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**

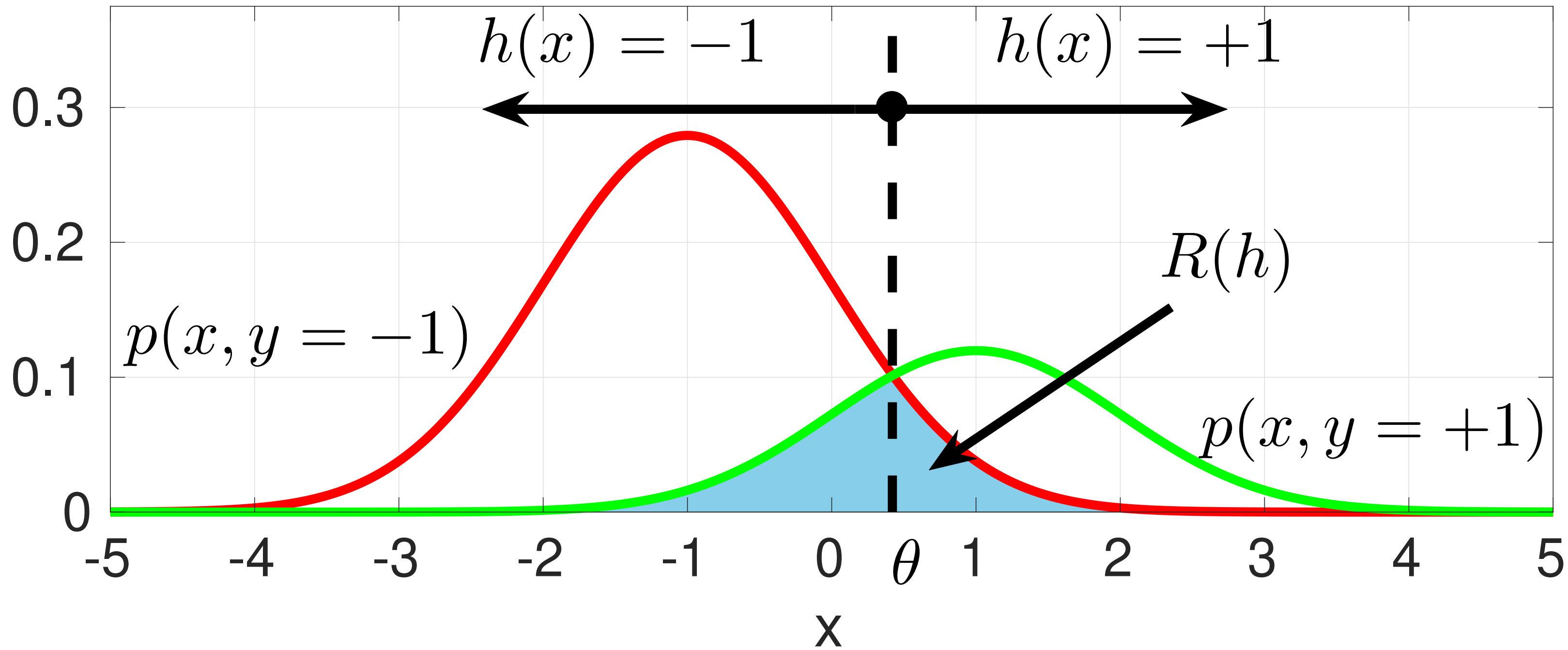
- $\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ a **loss function**

- **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) $\delta \in (0,1)$.

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

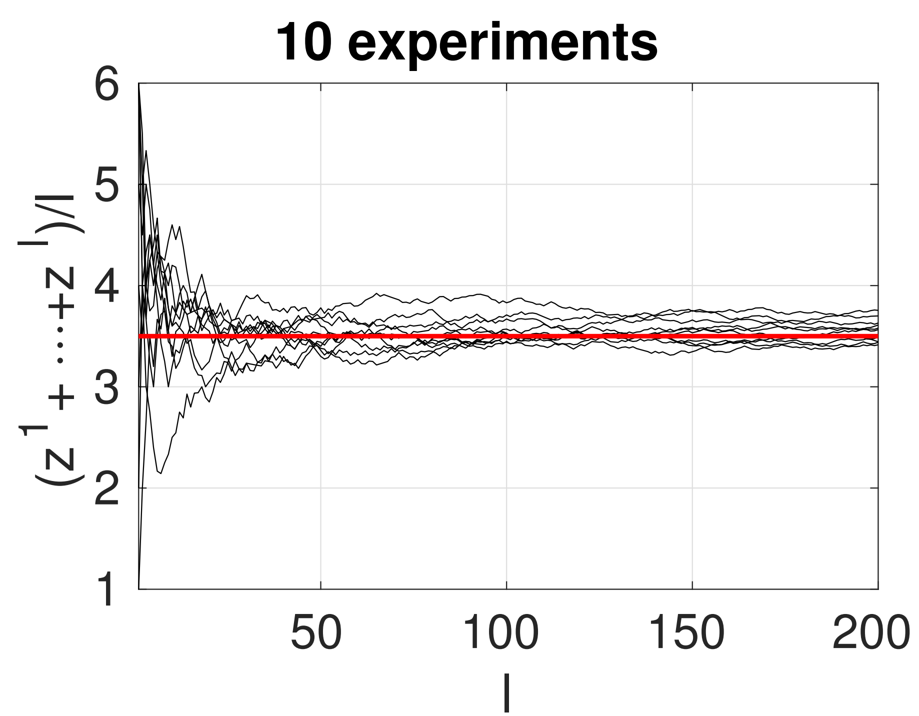
Law of large numbers

- 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

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

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

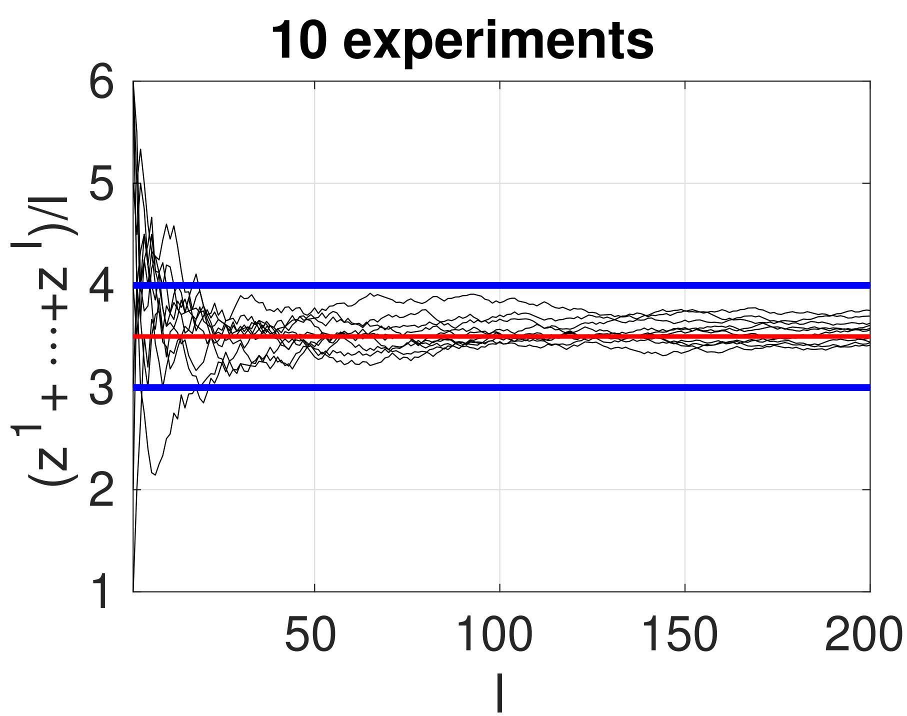


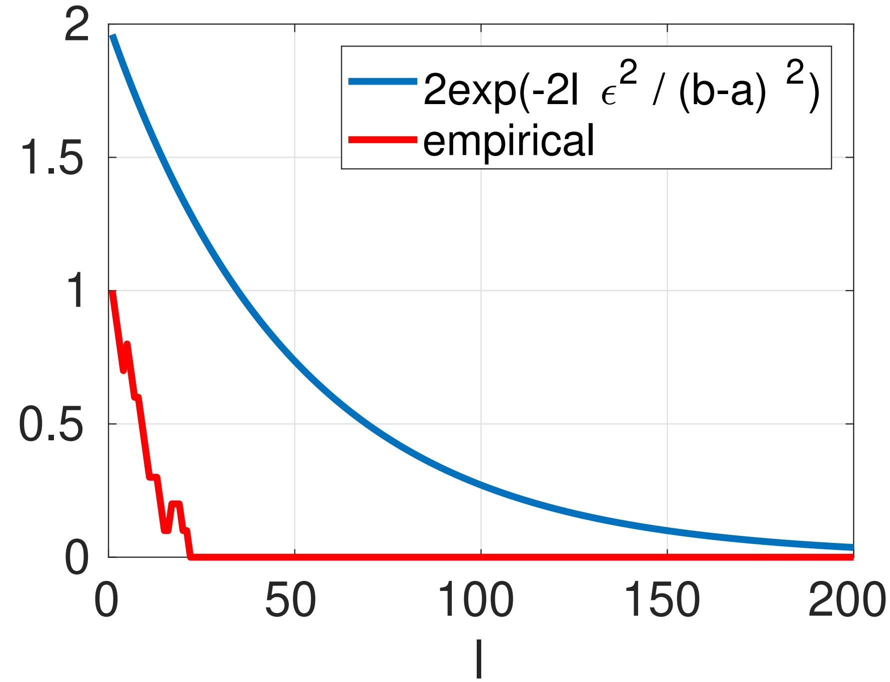
Hoeffding inequality

Theorem 1. *Let* $\{z^1, \ldots, z^l\} \in [a, b]^l$ *be realizations of independent random variables with the same expected value* μ. *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 $\mathcal{S}^l$ drawn i.i.d from $p(x, y)$.

- 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 $\mathcal{S}^l$ drawn i.i.d from $p(x,y)$.

- **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

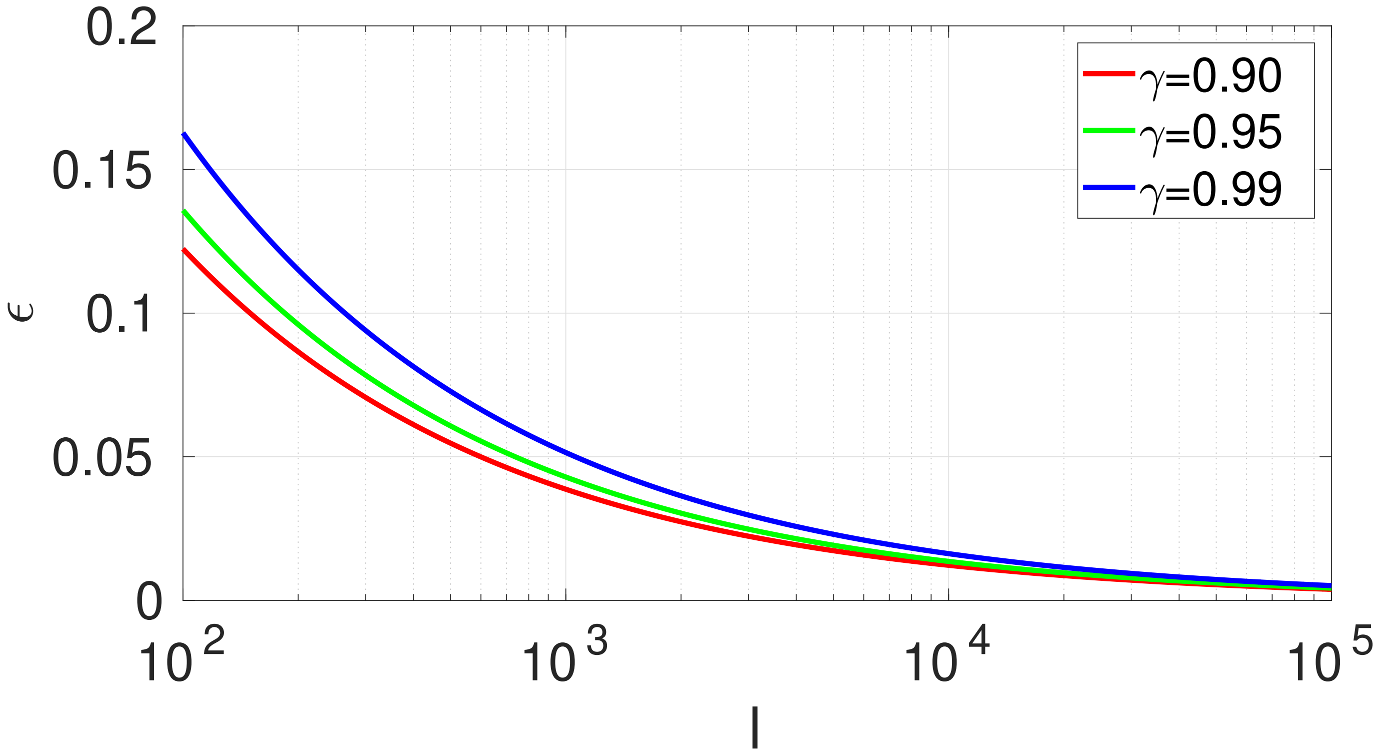
$$\varepsilon = (\ell_{\max} - \ell_{\min})\sqrt{\frac{\log(2) - \log(1-\gamma)}{2\,l}}\,.$$

- **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

Learning

- **The goal:** Find a strategy $h\colon \mathcal{X} \to \mathcal{Y}$ minimizing $R(h)$ using the training set of examples

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

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

- **Hypothesis space:** he have to use our knowledge to select

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

- **Learning algorithm:** a function

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

which returns a strategy $h_m = A(\mathcal{T}^m)$ for a training set $\mathcal{T}^m$

Learning: Empirical Risk Minimization approach

- The expected risk $R(h)$, i.e. the true but unknown objective, is replaced by the empirical risk computed from the training examples

$$R_{\mathcal{T}^m}(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(y^i, h(x^i))$$

- The ERM based algorithm returns h_m such that

$$h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \tag{1}$$

- Depending on the choince of $\mathcal{H}$ and ℓ and algorithm solving (1) we get individual instances e.g. Support Vector Machines, Linear Regression, Logistic Regression, Neural Networks learned by back-propagation, AdaBoost, Gradient Boosted Trees, ...

Errors to minimize when learning

The characters of the play:

- $R^* = \inf_{h \in \mathcal{Y}^\mathcal{X}} R(h)$ best attainable (Bayes) risk
- $R(h_\mathcal{H})$ best risk in $\mathcal{H}$ where $h_\mathcal{H} \in \mathrm{Argmin}_{h \in \mathcal{H}} R(h)$
- $R(h_m)$ risk of $h_m = A(\mathcal{T}_m)$ learned from $\mathcal{T}^m$

Excess error: the quantity we want to minimize

$$\underbrace{\Big(R(h_m) - R^*\Big)}_{\text{excess error}} = \underbrace{\Big(R(h_m) - R(h_\mathcal{H})\Big)}_{\text{estimation error}} + \underbrace{\Big(R(h_\mathcal{H}) - R^*\Big)}_{\text{approximation error}}$$

Questions:

- Which of the quantities are random and which are not ?
- What causes individual errors ?
- How do the errors depend on $\mathcal{H}$ and m?

Statistically consistent learning algorithm

- The statistically consistent algorithm can make the estimation error arbitrarily small if it has enough examples.

- Is the ERM algorithm statistically consistent ?

Definition 1. *The algorithm* $A\colon \cup_{m=1}^{\infty} (\mathcal{X}\times\mathcal{Y})^m \rightarrow \mathcal{H}$ *is statistically consistent in* $\mathcal{H}\subseteq\mathcal{Y}^{\mathcal{X}}$ *if for any* $p(x,y)$ *and* $\varepsilon > 0$ *it holds that*

$$\lim_{m\rightarrow\infty} \mathbb{P}\bigg(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \bigg) = 0$$

where $h_m = A(\mathcal{T}^m)$ *is the hypothesis returned by the algorithm* A *for training set* $\mathcal{T}^m$ *generated from* $p(x,y)$*.*

Example: ERM does not work if $\mathcal{H}$ is unconstrained

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

- The optimal strategy is $h(x) = +1$ with the Bayes risk $R^* = 0.2$.

- Consider learning algorithm which for a given training set $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ returns strategy

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

- The empirical risk is $R_{\mathcal{T}^m}(h_m) = 0$ with probability 1 for any m.

- The expected risk is $R(h_m) = 0.8$ for any m.

Uniform Law of Large Numbers

Definition 2. *The hypothesis space* $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ *satisfies the uniform law of large numbers if for all* $\varepsilon > 0$ *it holds that*

$$\lim_{m\to\infty} \mathbb{P}\bigg(\sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \geq \varepsilon \bigg) = 0$$

- ULLN says that the probability of seeing a “bad training set” can be made arbitrarily low if we have enough examples.

Theorem 2. *If* $\mathcal{H}$ *satisfies ULLN then ERM is statistically consistent in* $\mathcal{H}$.

Proof: ULLN implies consistency of ERM

For fixed $\mathcal{T}^m$ and $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ we have:

$$
\begin{aligned}
R(h_m) - R(h_{\mathcal{H}}) &= \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}})\Big) \\
&\leq \Big(R(h_m) - R_{\mathcal{T}^m}(h_m)\Big) + \Big(R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}})\Big) \\
&\leq 2 \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big|
\end{aligned}
$$

Therefore $\varepsilon \leq R(h_m) - R(h_{\mathcal{H}})$ implies $\frac{\varepsilon}{2} \leq \sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h)\Big|$ and

$$
\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \Big) \leq \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big| \geq \frac{\varepsilon}{2} \Big)
$$

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).

ULLN for finite hypothesis space

- Assume a finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- Define the set of all "bad" training sets for a hypothesis $h \in \mathcal{H}$ as

$$\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \,\middle|\, |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \right\}$$

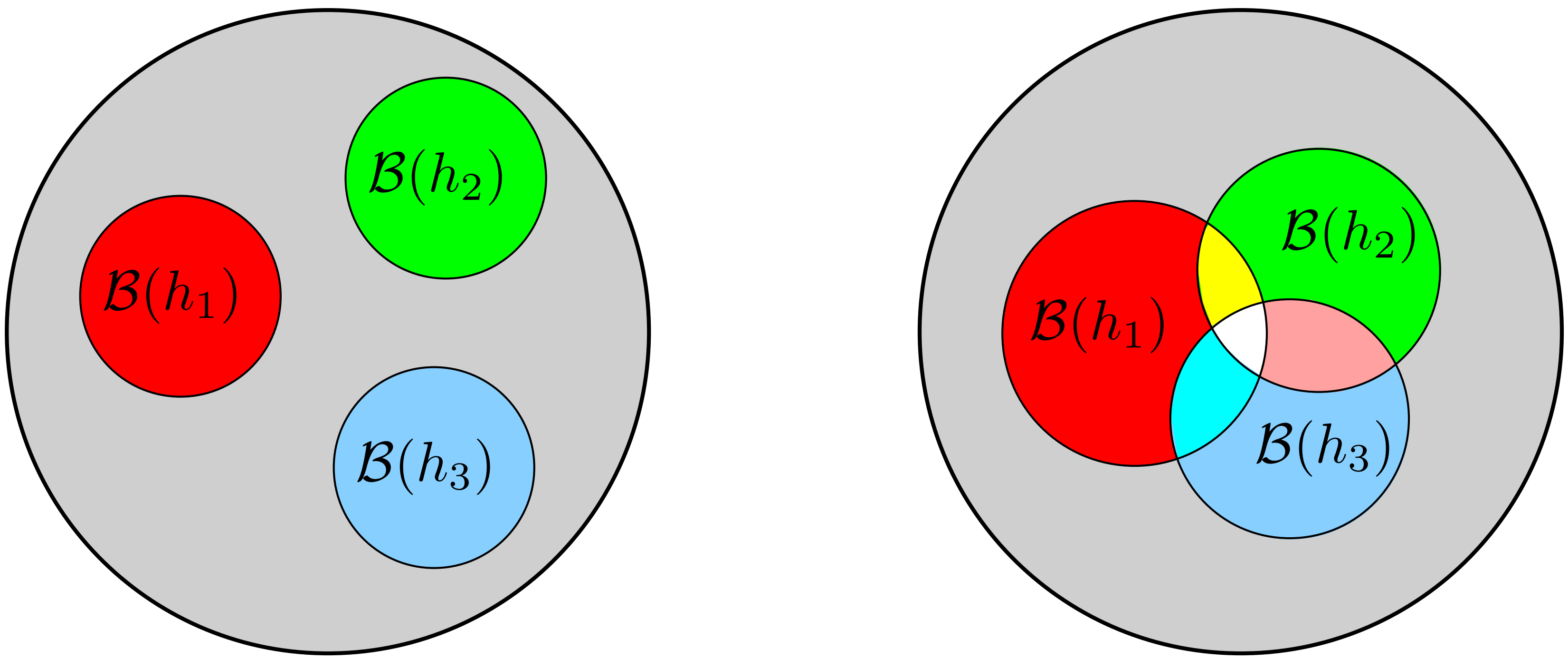
- Use the union bound to upper bound the probability of seeing a bad training set any hypothesis from $h \in \mathcal{H}$

$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big)$$
$$= \mathbb{P}\Big(\mathcal{T}^m \in \mathcal{B}(h_1) \bigvee \mathcal{T}^m \in \mathcal{B}(h_2) \bigvee \cdots \bigvee \mathcal{T}^m \in \mathcal{B}(h_K) \Big)$$
$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h))$$

ULLN for finite hypothesis space

- Example: the union bound for three hypotheses

$$\mathbb{P}\Big(\mathcal{T}^m \in \mathcal{B}(h_1) \bigvee \mathcal{T}^m \in \mathcal{B}(h_2) \bigvee \mathcal{T}^m \in \mathcal{B}(h_3)\Big) \leq \sum_{i=1}^{3} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_i))$$



mutually exclusive

not mutually exclusive

ULLN for finite hypothesis space

- Combining the union bound with the Hoeffding inequality yields

$$\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) \leq \sum_{h\in\mathcal{H}} \mathbb{P}(\underbrace{|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon}_{\mathcal{T}^m\in\mathcal{B}(h)}) \leq 2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

- Therefore we see that

$$\lim_{m\to\infty} \mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) = 0$$

Corollary 1. *The ULLN is satisfied for a finite hypothesis space.*

Generalization bound for finite hypothesis space

- Hoeffding inequality generalized for a finite hypothesis space $\mathcal{H}$:

$$\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \leq 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}}$$

- For which ε is $R(h)$ in the interval $(R_{\mathcal{T}^m}(h) - \varepsilon, R_{\mathcal{T}^m}(h) + \varepsilon)$ with the probability $1-\delta$ at least, regardless what $h \in \mathcal{H}$ we consider ?

$$\begin{aligned}
\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| < \varepsilon \Big) &= 1 - \mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \\
&\geq 1 - 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = 1 - \delta
\end{aligned}$$

and solving the last equality for ε yields

$$\varepsilon = (b-a)\sqrt{\frac{\log 2|\mathcal{H}| + \log\frac{1}{\delta}}{2m}}$$

Generalization bound for finite hypothesis space

Theorem 3. *Let $\mathcal{H}$ be a finite hypothesis space and $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ a training set draw from i.i.d. random variables with distribution $p(x, y)$. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality*

$$R(h) \leq R_{\mathcal{T}^m}(h) + (b - a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$$

holds for any $h \in \mathcal{H}$ and any loss function $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [a, b]$.

- The "worst-case" bound in Theorem 3 holds for any $h \in \mathcal{H}$, in particular, for the ERM algorithm which minimizes the first term.
- The second term suggests that we have to use $\mathcal{H}$ with appropriate cardinality (complexity); e.g. if m is small and $|\mathcal{H}|$ is high we can overfit.

Summary

Topics covered in the lecture:

- Prediction problem
- Test risk and its justification by the law of large numbers
- Empirical Risk Minimization
- Excess error = estimation error + approximation error
- Statistical consistency of learning algorithm
- Uniform law of large numbers
- Generalization bound for finite hypothesis space