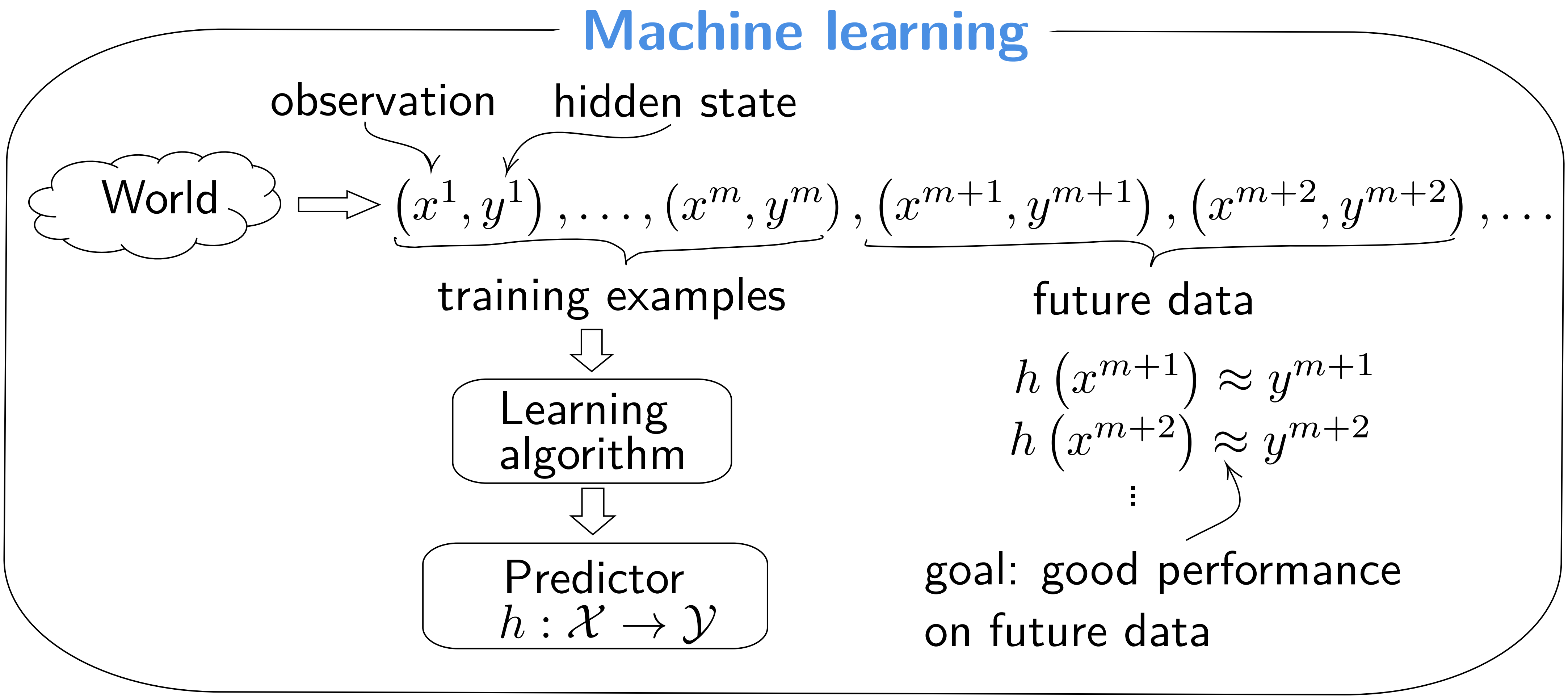
Statistical Machine Learning (BE4M33SSU)
Lecture 2: Predictor evaluation

Czech Technical University in Prague
V. Franc

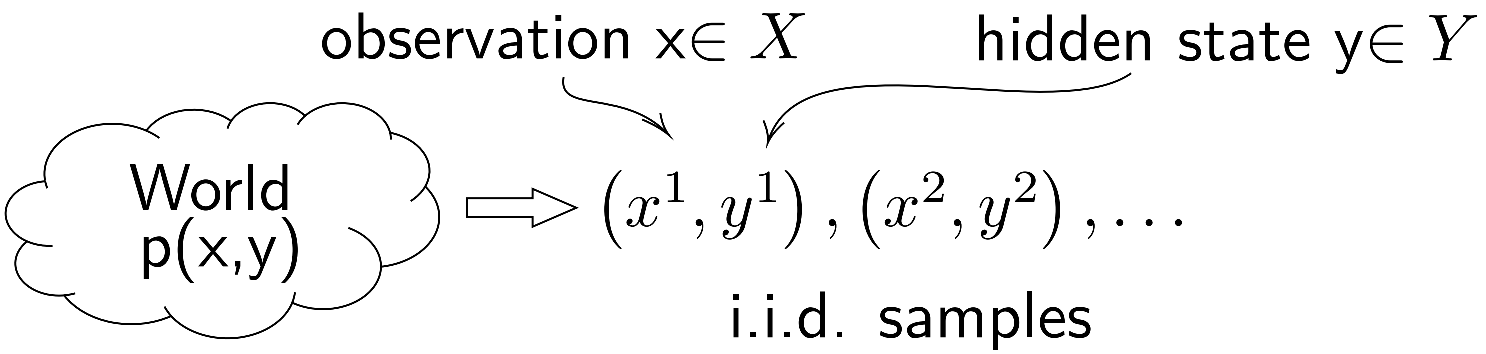
Content of the next three lectures: elements of machine learning theory



Machine learning theory: statistical framework which helps to clarify why and when the machine learning algorithms work.

Prediction problem and its optimal solution

- The main assumption: $(x, y) \in \mathcal{X} \times \mathcal{Y}$ are samples i.i.d. drawn from a random process with a distributon $p(x, y)$.



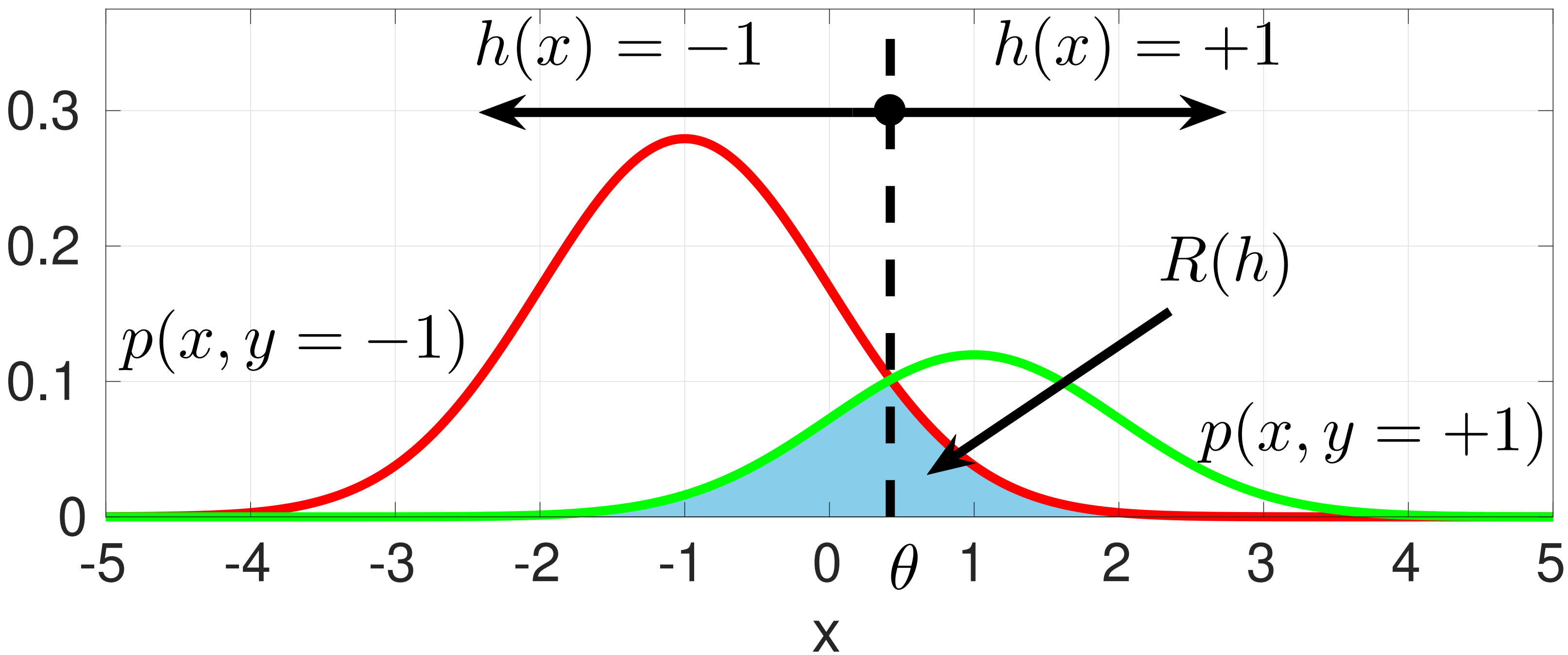
- We want to find a predictor (strategy, hypothesis, classifier) $h \colon \mathcal{X} \to \mathcal{Y}$
- Single prediction evaluated by loss function $\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
- The performance of h is evaluated by generalization error (expected risk)

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

- The optimal (Bayes) predictor: $h^* \in \min_{h \in \mathcal{Y}^{\mathcal{X}}} R(h)$

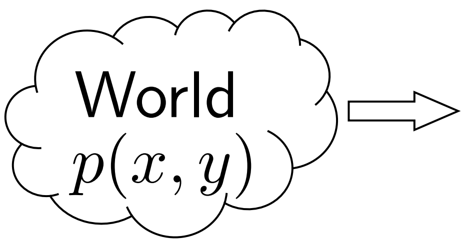
Example of a prediction problem

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



Machine learning: Learning and evaluation based on data

- **Setup:** we have only samples i.i.d drawn from an unknown $p(x, y)$.



$$\underbrace{(x^1, y^1), \dots, (x^m, y^m)}_{\text{training set}}, \underbrace{(x^{m+1}, y^{m+1}), \dots, (x^{m+l}, y^{m+l})}_{\text{test set}}, \dots$$

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

$$\mathcal{T}^m = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, m) \quad \text{drawn i.i.d. from} \quad p(x, y)$$

- **Evaluation**: estimate generalization error $R(h)$ of a given predictor $h\colon \mathcal{X} \to \mathcal{Y}$ using test (sequence) set

$$\mathcal{S}^l = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \dots, l) \quad \text{drawn i.i.d. from} \quad p(x, y)$$

Evaluation: estimation of the generalization error

- Given a predictor $h\colon \mathcal{X} \to \mathcal{Y}$ and a test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}[\ell(y, h(x))]$ by the test error

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

- Is the test error $R_{\mathcal{S}^l}(h)$ a good estimate of $R(h)$?
  - $R_{\mathcal{S}^l}(h)$ is a random number with an unknown distribution.
  - $R_{\mathcal{S}^l}(h)$ is an unbiased estimate of $R(h)$.

- **Problem:** With only knowledge of $\mathcal{S}^l$, can we confidently assess the difference between $R_{\mathcal{S}^l}(h)$ and $R(h)$?

Law of large numbers

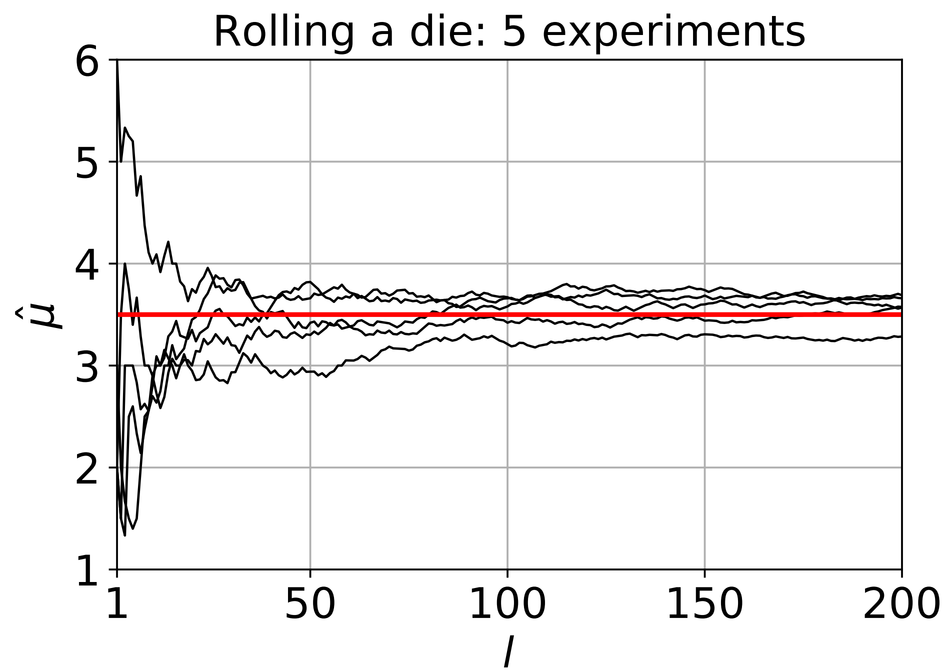
- Sample mean (arithmetic average) of the results of random trials gets closer to the expected value as more trials are performed.

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

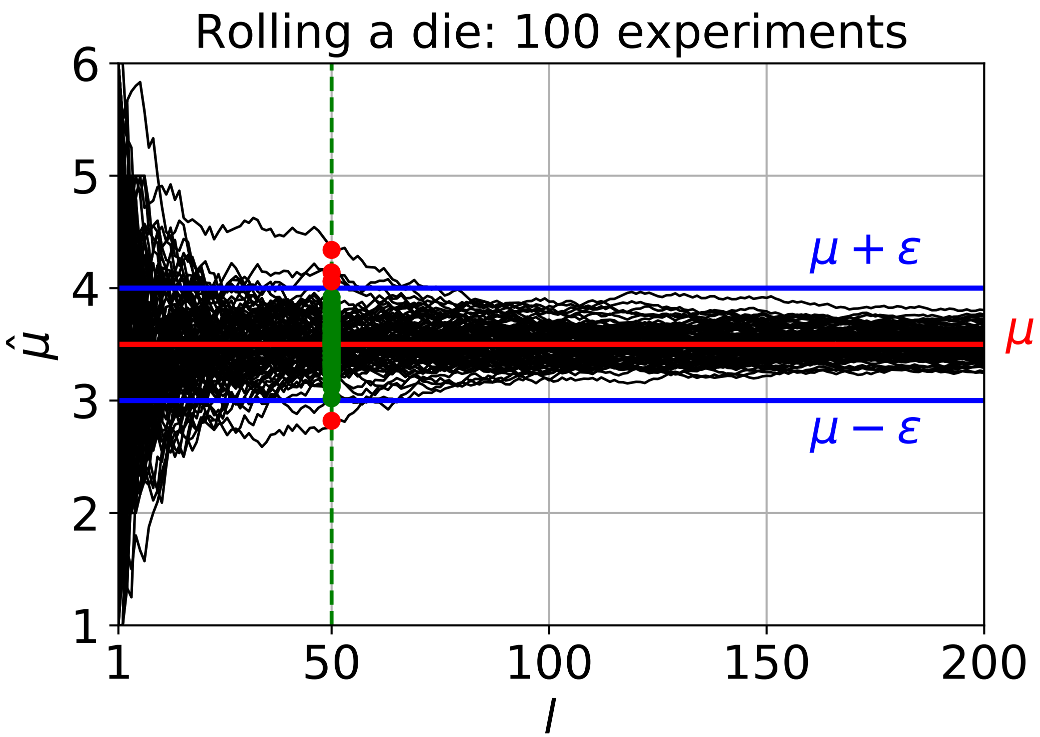
$$\mu = \mathbb{E}_{z\sim p}(z) = \sum_{z=1}^{6} z\, p(z) = \frac{1+2+3+4+5+6}{6} = 3.5$$

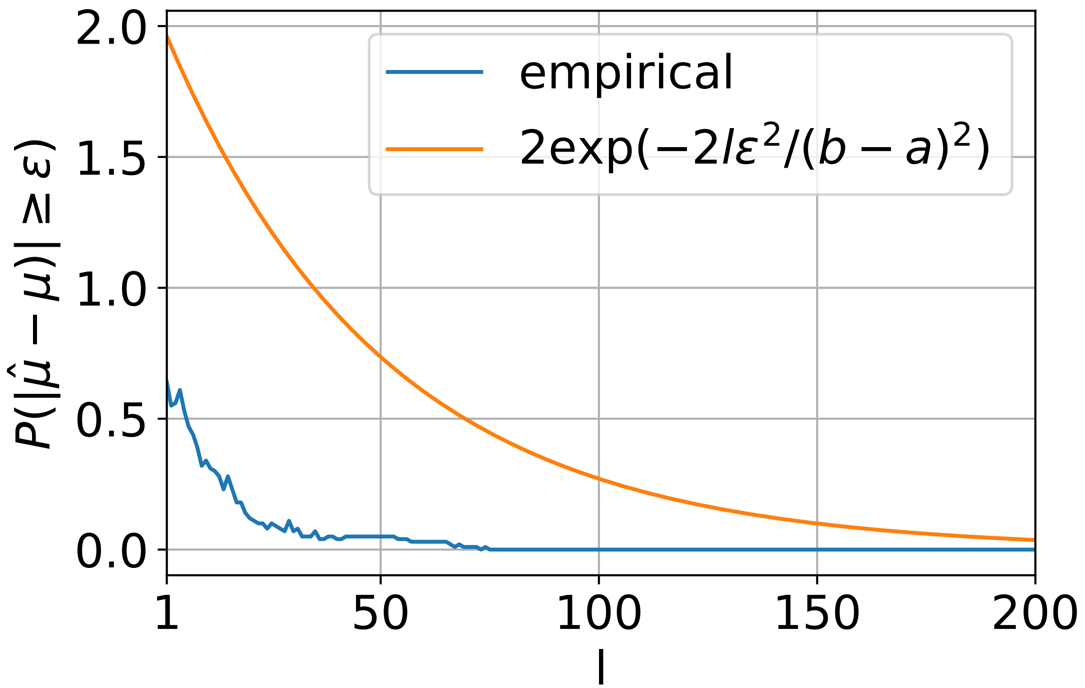
$$\hat{\mu}_l = \frac{1}{l}\sum_{i=1}^{l} z^i$$

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



Given a finite sample size l, how effectively does the sample mean $\hat{\mu}_l$ estime the expected value μ?





sample size $l = 50$, deviation $\varepsilon = 0.5$

Hoeffding inequality

$$\frac{\#\left(|\hat{\mu}_l - \mu| \geq \varepsilon\right)}{\#\text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P}\Big(|\hat{\mu}_l - \mu| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(b-a)^2}}$$

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

Hoeffding inequality

Theorem: Let $(z^1, \ldots, z^l)$ be a sample from independent r.v. from $[a, b]$ with expected value μ. Let $\hat{\mu}_l = \frac{1}{l}\sum_{i=1}^{l} z^i$. Then for any $\varepsilon > 0$ it holds that

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

Properties:

- (-) Conservative: the bound may not be tight.
- (+) General: the bound holds for any distribution.
- (+) Cheap: The bound is simple and easy to compute.

Confidence interval for the generalization error

- Given $h\colon \mathcal{X} \to \mathcal{Y}$ and test set $\mathcal{S}^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y)\sim p}[\ell(y, h(x))]$ by test error $R_{\mathcal{S}^l}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell(y^i, h(x^i))$.
- We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\Big(|R_{\mathcal{S}^l}(h) - R(h)| \geq \varepsilon\Big) \leq 2e^{-\frac{2l\,\varepsilon^2}{(\ell_{\max}-\ell_{\min})^2}} \qquad \forall\,\varepsilon > 0$$

- We use Hoeffding inequality to construct the confidence interval:

$$R(h) \in \big(R_{\mathcal{S}^l}(h) - \varepsilon, R_{\mathcal{S}^l}(h) + \varepsilon\big) \quad \text{holds with prob.} \quad 1 - \delta \quad \text{at least.}$$

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

Summary: We have derived a procedure to confidently assess the difference between $R_{\mathcal{S}^l}(h)$ and $R(h)$ knowing only the test examples $\mathcal{S}^l$.

Summary

- Formulation of the prediction problem.
- Evaluation vs learning.
- Law of Large numbers.
- Hoeffding inequality.
- Confidence intervals to estimate the generalization error.