Statistical Machine Learning (BE4M33SSU)
Lecture 2: Predictor Evaluation and Empirical Risk Minimization.

Czech Technical University in Prague
V. Franc

Law of Large Numbers

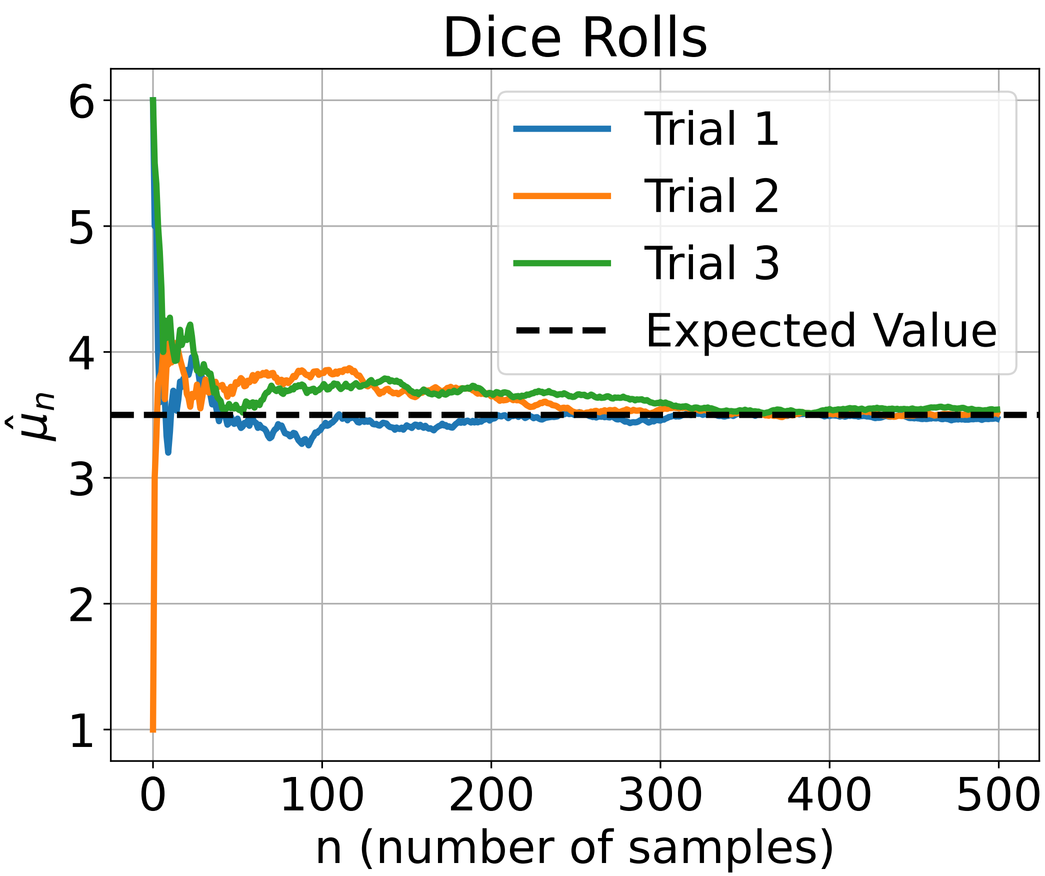
- The sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$ of i.i.d. sample $Z_n = (z_1, z_2, \ldots, z_n)$ generated from $q(z)$ gets closer to the expected value $\mu = \mathbb{E}_{z\sim q}[z]$ as the sample size n increases.

- Example: The expected value of a single roll of a fair die is

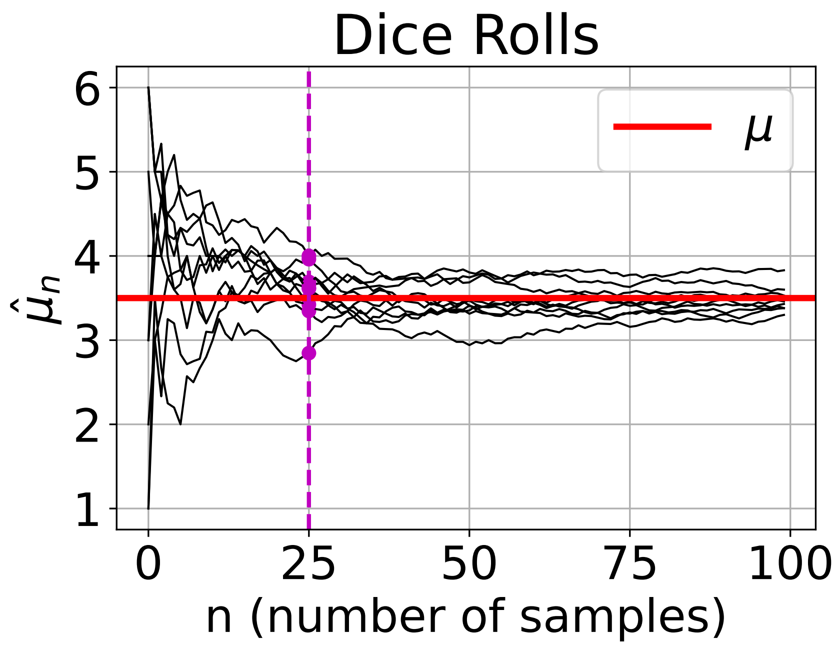
$$\mu = \mathbb{E}_{z\sim q}[z] = \sum_{z=1}^{6} z\, q(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$z_1 = 3$ $z_2 = 1$ $z_3 = 5$ $z_n = 2$



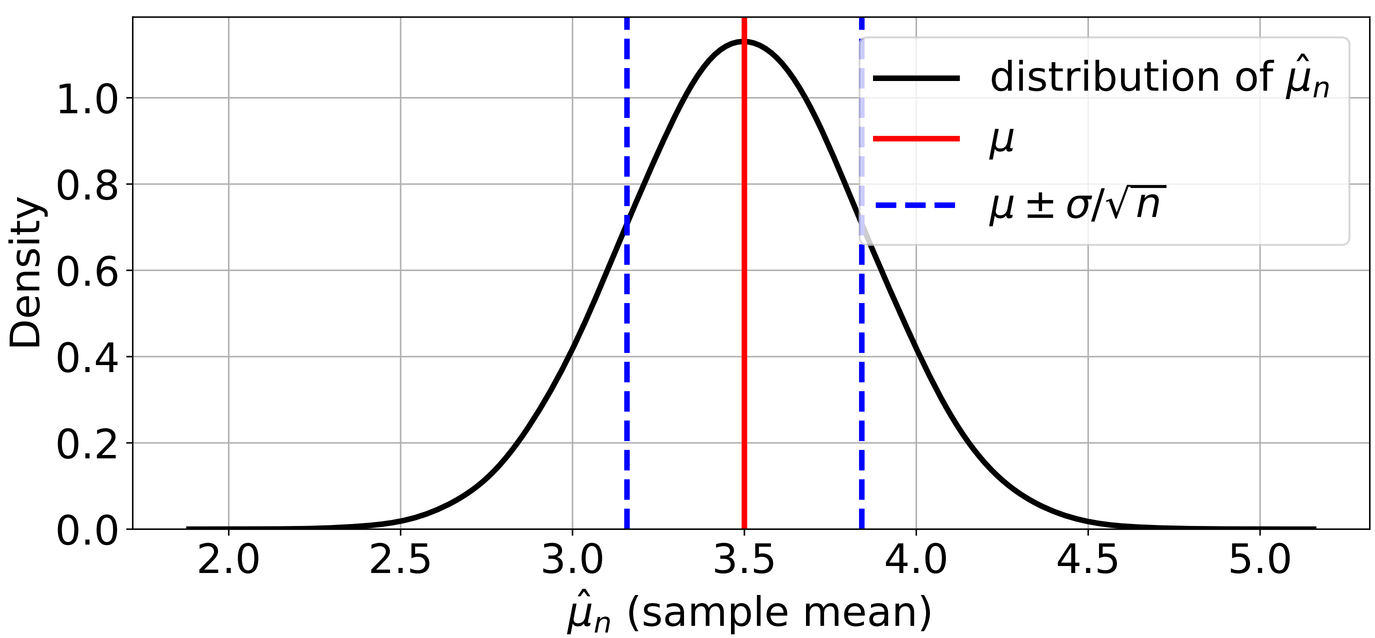
Law of large numbers



$$\mu = \mathbb{E}_{z\sim q}[z] = \frac{1+2+3+4+5+6}{6} = 3.5$$
$$\sigma^2 = \mathbb{V}_{z\sim q}[z] = \frac{(1-3.5)^2+\cdots+(6-3.5)^2}{6} \approx 2.917$$

The sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} z_i$ is an unbiased estimator of the expected value μ:

$$\mathbb{E}_{(z_1,\ldots,z_n)\sim q^n}[\hat{\mu}_n] = \mu$$

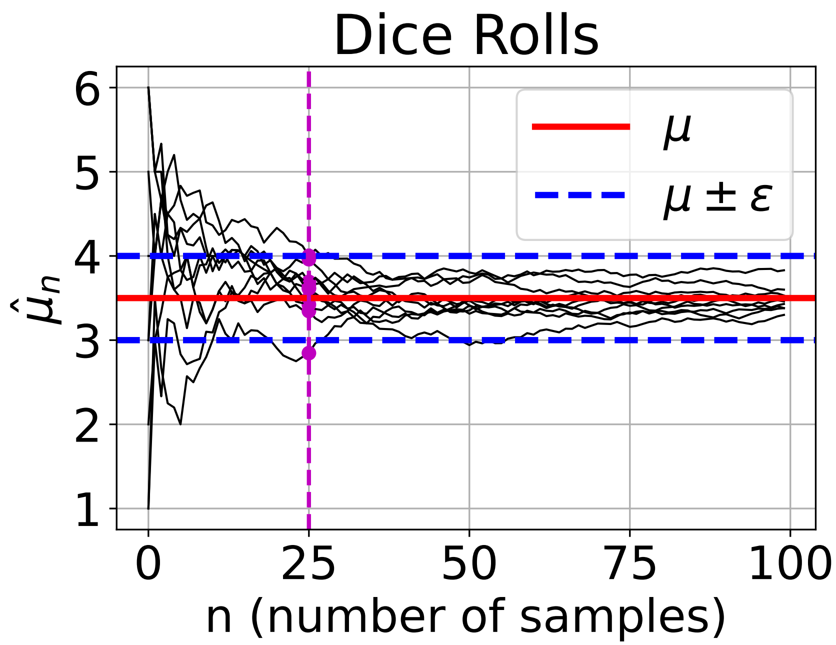


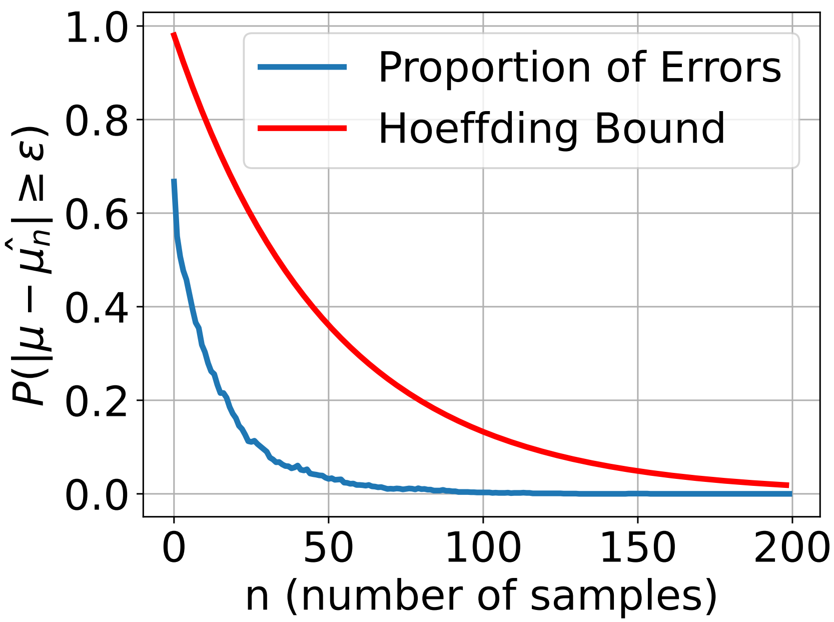
The variance of $\hat{\mu}_n$ decays with $\frac{1}{n}$:

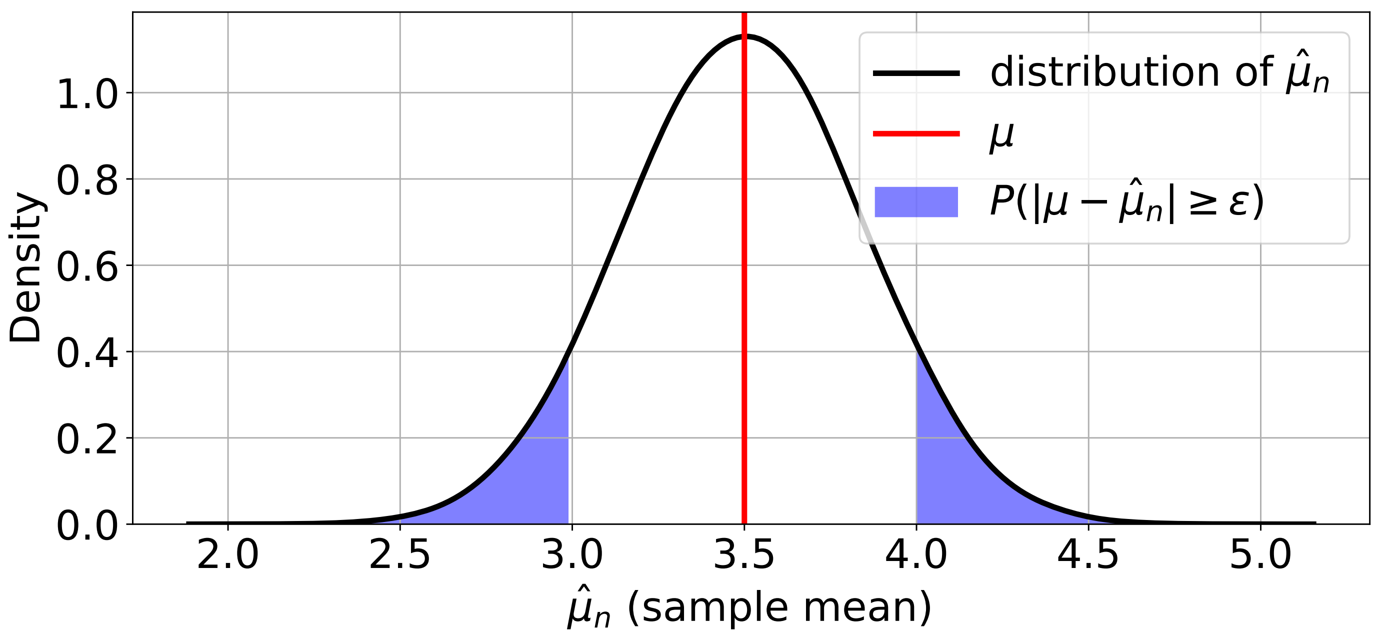
$$\mathbb{V}_{(z_1,\ldots,z_n)\sim q^n}[\hat{\mu}_n] = \frac{\sigma^2}{n}$$

$$n = 25 \rightarrow \frac{\sigma^2}{n} = 0.116$$

Hoeffding inequality







Hoeffding inequality:

$$\mathbb{P}(|\mu - \hat{\mu}_n| \geq \varepsilon) \leq 2\, e^{-\frac{2\,n\,\varepsilon^2}{(b-a)^2}}$$

$$a = 1,\ b = 6,\ \varepsilon = 0.5$$

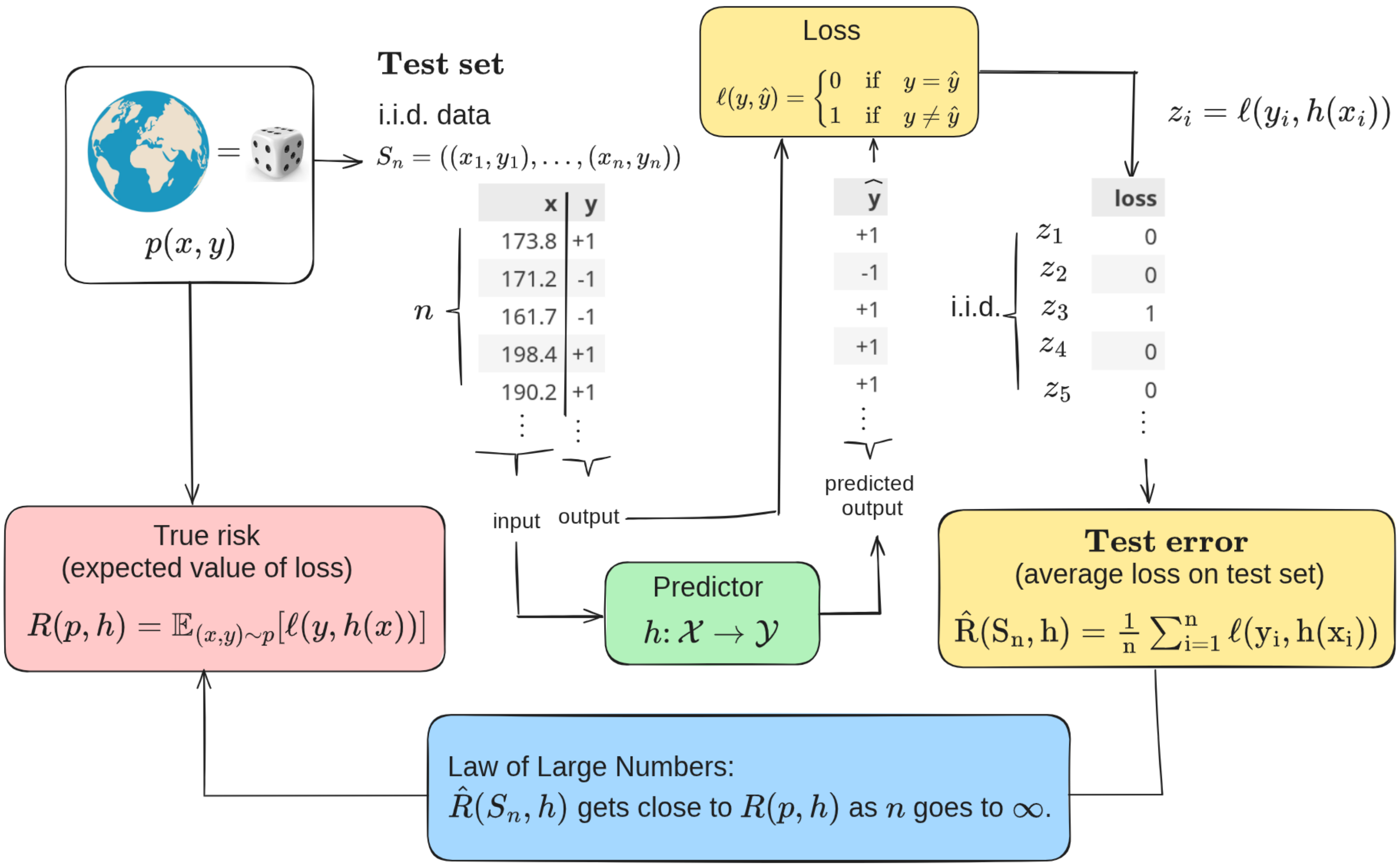
Hoeffding inequality

- **Theorem:** Let $Z_n = (z_1, \ldots, z_n)$ be i.i.d. sample generated from r.v. with distribution $q(z)$. Let the random variables attain values from an interval $[a, b]$. Let the expected value of the r.v. be $\mu = \mathbb{E}_{z \sim q}[z]$. Let $\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} z_i$. Then, for any $\varepsilon > 0$:

$$\mathbb{P}_{Z_n \sim q^n}\Big(|\hat{\mu}_n - \mu| \geq \varepsilon \Big) \leq 2e^{-\frac{2\,n\,\varepsilon^2}{(b-a)^2}}$$

- **Key Properties:**
  - $(+)$ **General:** the bound holds for any bounded i.i.d. random variables.
  - (-) **Conservative:** the bound is typically not tight.
  - $(+)$ **Vanishing:** the bound $\to 0$ as $n \to \infty$.
  - $(+)$ **Cheap:** the bound is simple and easy to compute.

Predictor evaluation



Application of the Hoeffding inequality:

$$\mathbb{P}_{Z_n \sim q^n}\Big(\big|\hat{\mu}_n - \mu\big| \geq \varepsilon\Big) \leq 2e^{-\frac{2\,n\,\varepsilon^2}{(b-a)^2}} \quad \Rightarrow \quad \mathbb{P}_{S_n \sim p^n}\Big(\big|\hat{R}(S_n, h) - R(p, h)\big| \geq \varepsilon\Big) \leq 2e^{-\frac{2\,n\,\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}}$$

Confidence interval

- **Goal:** Characterize the deviation between the true risk $R(p,h) = \mathbb{E}_{(x,y)\sim p}[\ell(y,h(x))]$ and the test error $\hat{R}(S_n,h) = \frac{1}{n}\sum_{i=1}^n \ell(y^i,h(x^i))$ computed on i.i.d. data $S_n \sim p^n$.

- **Hoeffding inequality:** provides a probabistic bound on the deviation between the true risk $R(p,h)$ and the test errork $\hat{R}(S_n,h)$:

$$\mathbb{P}_{\mathcal{S}^n\sim p^n}\Big(|R(p,h) - \hat{R}(S_n,h)| \geq \varepsilon\Big) \leq 2e^{-\frac{2\,n\varepsilon^2}{(\ell_{\max}-\ell_{\min})^2}} \qquad \forall \varepsilon > 0$$

- **(1-δ)-Confidence interval:** (derived from the Hoeffding inequality)

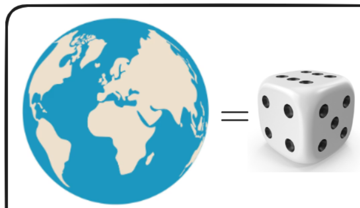
$$R(p,h) \in \big(\hat{R}(S_n,h) - \varepsilon, \hat{R}(S_n,h) + \varepsilon\big) \quad \text{holds with probability} \quad 1-\delta \quad \text{at least,}$$

  where $1-\delta$ is called the confidence level.

  - For fixed n and $\delta \in [0,1]$, compute $\varepsilon = (\ell_{\max} - \ell_{\min})\sqrt{\frac{\log(2)-\log(\delta)}{2\,n}}$
  - For fixed ε and $\delta \in [0,1]$, compute $n = \frac{\log(2)-\log(\delta)}{2\varepsilon^2}\,(\ell_{\max} - \ell_{\min})^2$

Empirical Risk Minimization

ERM: a principle to construct algorithms learning predictors from data.



$p(x, y)$

Training set

i.i.d. data

$T_m = ((x_1, y_1), \dots, (x_m, y_m))$

x	**y**
179.5	+1
161.5	+1
176.9	+1
157.7	-1
184.0	+1
167.4	+1
175.1	+1

Hypothesis class

$\mathcal{H} = \{h(x; \theta) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\}$

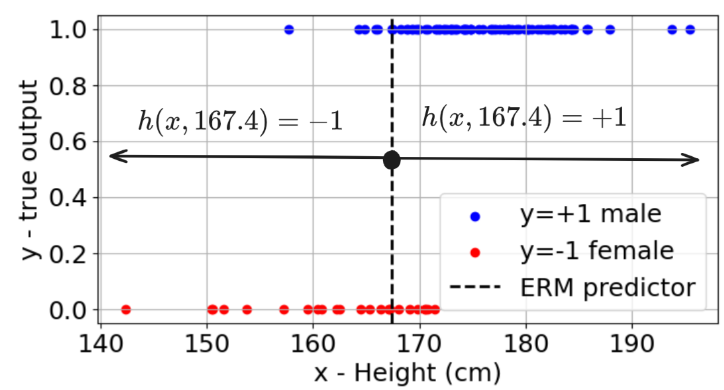
Loss function

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

Learning algorithm based on ERM

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

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



Predictor

$h_m(x, 167.4) = \text{sign}(x - 167.4)$

Instances: Linear regression, Logistic Regression, Neural Networks learn by back-propagation, Gradient Boosted Trees, ...

Empirical Risk Minimization

- **Goal:** Given a training set $T_m = ((x_1, y_1), \ldots, (x_m, y_m)) \sim p^m$, learn a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing the expected risk $R(p, h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x))]$.

- **Hypothesis class (space):** is fixed before learning based on prior knowledge

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

- **Learning algorithm:** is a function

$$A\colon \cup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}$$

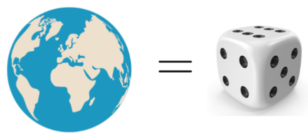
- **Emprirical risk** evaluated on T_m (a.k.a training error):

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

- **ERM based learning algorithm**:

$$h_m = A(T_m) = \operatorname*{Argmin}_{h \in \mathcal{H}} \hat{R}(T_m, h)$$

When does ERM work?



$p(x, y)$

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

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

Learning algorithm based on ERM

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

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

Plan for the next lectures:

PAC learning

Successful PAC learner = with a high probability it finds close approximation of the best predictor in $\mathcal{H}$.

"Too complex" hypothesis space

$h \colon \mathcal{X} \to \mathcal{Y}$

ERM is not PAC learner

Finite hypothesis space

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

ERM is PAC learner

VC dimension

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

$\mathrm{VCdim}(\mathcal{H}) < \infty$ $\iff$ ERM is PAC learner

ERM can fail when the hypothesis class is too complex

- Setup: $\mathcal{X} = [a,b] \subset \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = [\![y \neq y']\!]$, $p(x \mid y = +1)$ and $p(x \mid y = -1)$ uniform on $\mathcal{X}$, with $p(y = +1) = 0.8$.

- Optimal predictor: $h^*(x) = +1$ with true risk $R(p, h^*) = 0.2$.

- "Memorizer" learning rule: for training set $T_m = ((x_1, y_1), \ldots, (x_m, y_m))$, define

$$h_m(x) = \begin{cases} y_j & \text{if } x = x_j \text{ for some } j \in \{1, \ldots, m\}, \\ -1 & \text{otherwise.} \end{cases}$$

  - Implements ERM: $\mathbb{P}\big(\hat{R}(T_m, h_m) = 0\big) = 1$.
  - Performs poorly: $\mathbb{P}\big(R(p, h_m) = 0.8\big) = 1$ for any finite m.

- **Overfitting**: occurs when $h_m = A(T_m)$ achieves low empirical risk $\hat{R}(T_m, h_m)$ while the true risk $R(p, h_m)$ is high.

Summary of Key Concepts

- Law of Large Numbers
  - The sample mean converges to the expected value as the number of samples increases.
  - Hoeffding's inequality: bounds the probability of deviation between the sample mean and the expected value.
- Predictor Evaluation
  - For a fixed $h\colon \mathcal{X} \to \mathcal{Y}$, the losses $\ell(y_i, h(x_i))$ on an i.i.d. sample S_n are themselves i.i.d. random variables.
  - Confidence intervals provide bounds on the estimation error.
- Empirical Risk Minimization (ERM)
  - ERM selects a predictor that minimizes empirical error on the training set.
  - ERM may fail if the hypothesis class $\mathcal{H}$ is too complex.