Statistical Machine Learning (BE4M33SSU)
Lecture 3: Probably Approximately Correct Learning

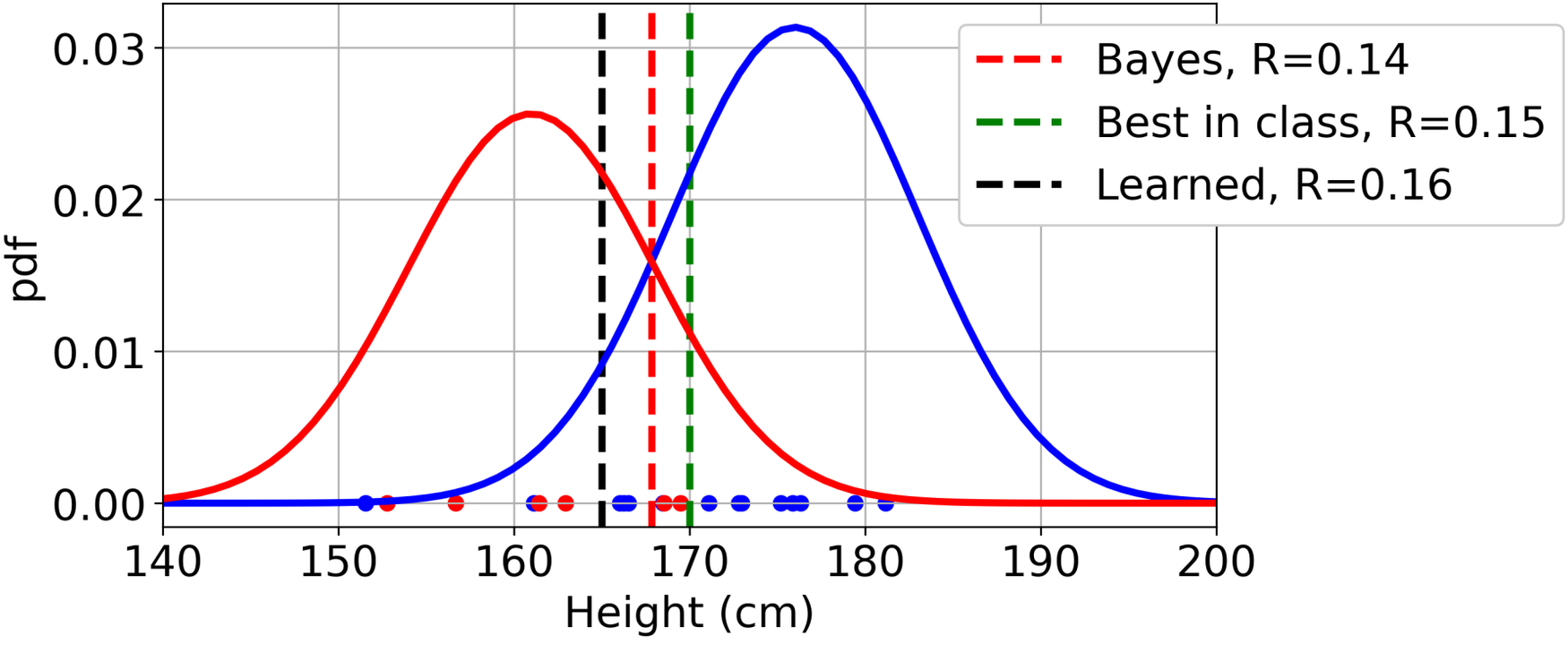
Czech Technical University in Prague
V. Franc

Error decomposition

Errors:

1. Best (Bayes) attainable risk $R(p, h_*)$, where $h_*(x) = \arg\min_{h \in \mathcal{Y}^{\mathcal{X}}} R(p, h)$
2. Best risk in the class $R(p, h_{\mathcal{H}})$, where $h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(p, h)$
3. Risk of the learned predictor $R(p, h_m)$, where $h_m = A(T_m)$

$$\mathcal{H} = \{h(x, \theta) = \text{sign}(x - \theta) \mid \theta \in \{140, 145, \ldots, 200\}\}$$



Error decomposition

Errors:

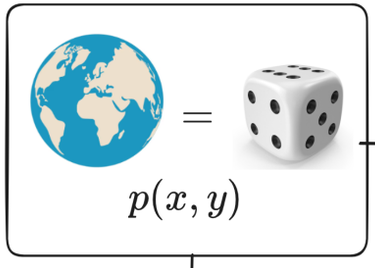
1. Best (Bayes) attainable risk $R(p, h_*)$, where $h_*(x) = \arg\min_{h \in \mathcal{Y}^{\mathcal{X}}} R(p, h)$
2. Best risk in the class $R(p, h_{\mathcal{H}})$, where $h_{\mathcal{H}} = \arg\min_{h \in \mathcal{H}} R(p, h)$
3. Risk of the learned predictor $R(p, h_m)$, where $h_m = A(T_m)$

Error decomposition:

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

- The approximation error: depends on $\mathcal{H}$ chosen prior to learning.
- The estimation error: depends on $\mathcal{H}$, training data T_m and the algorithm A.
- Best (Bayes) attainable risk: irreducible error.

Probably Approximately Correct



$p(x, y)$

i.i.d.

Training set

T_m^1

x	y
173.8	+1
171.2	-1
161.7	-1
198.4	+1
190.2	+1

T_m^2

x	y
182.0	+1
157.1	-1
187.0	+1
172.4	+1
151.1	-1

Hypothesis class

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

Loss function

$\ell(y, \hat{y}) = \begin{cases} 0 & \text{if } y = \hat{y} \\ 1 & \text{if } y \neq \hat{y} \end{cases}$

True risk

$R(p, h) = \mathbb{E}_{x,y\sim p}[\ell(y, h(x))]$

Best in class predictor

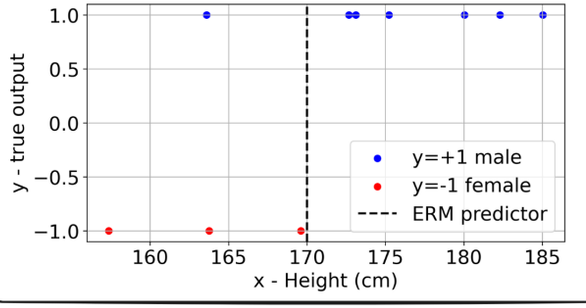
$h_{\mathcal{H}} = \text{argmin}_{h\in\mathcal{H}} R(p, h) = \text{sign}(x - 166)$

$R(p, h_{\mathcal{H}}) = 0.125$

Learning algorithm based on ERM

Find a predictor from $\mathcal{H}$ which minimizes the empirical error:

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



Learned predictor

$h_m^1(x, 170) = \text{sign}(x - 170)$

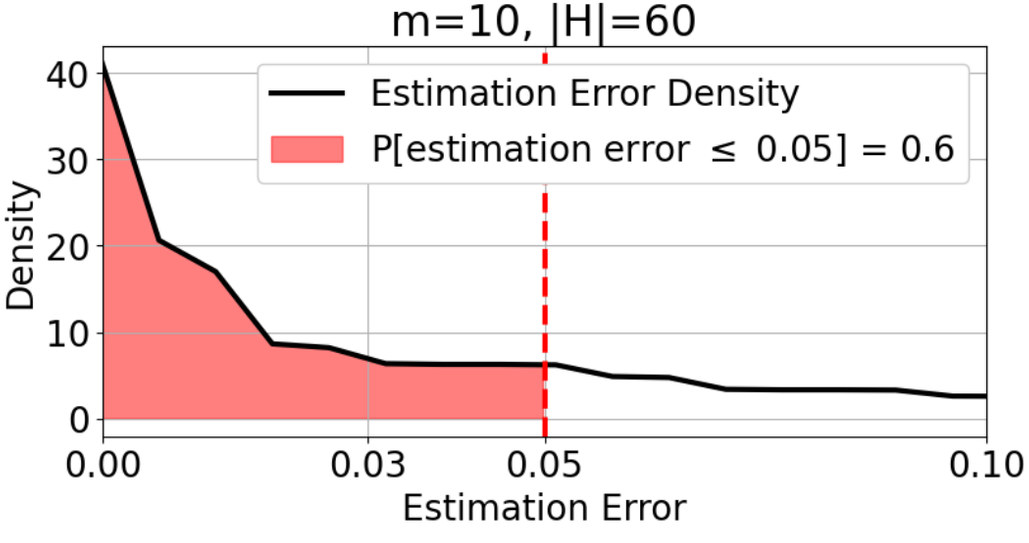
$R(p, h_m^1) = 0.167$

$h_m^2(x, 165) = \text{sign}(x - 165)$

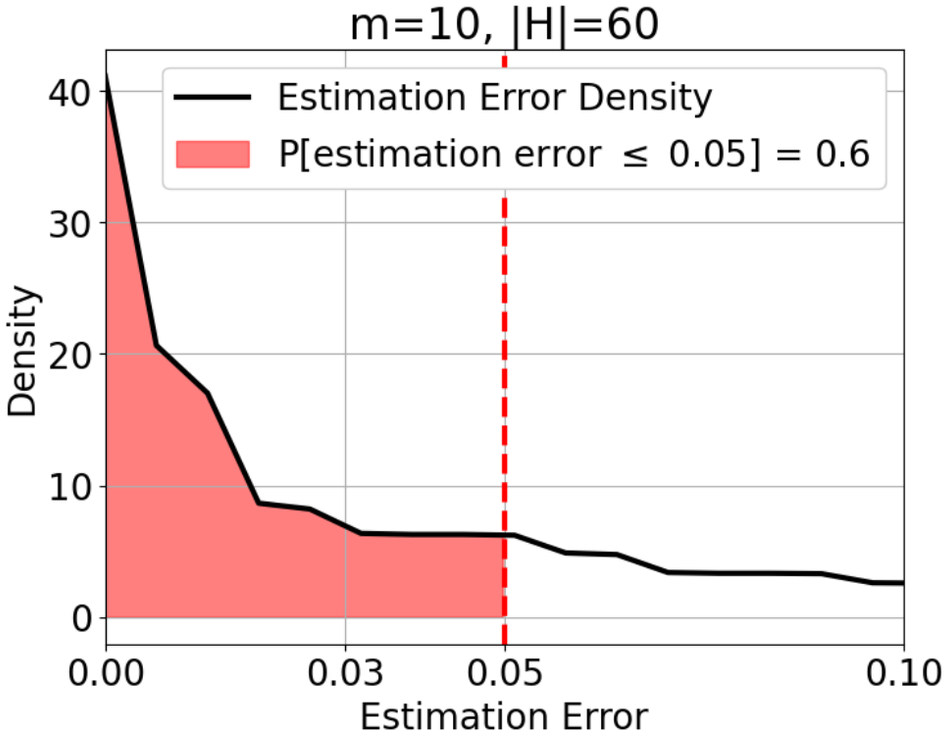
$R(p, h_m^2) = 0.126$

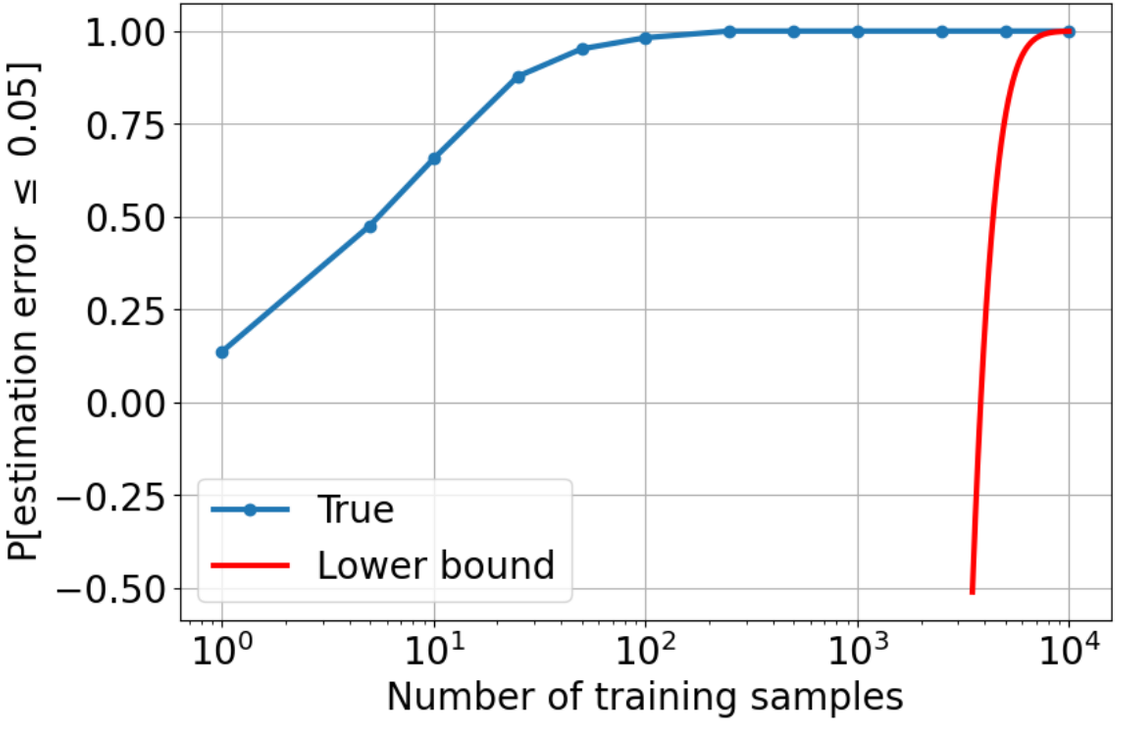
Given $\varepsilon > 0$, learned predictor h_m is approximately correct provided:

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



Probably Approximately Correct





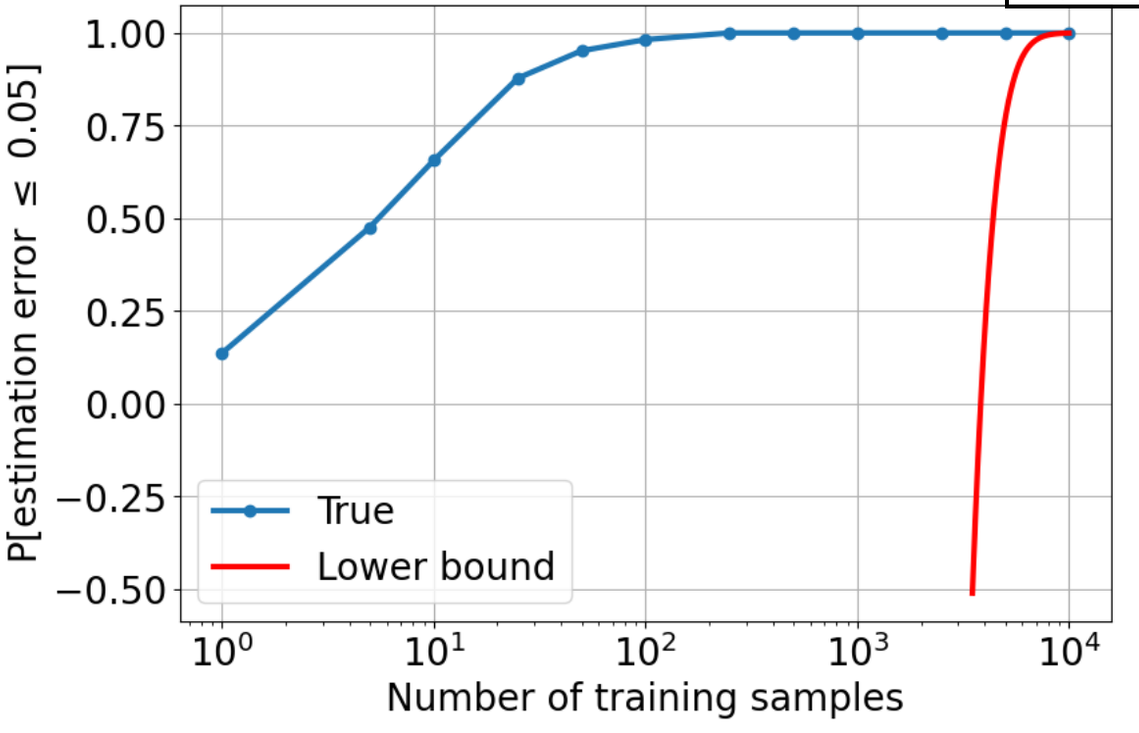
- ERM algorithm: $h_m = A(T_m) = \arg\min_{h \in \mathcal{H}} \left[\frac{1}{m} \sum_{i=1}^{m} [\![y_i \neq h(x_i)]\!] \right]$
- We derive a lower bound valid for any distribution $p(x, y)$ and finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_H\}$:

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

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:
  $\mathcal{H} = \{ h^i \colon \mathcal{X} \to \mathcal{Y} \mid i \in \{1, \ldots, H\} \}$



- 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:

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

Probably Approximately Correct learning

- **Successful PAC learning algorithm:** An algorithm can learn a hypothesis that is likely ("probably") to be approximately correct, given a sufficient number of training examples.

- **Definition:** Algorithm is a successful PAC learner for hypothesis class $\mathcal{H}$ if there exists a function $m_{\text{pac}}^{\mathcal{H}}\colon (0,1) \times (0,1) \to \mathbb{N}$ such that: For every $\varepsilon \in (0,1)$, $\delta \in (0,1)$, and every distribution $p(x,y)$, when running the algorithm on $m \geq m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$ examples T_m i.i.d. drawn from $p(x,y)$, then the algorithm returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}_{T_m \sim p^m}\Big(\underbrace{R(p,h_m) - R(p,h_{\mathcal{H}}) \leq \varepsilon}_{\textit{approximately correct}} \Big) \geq \underbrace{1-\delta}_{\textit{probably}}$$

- **Key Concepts:**
  - **Approximately correct:** The learned predictor's risk is at most ε greater than the risk of the best possible predictor in the class $\mathcal{H}$.
  - **Probably:** The probability of the algorithm failing to produce the approximately correct predictor is at most δ.
  - **Sample complexity** $m_{\text{pac}}^{\mathcal{H}}(\varepsilon, \delta)$**:** The minimum number of examples required to guaranteed the desired accuracy ε and the confidence $1-\delta$.
  - **Distribution independence:** The guarantees hold for any data distribution $p(x,y)$.

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 the Empirical Risk Minimization (ERM) algorithm

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

  is a successful PAC learner with sample complexity

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

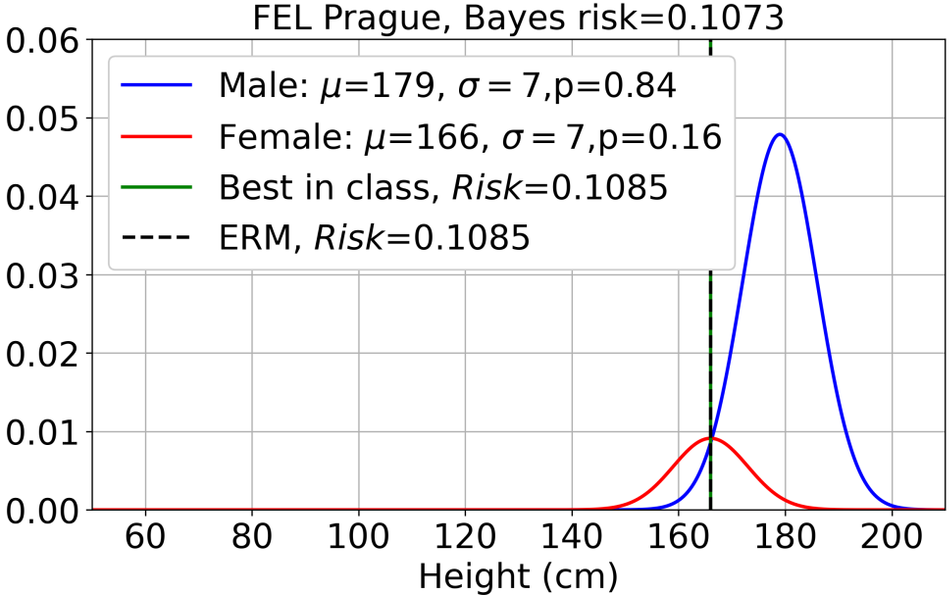
- The theorem is a consequence of the bound we introduced earlier (however, have not yet proved):

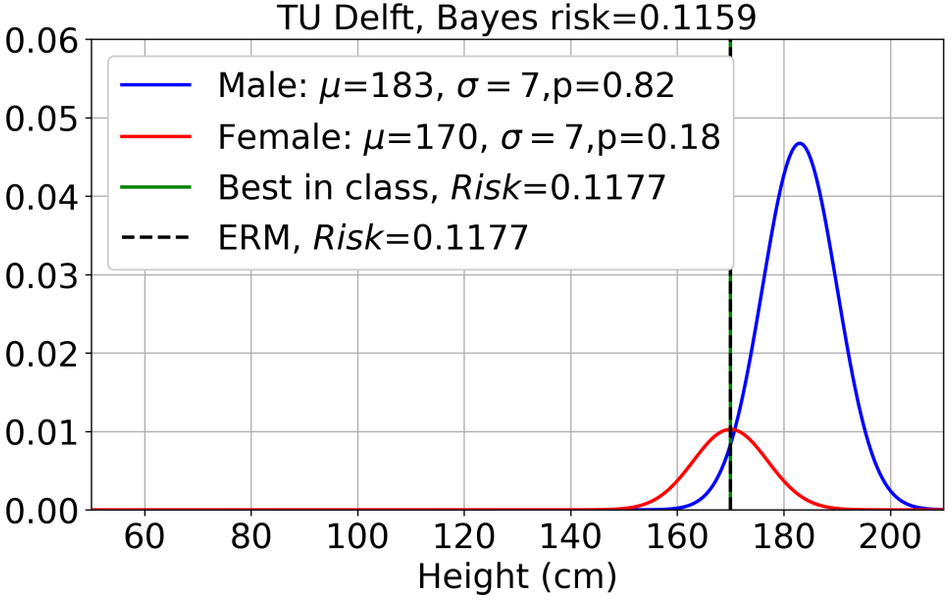
$$\mathbb{P}_{T_m \sim p^m}\Big[\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon}_{\text{approximately correct}} \Big] \geq 1 - 2\,|\mathcal{H}|\, e^{-\frac{1}{2} m \varepsilon^2} = \underbrace{1 - \delta}_{\text{probably}}$$

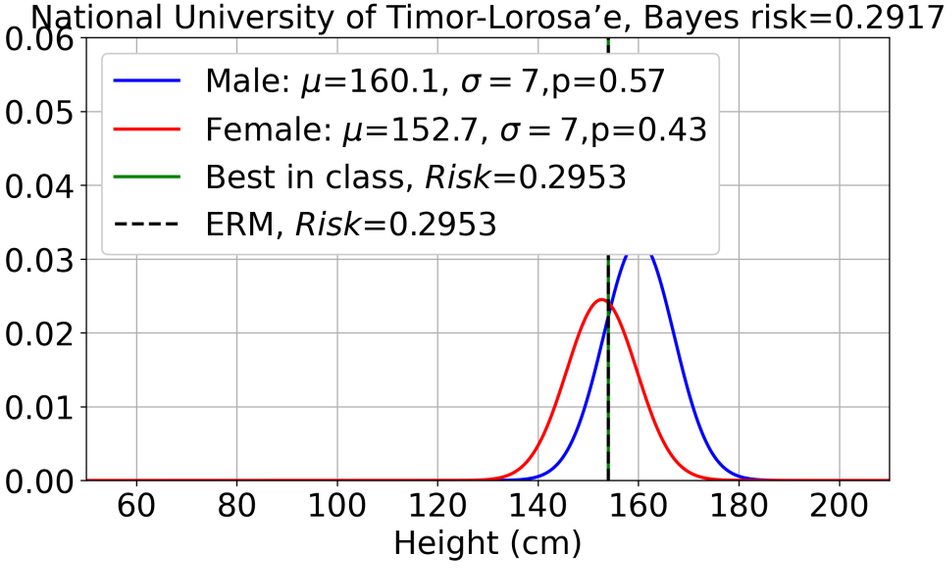
Probably Approximately Correct learning

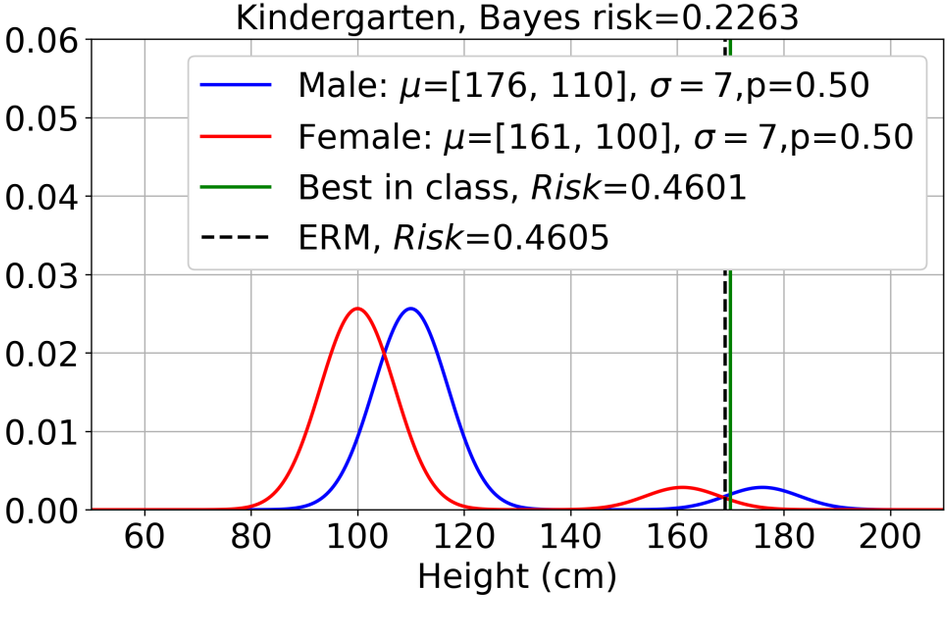
Example: $\varepsilon = 0.05$, $\delta = 0.1$, $\mathcal{H} = \{h(x;\theta) = \text{sign}(x - \theta) \mid \theta \in \{140, 141, \ldots, 200\}\}$

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

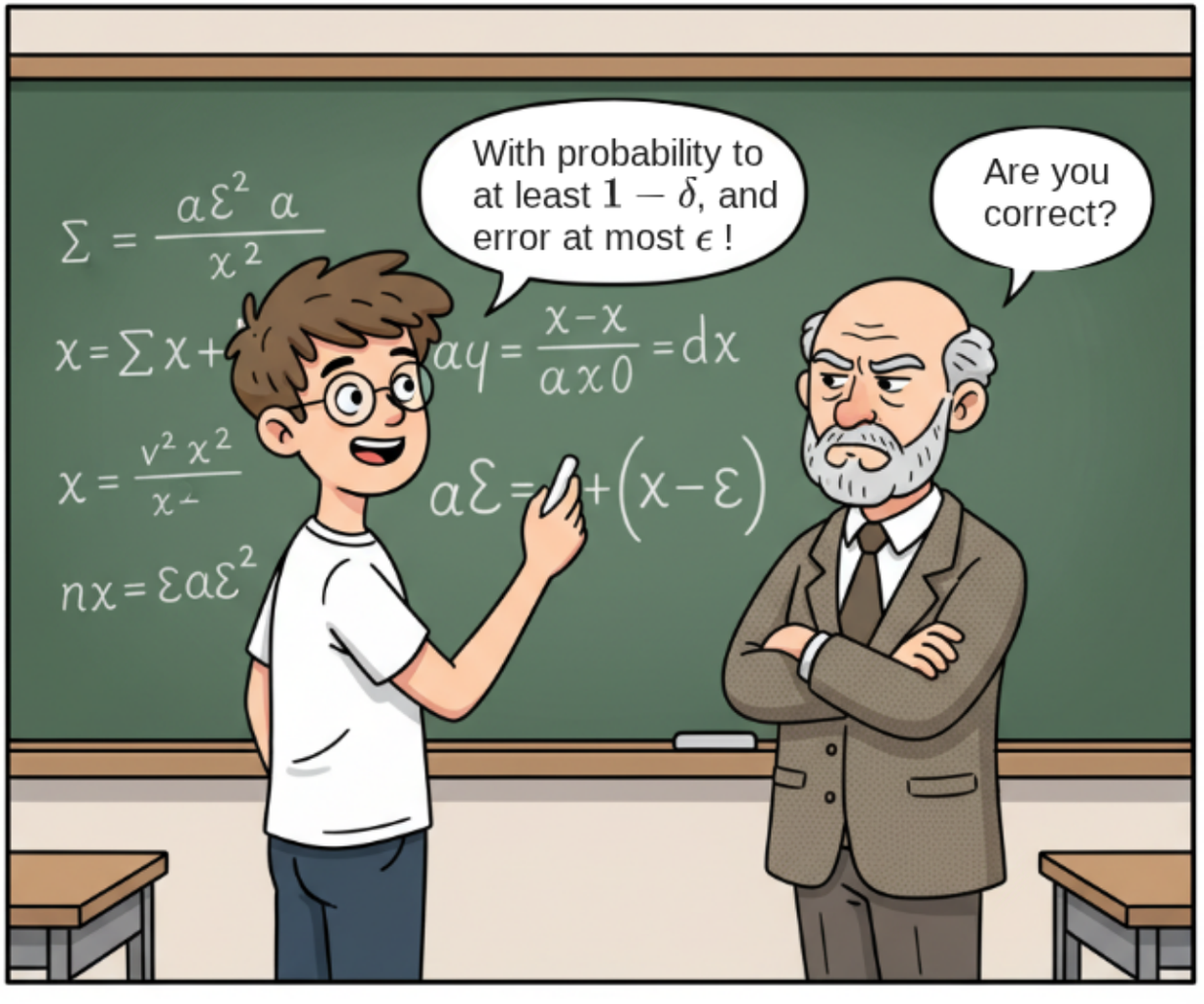








Probably Approximately Correct Joke



BE4M33SSU exam edition.

Joke: Chat GPT-5

Image: Nano Banana

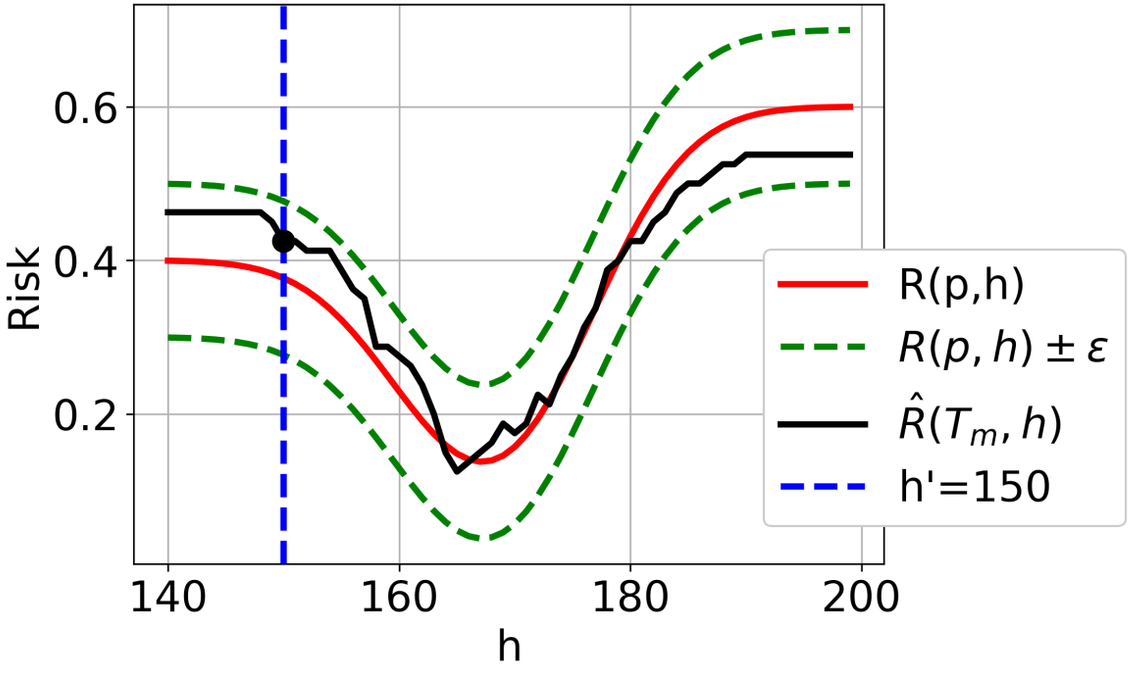
Uniform Law of Large Numbers

- **LLN**: $m \geq \frac{1}{2\varepsilon^2} \log\left(\frac{2}{\delta}\right) \quad \Rightarrow \quad \mathbb{P}\Big(\underbrace{\big|R(p,h) - \hat{R}(T_m,h)\big| > \varepsilon}_{\text{generalization error of } h \text{ is high}} \Big) \leq \delta$

  LLN applies for any $h\colon \mathcal{X} \to \mathcal{Y}$ fixed prior to observating the data T_m

- **ULLN:** $m \geq m_{\text{ul}}^{\mathcal{H}}(\varepsilon,\delta) \quad \Rightarrow \quad \mathbb{P}\Big(\underbrace{\max_{h\in\mathcal{H}} \big|R(p,h) - \hat{R}(T_m,h)\big| > \varepsilon}_{\text{generalization error of some } h\in\mathcal{H} \text{ is high}} \Big) \leq \delta$

  ULLN applies only for some $\mathcal{H}$, e.g., when $\mathcal{H}$ is finite $m_{\text{ul}}^{\mathcal{H}}(\varepsilon,\delta) = \frac{1}{2\varepsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$



Example:

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

$\varepsilon = 0.1,\ \delta = 0.05,\ |\mathcal{H}| = 60$

$m_{\text{ul}}^{\mathcal{H}}(0.1, 0.05) = 389.2$

Uniform Law of Large Numbers

- Assume a finite hypothesis class $\mathcal{H} = \{h^i \colon \mathcal{X} \to \mathcal{Y} \mid i \in \{1, 2, \ldots, H\}\}$.

$$\mathbb{P}\Big(\underbrace{\max_{h \in \mathcal{H}} \big|R(p,h) - \hat{R}(T_m, h)\big| \geq \varepsilon}_{\text{generalization error of some } h \in \mathcal{H} \text{ is high}} \Big) \overset{(1)}{=} \mathbb{P}\left(\begin{array}{ll} \big|R(p,h^1) - \hat{R}(T_m,h^1)\big| \geq \varepsilon & \text{or} \\ \big|R(p,h^2) - \hat{R}(T_m,h^2)\big| \geq \varepsilon & \text{or} \\ \vdots & \\ \big|R(p,h^H) - \hat{R}(T_m,h^H)\big| \geq \varepsilon & \end{array} \right)$$
$$\overset{(2)}{\leq} \sum_{h \in \mathcal{H}} \mathbb{P}\Big(\big|R(p,h) - \hat{R}(T_m,h)\big| \geq \varepsilon \Big)$$
$$\overset{(3)}{\leq} 2\,|\mathcal{H}|\, e^{-2m\varepsilon^2}$$

1. $a \geq \varepsilon$ or $b \geq \varepsilon \iff \max\{a,b\} \geq \varepsilon$
2. Union bound: $\mathbb{P}\Big(A_1 \text{ or } A_2 \text{ or } \cdots \text{ or } A_n\Big) \leq \sum_{i=1}^{n} \mathbb{P}(A_i)$
3. Hoeffding inequality: $\mathbb{P}\Big(\big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon\Big) \leq 2\, e^{-2\,m\,\varepsilon^2}$

- Setting $2\,|\mathcal{H}|\, e^{-2m\varepsilon^2} = \delta$ and solving for m, we get

$$m_{\text{ul}}^{\mathcal{H}}(\varepsilon, \delta) = \frac{1}{2\varepsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right) \qquad \Rightarrow \qquad \mathbb{P}\Big(\max_{h \in \mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon \Big) \leq \delta$$

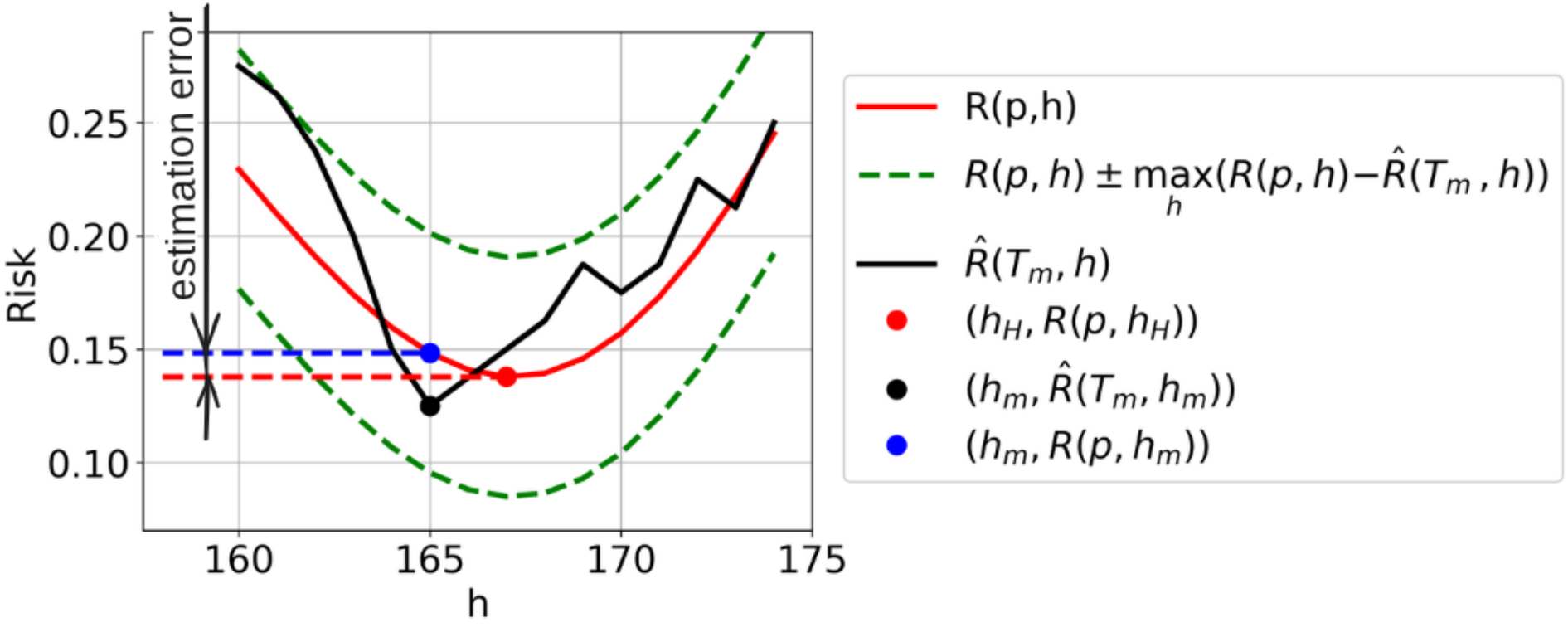
Bound on Estimation Error

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

$$\underbrace{R(p, h_m) - R(p, h_{\mathcal{H}})}_{\textit{estimation error}} = \Big(R(p, h_m) - \hat{R}(T_m, h_m)\Big) + \Big(\hat{R}(T_m, h_m) - R(p, h_{\mathcal{H}})\Big)$$

$$\leq \Big(R(p, h_m) - \hat{R}(T_m, h_m)\Big) + \Big(\hat{R}(T_m, h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big)$$

$$\leq 2 \underbrace{\max_{h \in \mathcal{H}} \big|R(p, h) - \hat{R}(T_m, h)\big|}_{\text{Maximal generalization error}}$$



ULLN implies ERM is PAC learner

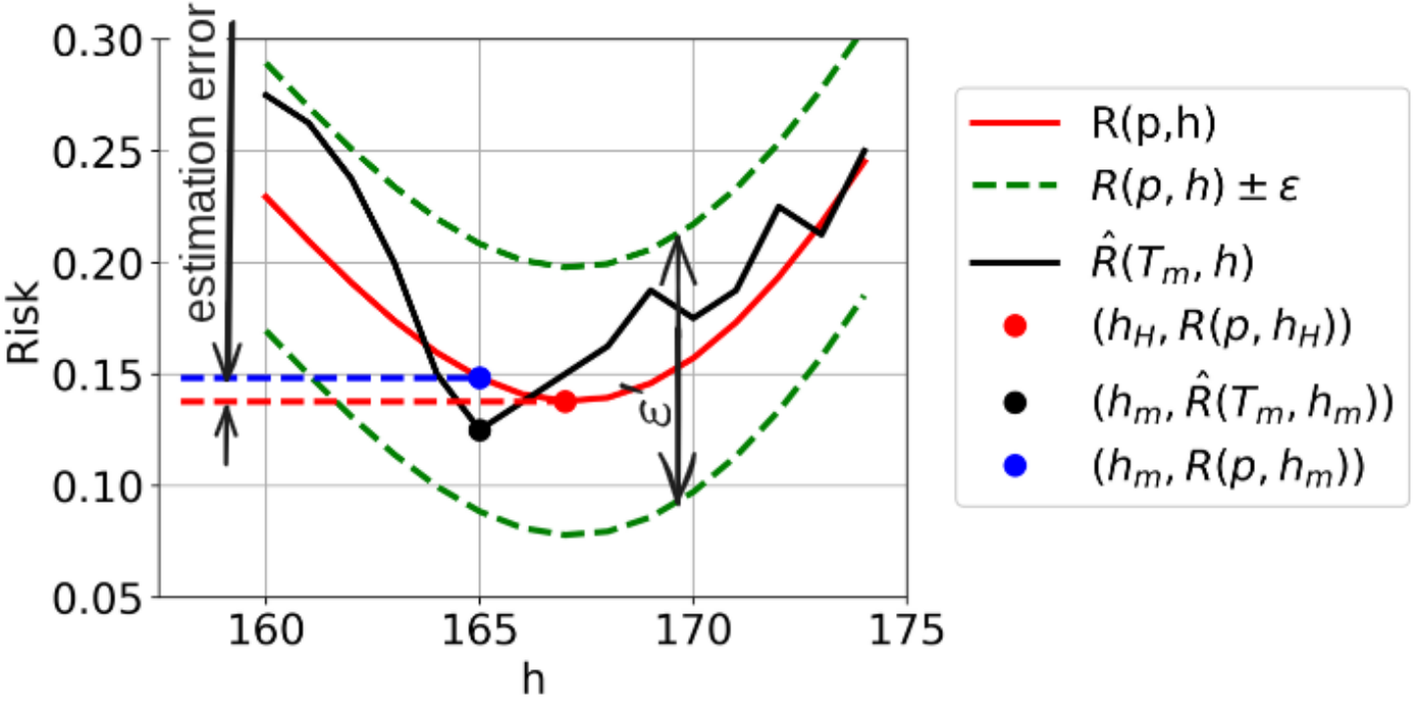
- **ULLN:** $m \geq m_{\text{ul}}^{\mathcal{H}}(\varepsilon, \delta) \quad \Rightarrow \quad \mathbb{P}\Big(\max_{h \in \mathcal{H}} \big| R(p,h) - \hat{R}(T_m, h) \big| > \varepsilon \Big) \leq \delta$

- **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 + Bound on estimation error = ERM is succesfull PAC learner**:

$$m \geq m_{\text{pac}}^{\mathcal{H}}(\varepsilon', \delta) = m_{\text{ul}}^{\mathcal{H}}(\varepsilon'/2, \delta) \quad \Rightarrow \quad \mathbb{P}\Big(R(p, h_m) - R(p, h_{\mathcal{H}}) \leq \varepsilon' \Big) \geq 1 - \delta$$



Example:

For finite $\mathcal{H}$, we have:

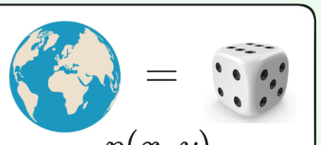
$$m_{\text{ul}}^{\mathcal{H}}(\varepsilon, \delta) = \frac{1}{2\varepsilon^2} \log\left(\frac{2|\mathcal{H}|}{\delta}\right)$$

Hence, the sample complexity is:

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

Summary of Key Concepts

Empirical Risk Minimization



$p(x, y)$

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

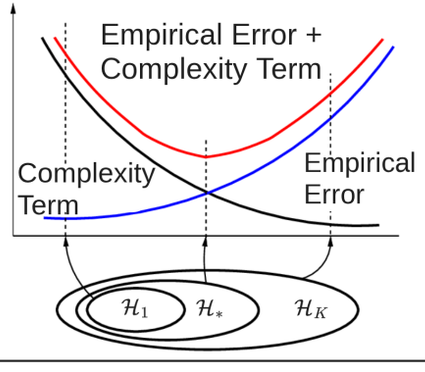
ERM is not PAC learner

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

- **Error Decomposition**
  - Error of the learned predictor = Estimation Error + Approximation Error + Bayes Risk
- **Probably Approximately Correct (PAC) Learning**
  - A successful PAC learner, with high probability, finds a close approximation of the best predictor in the class, given enough examples.
  - Sample complexity: number of examples needed for PAC guarantees.
- **Empirical Risk Minimization:**
  - ERM over a finite hypothesis space is a successful PAC learner.
- **Uniform Law of Large Numbers**
  - Guarantees uniform convergence of empirical risk to the expected risk over the hypothesis space.