\,m\,\varepsilon^2}$$



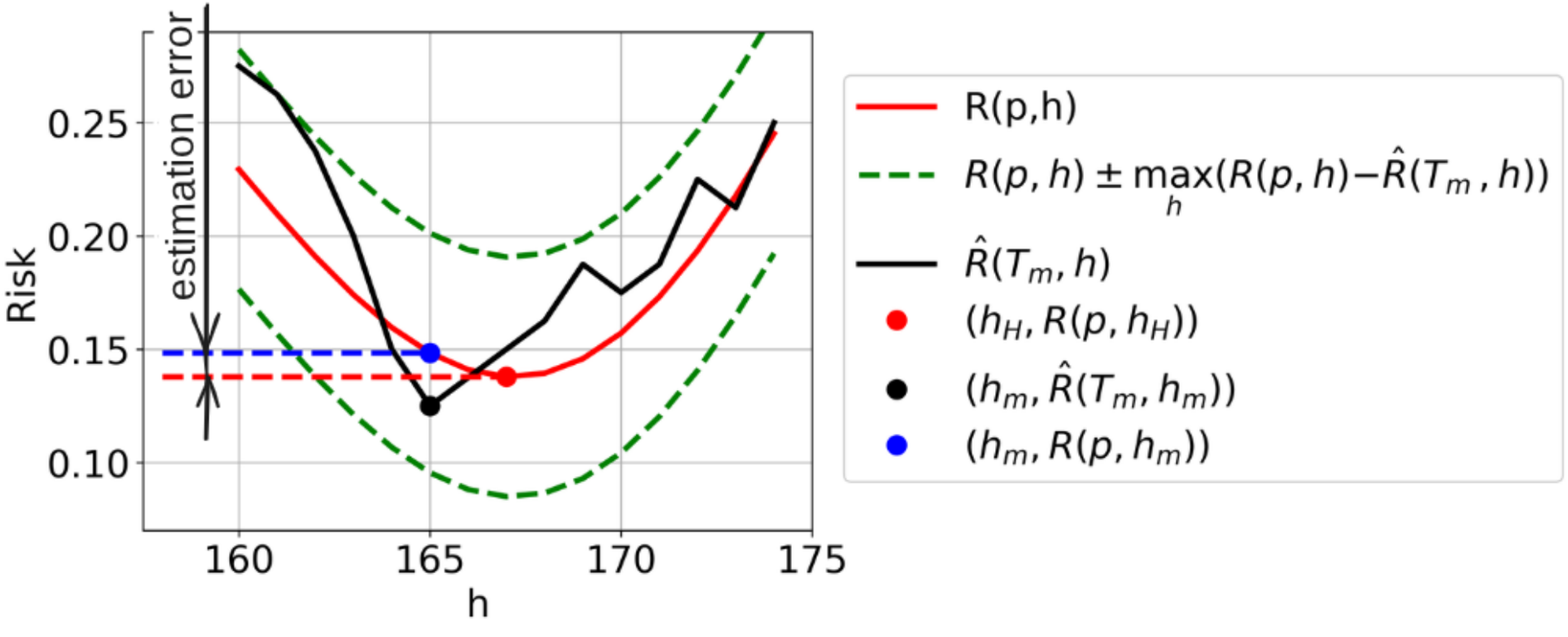
Example:

$\mathcal{H} = \{h(x;\theta) = \text{sign}(x - \theta) \mid \theta \in \{140, 141, \ldots, 200\}\}$, $\varepsilon = 0.1$

Bound on Estimation Error

- **Bound on estimation error for ERM algorithm** $h_m = \arg\min_{h\in\mathcal{H}} \hat{R}(T_m, h)$:

$$\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}})}_{\textit{estimation error}} = \Big(R(p, h_m) - \hat{R}(T_m, h_m)\Big) + \Big(\hat{R}(T_m, h_m) - R(p, h_{\mathcal{H}})\Big)$$
$$\leq \Big(R(p, h_m) - \hat{R}(T_m, h_m)\Big) + \Big(\hat{R}(T_m, h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big)$$
$$\leq 2 \underbrace{\max_{h\in\mathcal{H}} \big|R(p, h) - \hat{R}(T_m, h)\big|}_{\textit{maximal generalization gap}}$$



ULLN implies $\mathcal{H}$ is PAC learnable with ERM algorithm

- **ULLN** for finite hypothesis class $\mathcal{H} = \{h_1, h_2, \ldots, h_H\}$:

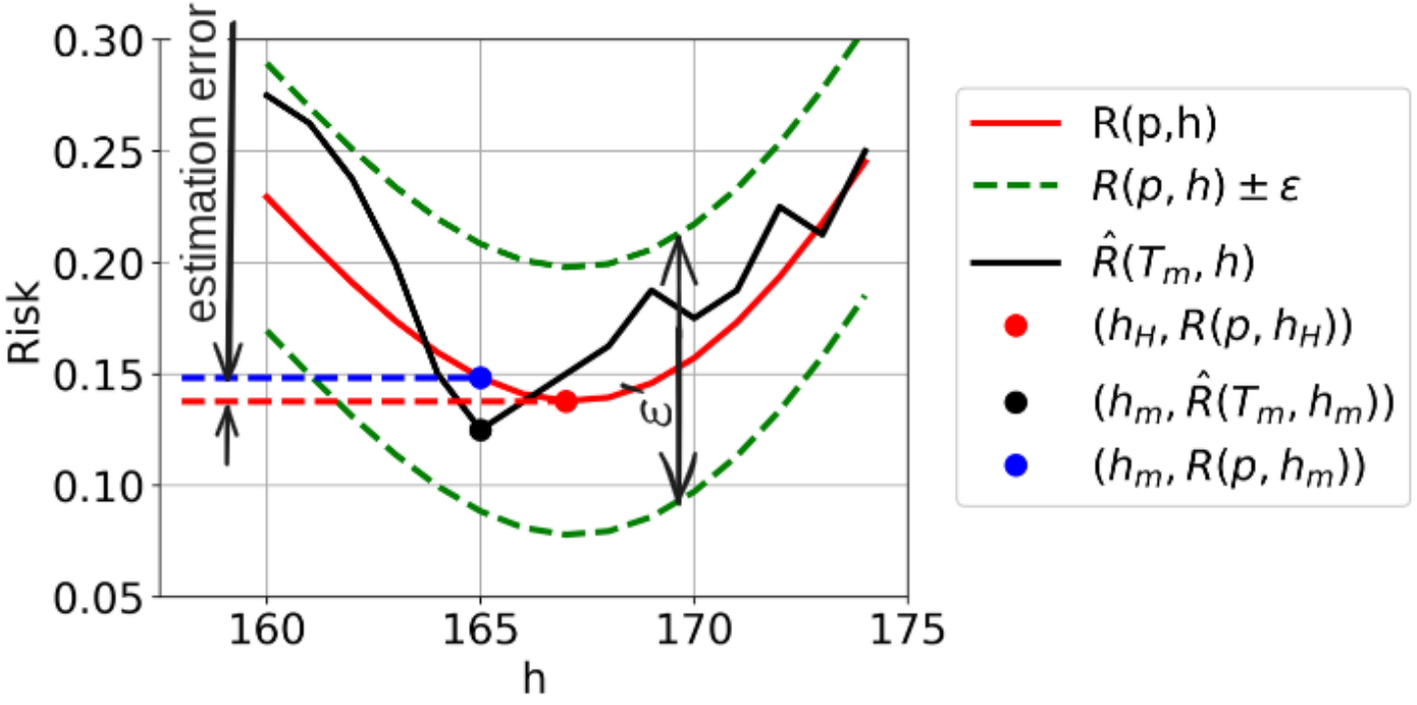
$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big| \geq \varepsilon \Big) \leq 2\,|\mathcal{H}|\, e^{-2\,m\,\varepsilon^2}$$

- **Bound on estimation error of** $h_m = \arg\min_{h \in \mathcal{H}} \hat{R}(T_m, h)$:

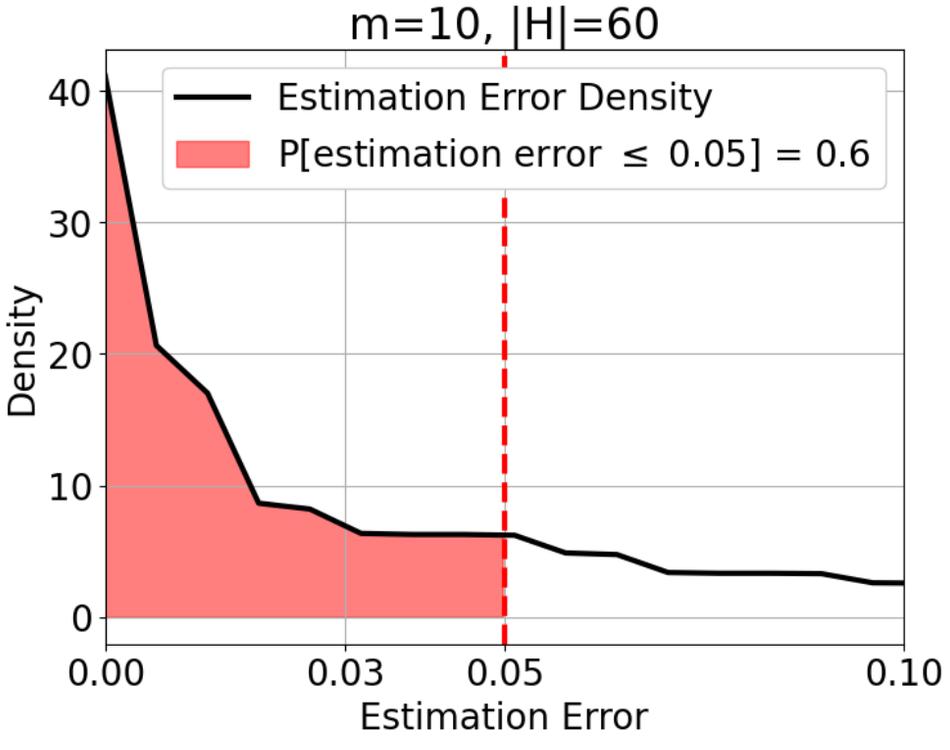
$$R(p, h_m) - R(p, h_{\mathcal{H}}) \leq 2 \max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big|$$

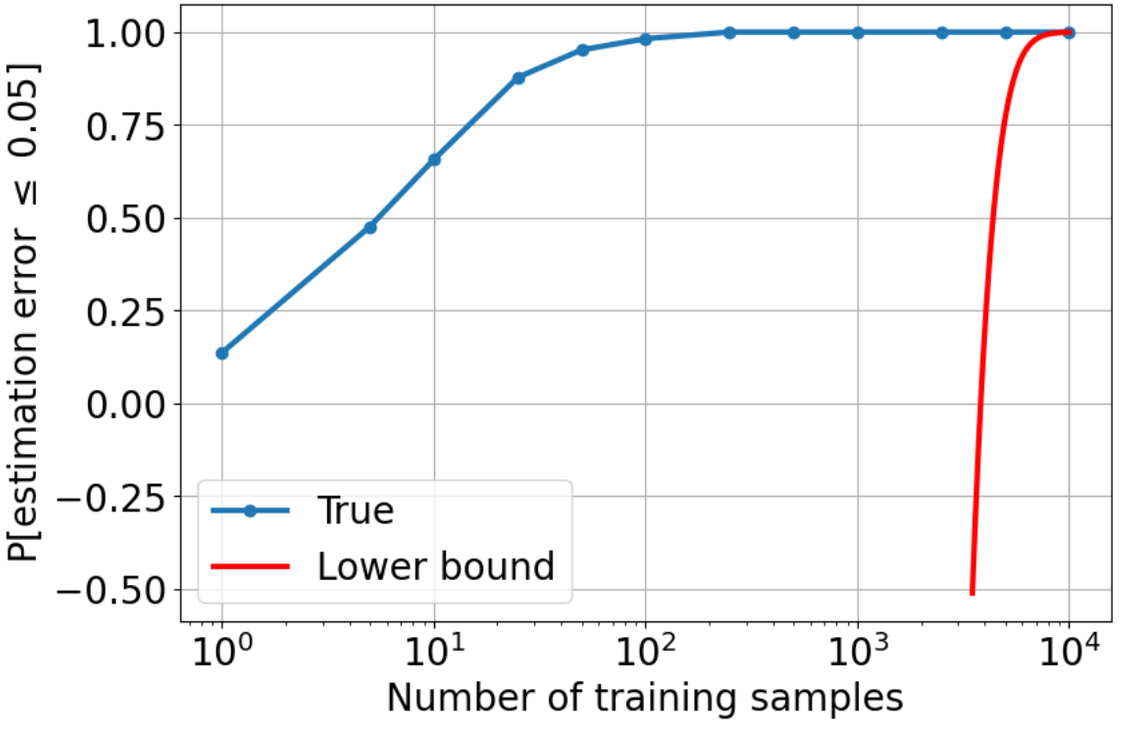
- **ULLN applies for** $\mathcal{H}$ **+ Bound on estimation error** $= \mathcal{H}$ **is PAC learnable with ERM**:

$$\mathbb{P}\Big(R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon' \Big) \geq 1 - 2\,|\mathcal{H}|\, e^{-\frac{1}{2}\,m\,(\varepsilon')^2}$$



Probably Approximately Correct





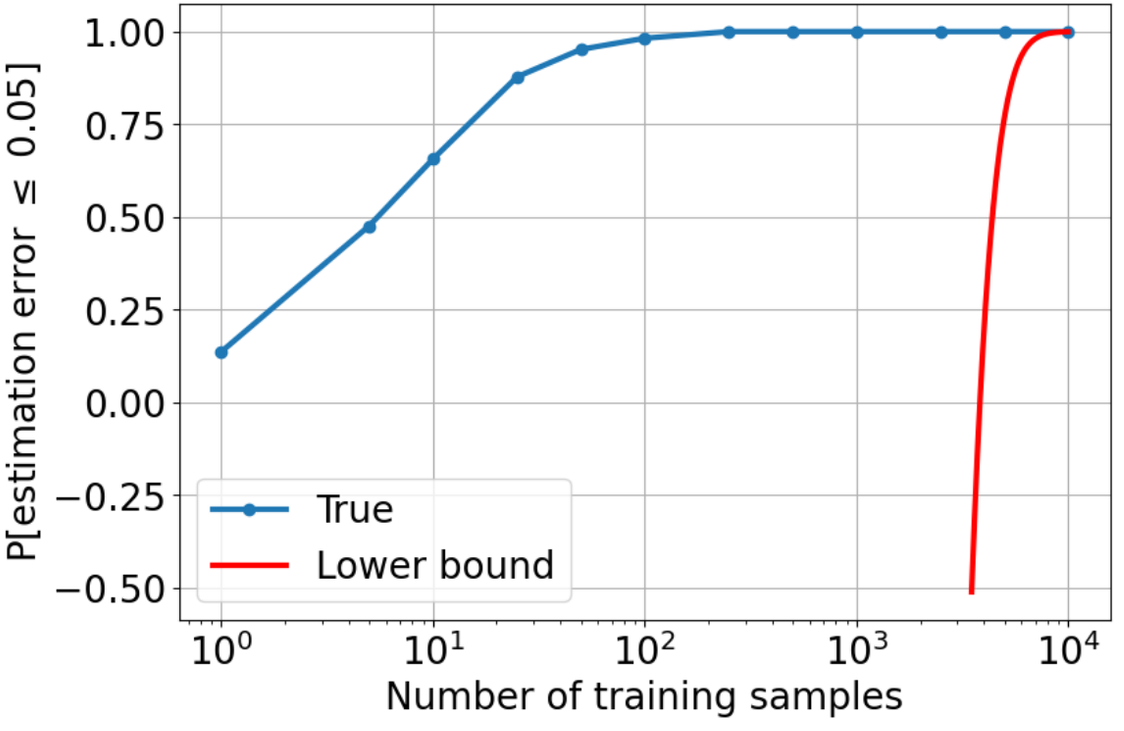
- ERM algorithm: $h_m = A(T_m) = \arg\min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h(x_i)]\!] \right]$
- Distribution-free lower bound valid for a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_H\}$ and $0/1$-loss:

$$\mathbb{P}_{T_m \sim p^m} \Big(\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon}_{\text{approximately correct}} \Big) \geq \underbrace{1 - 2\,|\mathcal{H}|\, e^{-\frac{1}{2} m\, \varepsilon^2}}_{\text{lower bound increasing with } m}$$

Probably Approximately Correct

Setup:

- $h_m = \arg\min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h(x_i)]\!] \right]$
- Finite hypothesis class: $\mathcal{H} = \{h_1, h_2, \ldots, h_H\}$
- 0/1-loss: $\ell(y, y') = [\![y \neq y']\!]$



- Distribution-free lower bound:

$$\mathbb{P}_{T_m \sim p^m} \Big[\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon}_{\text{approximately correct}} \Big] \geq 1 - 2\,|\mathcal{H}|\, e^{-\frac{1}{2} m \varepsilon^2} = \underbrace{1 - \delta}_{\text{probably}}$$

- Given $\varepsilon > 0$, probability of failure $\delta > 0$, we can compute the sample complexity:

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

ERM on Finite Hypothesis Space is a Successful PAC learner

- **Theorem.** Let $\mathcal{H} = \{h_i : \mathcal{X} \to \mathcal{Y} \mid i \in \{1, \ldots, H\}\}$ be a finite hypothesis class. Then, $\mathcal{H}$ is PAC learnable with the Empirical Risk Minimization (ERM) algorithm

$$h_m = \arg\min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h(x_i)]\!] \right]$$

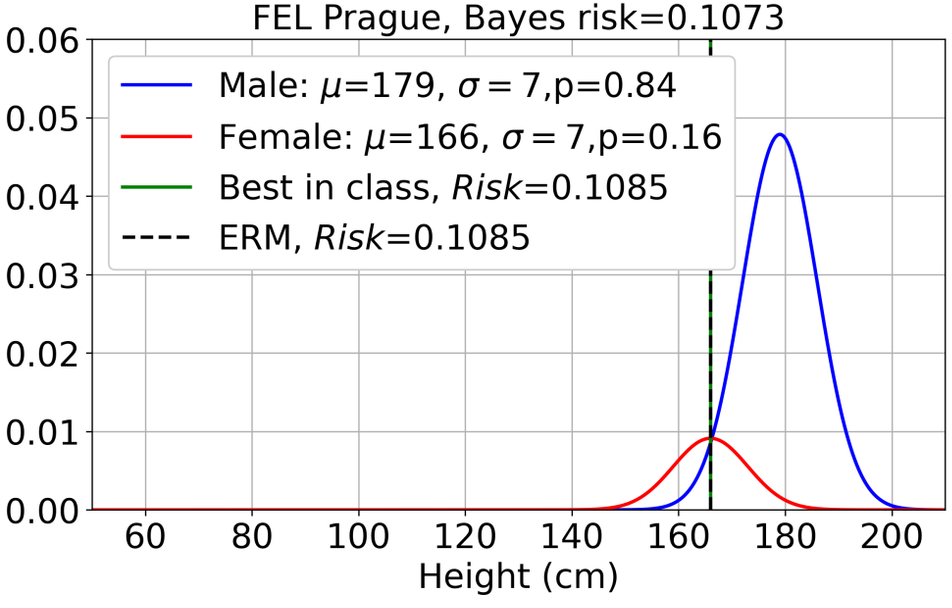
with sample complexity

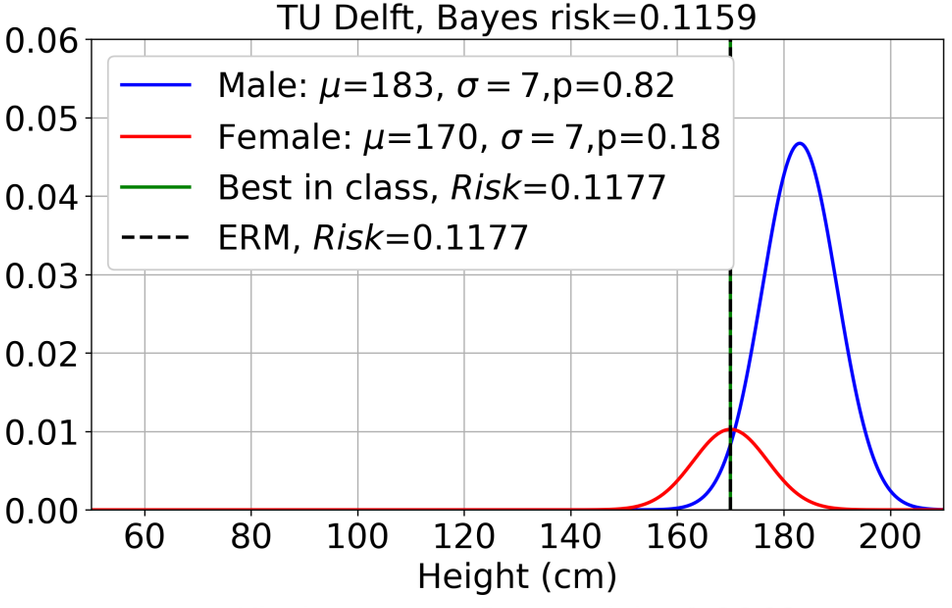
$$\boxed{m^{\mathcal{H}}_{\text{PAC}}(\varepsilon, \delta) = \frac{2}{\varepsilon^2} \ln\left(\frac{2\,|\mathcal{H}|}{\delta}\right)}$$

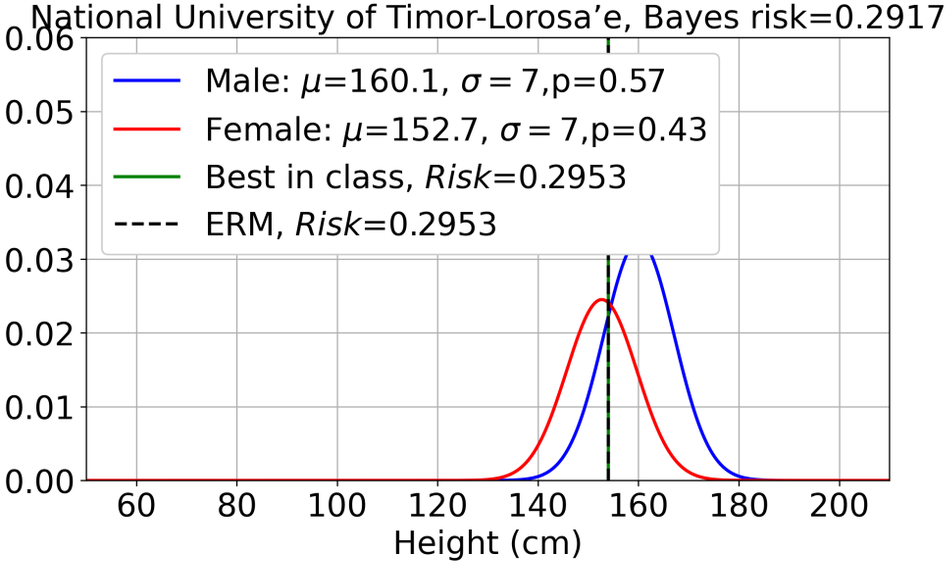
Probably Approximately Correct learning

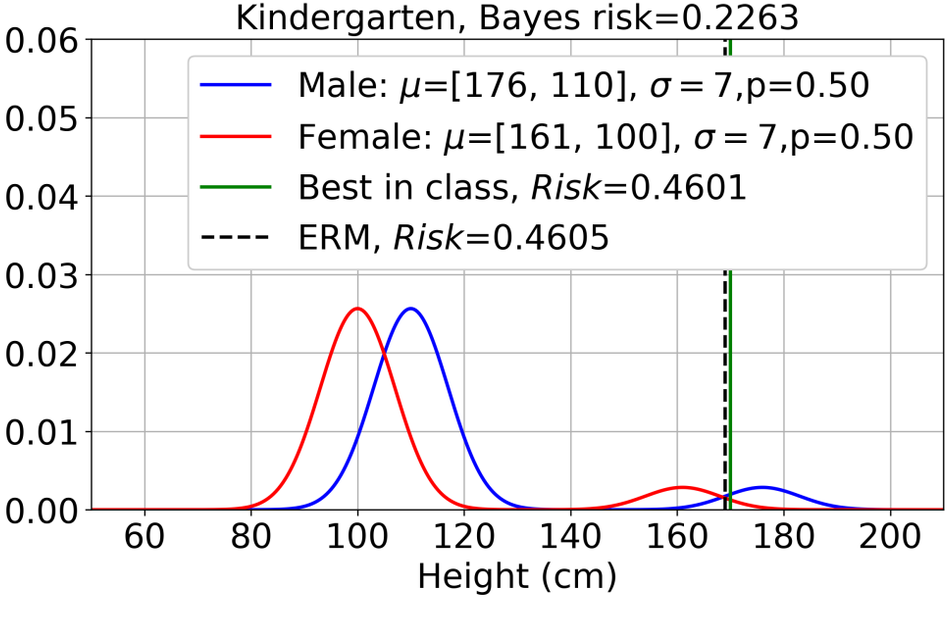
Example: $\varepsilon = 0.05$, $\delta = 0.1$, $\mathcal{H} = \{h(x;\theta) = \text{sign}(x - \theta) \mid \theta \in \{140, 141, \ldots, 200\}\}$

The sample complexity: $m^{\mathcal{H}}_{\text{pac}}(\varepsilon, \delta) = \frac{2}{\varepsilon^2} \ln\left(\frac{2\,|\mathcal{H}|}{\delta}\right) = \frac{2}{0.05^2} \ln\left(\frac{2 \cdot 60}{0.1}\right) \approx 5,673$









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

Hypotheisis class represents our prior knowledge

"Too complex" hypothesis class

E.g. Memorizer

$\mathcal{H}$ is not PAC learner

$\ldots$

Finite hypothesis class

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

ULLN holds for $\mathcal{H}$

Summary

- **Probably Approximately Correct (PAC) Learning**
  - A hypothesis class $\mathcal{H}$ is **PAC learnable** if there exists a learning algorithm that, given enough i.i.d. training examples, returns with probability at least $1 - \delta$ a predictor whose risk is at most ε larger than the risk of the best predictor in $\mathcal{H}$.
  - The guarantee is **distribution-free**: it holds for every data distribution $p(x, y)$.
  - The function $m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$ is the **sample complexity**, i.e., the number of examples sufficient to guarantee accuracy ε with confidence $1 - \delta$.
- **Uniform Law of Large Numbers is Sufficient for PAC Learnability**
  - If ULLN holds for the hypothesis class $\mathcal{H}$, then $\mathcal{H}$ is PAC learnable and ERM is a successful PAC learner.
  - Every finite hypothesis class satisfies ULLN.
  - Therefore, every finite hypothesis class is **PAC learnable**, and ERM is a successful PAC learner for it.
  - We derived an explicit sample-complexity bound for finite hypothesis classes depending on $|\mathcal{H}|$.