Statistical Machine Learning (BE4M33SSU)
Lecture 2: Empirical Risk Minimization I

Czech Technical University in Prague
V. Franc

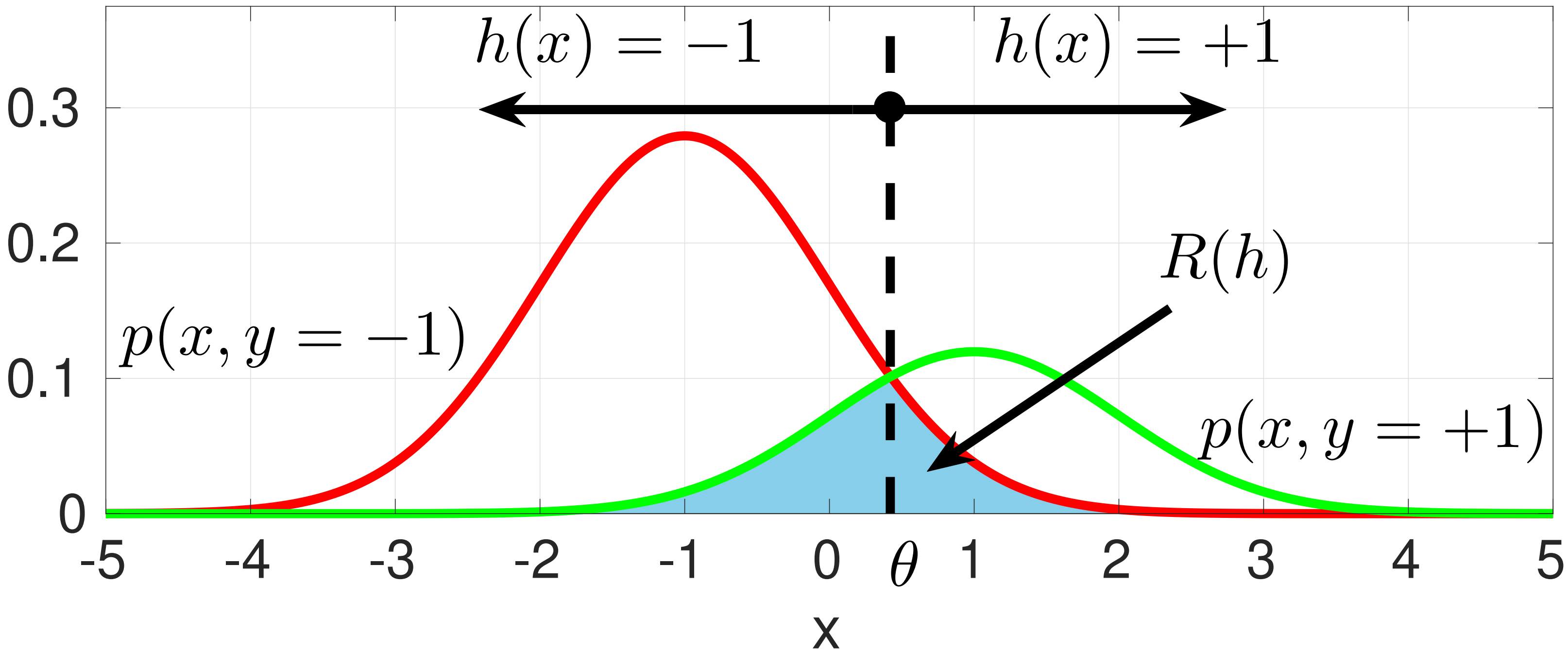
Prediction problem: the definition

- $\mathcal{X}$ a set of input **observations/features**
- $\mathcal{Y}$ a finite set of **hidden states**
- $(x, y) \in \mathcal{X} \times \mathcal{Y}$ samples **randomly drawn** from r.v. with p.d.f. $p(x, y)$
- $h \colon \mathcal{X} \to \mathcal{Y}$ a **prediction strategy**
- $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a **loss function**
- **Task** is to find a strategy with the minimal **expected risk**

$$R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x))\, p(x, y)\, \mathrm{d}x = \mathbb{E}_{(x,y) \sim p}\Big(\ell(y, h(x))\Big)$$

Example of a prediction problem

- The statistical model:
  - $\mathcal{X} = \mathbb{R}$, $\mathcal{Y} = \{+1, -1\}$, $\ell(y, y') = \begin{cases} 0 & \text{if} \quad y = y' \\ 1 & \text{if} \quad y \neq y' \end{cases}$
  - $p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}$, $y \in \mathcal{Y}$.



Solving the prediction problem from examples

- **Assumption**: we have an access to examples

$$\{(x^1, y^1), (x^2, y^2), \ldots\}$$

drawn from i.i.d. r.v. distributed according to unknown $p(x, y)$.

- 1) **Testing**: a given $h\colon \mathcal{X} \to \mathcal{Y}$ estimate its $R(h)$ using **test set**

$$\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\}$$

drawn i.i.d. from $p(x, y)$.

- 2) **Learning**: find $h\colon \mathcal{X} \to \mathcal{Y}$ with small $R(h)$ using **training set**

$$\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}$$

drawn i.i.d. from $p(x, y)$.

Testing: estimation of the expected risk

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l$ draw i.i.d. from distribution $p(x,y)$, compute the **empirical risk**

$$R_{\mathcal{S}^l}(h) = \frac{1}{l}\big(\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l))\big) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$$

and use it as an estimate of $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$.

- The empirical risk $R_{\mathcal{S}^l}(h)$ is a random variable.

- We will show how to compute an interval such that

$$R(h) \in (R_{\mathcal{S}^l(h)} - \varepsilon, R_{\mathcal{S}^l(h)} + \varepsilon)$$

holds with a prescribed probability (confidence) $\gamma \in (0,1)$.

- We show how the interval width ε depends on l and γ.

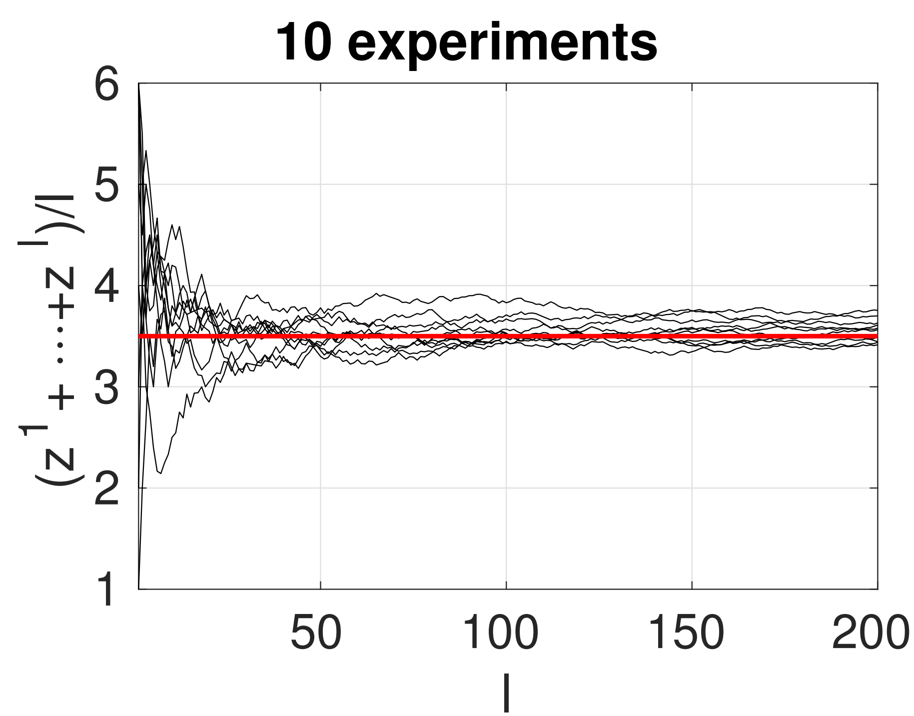
Law of large numbers

- Arithmetic mean of the results of random trials gets closer to the expected value as more trials are performed.

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

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

$z^1 = 3$ $z^2 = 1$ $z^3 = 5$ $z^l = 2$

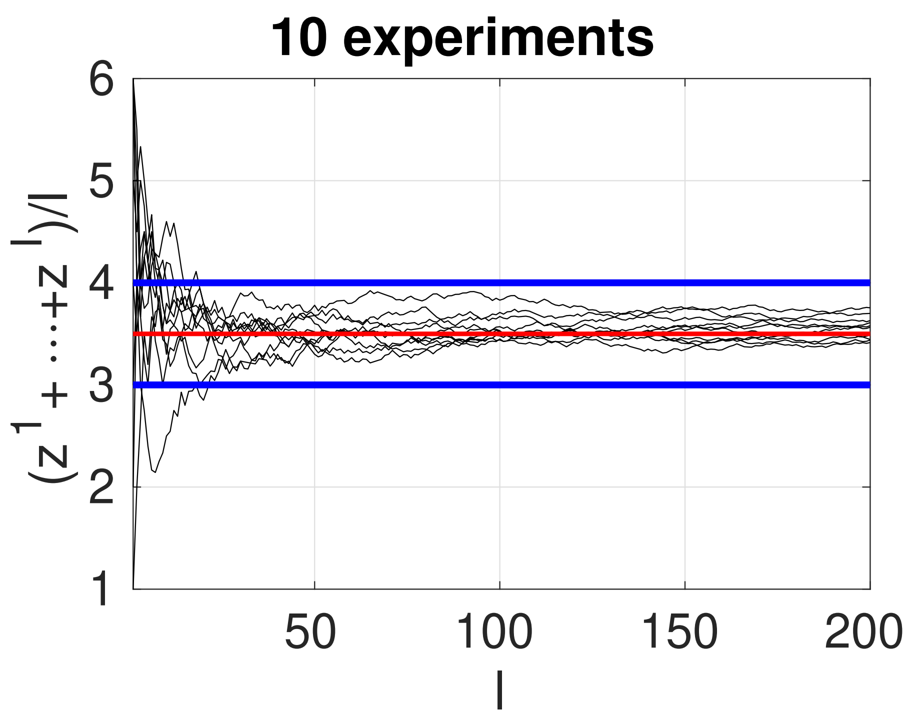


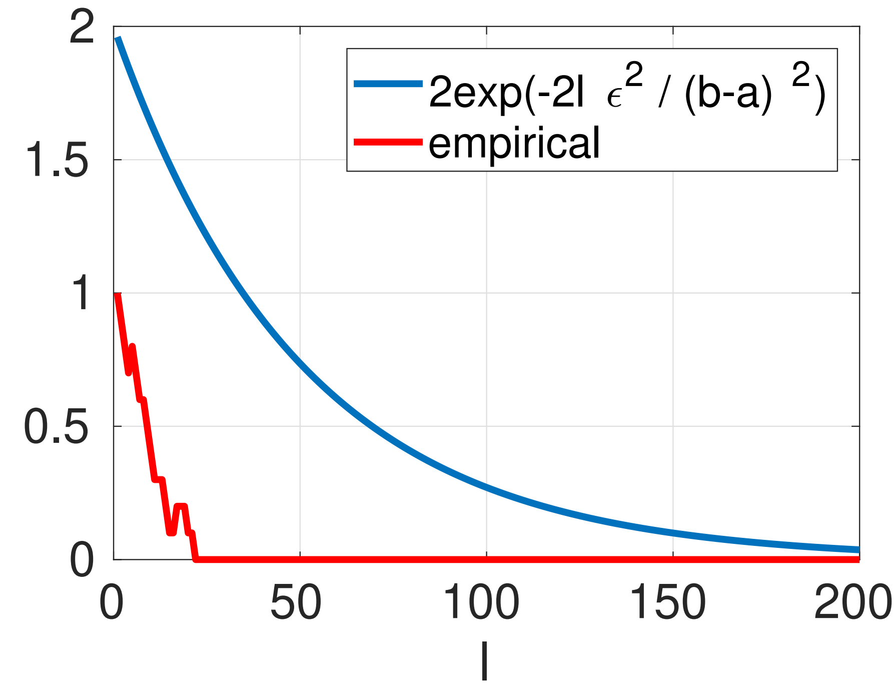
Hoeffding inequality

Theorem 1. *Let $\{z^1, \ldots, z^l\}$ be realizations of independent random variables with the same expected value μ and their values are bounded by an interval $[a, b]$. Then for any $\varepsilon > 0$ it holds that*

$$\mathbb{P}\left(\left|\frac{1}{l}\sum_{i=1}^{l} z^i - \mu\right| \geq \varepsilon\right) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

- Example (rolling a die): $\mu = 3.5$, $z_i \in [1, 6]$, $\varepsilon = 0.5$.





Confidence intervals

- Let $\mu_l = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the arithmetic average computed from $\{z^1, \ldots, z^l\} \in [a,b]^l$ sampled from r.v. with expected value μ.

- Find ε such that $\mu \in (\mu_l - \varepsilon, \mu_l + \varepsilon)$ with probability at least γ.

  Using the Hoeffding inequality we can write

$$\mathbb{P}\Big(|\mu_l - \mu| < \varepsilon\Big) = 1 - \mathbb{P}\Big(|\mu_l - \mu| \geq \varepsilon\Big) \geq 1 - 2e^{-\frac{2\,l\,\varepsilon^2}{(b-a)^2}} = \gamma$$

  and solving the last equation for ε yields

$$\varepsilon = |b-a|\sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}$$

Confidence intervals

- Let $\mu_l = \frac{1}{l}\sum_{i=1}^{l} z^i$ be the arithmetic average computed from $\{z^1,\ldots,z^l\} \in [a,b]^l$ sampled from r.v. with expected value μ.

- Given a fixed $\varepsilon > 0$ and $\gamma \in (0,1)$, what is the minimal number of examples l such that $\mu \in (\mu_l - \varepsilon, \mu_l + \varepsilon)$ with probability γ at least ?

  Starting from

$$\mathbb{P}\Big(|\mu_l - \mu| < \varepsilon\Big) = 1 - \mathbb{P}\Big(|\mu_l - \mu| \geq \varepsilon\Big) \geq 1 - 2e^{-\frac{2\,l\,\varepsilon^2}{(b-a)^2}} = \gamma$$

  and solving for l yields

$$l = \frac{\log(2) - \log(1-\gamma)}{2\,\varepsilon^2}\,(b-a)^2$$

Testing: estimation of the expected risk

- Given $h\colon \mathcal{X} \to \mathcal{Y}$ estimate the expected risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ by the empirical risk $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$ using the test set $\mathcal{S}^l$.

- The incurred losses $z^i = \ell(y^i, h(x^i)) \in [\ell_{\min}, \ell_{\max}]$, $i \in \{1,\ldots,l\}$, are realizations of i.i.d. r.v. with the expected value $\mu = R(h)$.

- According to the Hoeffding inequality, for any $\varepsilon > 0$ the probability of seeing a "bad test set" can be bound by

$$\mathbb{P}\left(\left|R_{\mathcal{S}^l}(h) - R(h)\right| \geq \varepsilon\right) \leq 2e^{-\frac{2l\,\varepsilon^2}{(\ell_{\min}-\ell_{\max})^2}}$$

Testing: confidence intervals

- Given $h\colon \mathcal{X} \to \mathcal{Y}$ estimate the expected risk $R(h) = \mathbb{E}_{(x,y)\sim p}(\ell(y, h(x)))$ by the empirical risk $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$ using the test set $\mathcal{S}^l$.

- **Confidence interval:** the expected risk is

$$R(h) \in \big(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon \,\big)$$

with the probability (confidence) $\gamma \in (0,1)$ at least.

- **Interval width:** For fixed l and $\gamma \in (0,1)$ compute

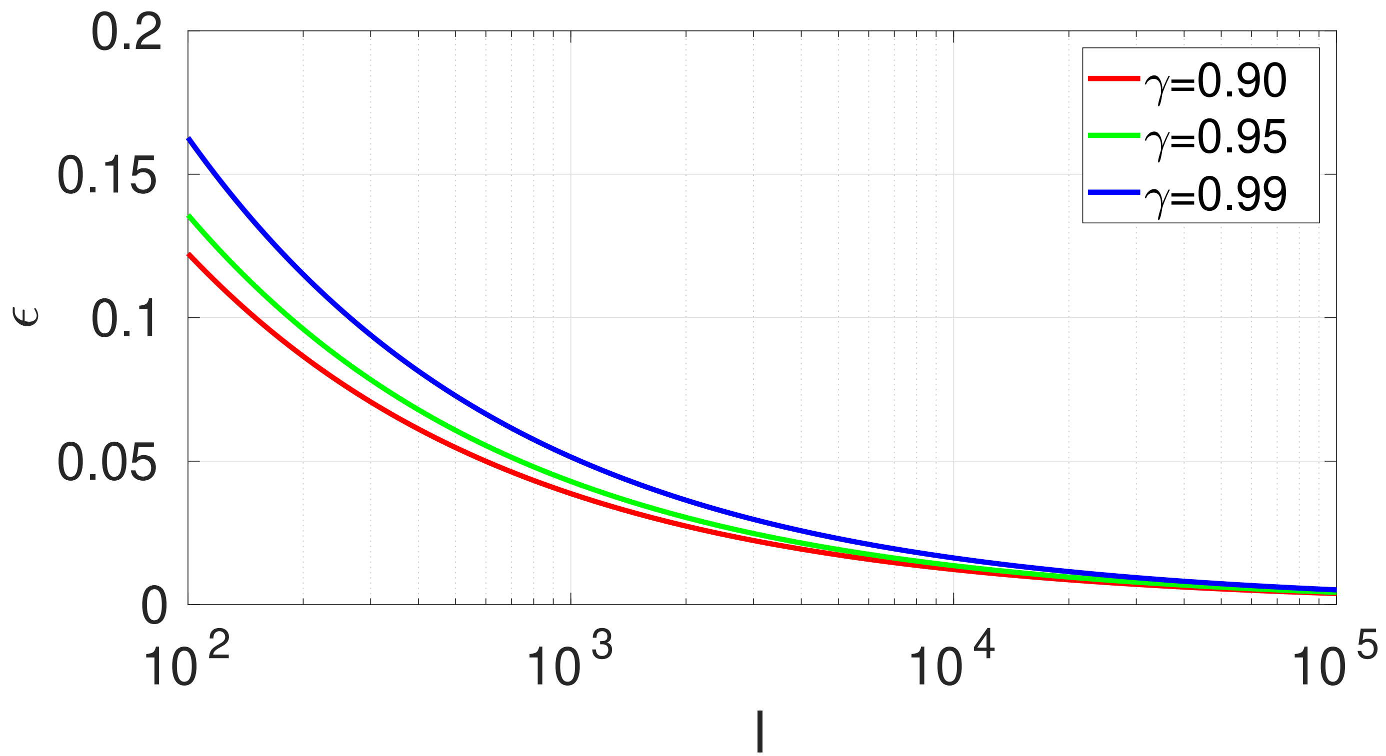
$$\varepsilon = (\ell_{\max} - \ell_{\min}) \sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}\,.$$

- **Number of examples:** For fixed ε and $\gamma \in (0,1)$ compute

$$l = \frac{\log(2) - \log(1-\gamma)}{2\,\varepsilon^2}\,(\ell_{\max} - \ell_{\min})^2$$

Example: confidence intervals

- The width of $R(h) \in \big(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon\big)$ is for $\ell(y,y') = [\![y \neq y']\!]$ given by $\varepsilon = \sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}$



for $\gamma = 0.95$

l	100	1,000	10,000	18,445
ε	0.135	0.043	0.014	0.01