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

**Machine learning**

World \[\rightarrow\] \((x^1, y^1), \ldots, (x^m, y^m), (x^{m+1}, y^{m+1}), (x^{m+2}, y^{m+2}), \ldots\)

training examples

Learning algorithm

Predictor \(h : \mathcal{X} \rightarrow \mathcal{Y}\)

future data

\[
\begin{align*}
h(x^{m+1}) &\approx y^{m+1} \\
h(x^{m+2}) &\approx y^{m+2} \\
\vdots
\end{align*}
\]

goal: good performance on future data
Content of the next three lectures:
 elements of machine learning theory

**Machine learning**

- **Observation**
- **Hidden State**
- **Training Examples**
- **Future Data**

- **Learning Algorithm**
- **Predictor** \( h : \mathcal{X} \rightarrow \mathcal{Y} \)

**Machine learning theory**: statistical framework which helps to clarify why and when the machine learning algorithms work.
The main assumption: \((x, y) \in X \times Y\) are samples i.i.d. drawn from a random process with a distribution \(p(x, y)\).

- observation \(x \in X\)
- hidden state \(y \in Y\)

\[\text{World } p(x, y) \rightarrow (x^1, y^1), (x^2, y^2), \ldots\]

i.i.d. samples
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 distribution \(p(x, y)\).

  - observation \(x \in \mathcal{X}\)
  - hidden state \(y \in \mathcal{Y}\)

  - \((x^1, y^1), (x^2, y^2), \ldots\)
  - i.i.d. samples

- We want to find a predictor (strategy, hypothesis, classifier) \(h: \mathcal{X} \rightarrow \mathcal{Y}\)
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 distribution \(p(x, y)\).

  - Observation \(x \in X\) hidden state \(y \in Y\)

- **World** \(p(x, y)\) \(\rightarrow (x^1, y^1), (x^2, y^2), \ldots\)

  - i.i.d. samples

- **We want to find a predictor** (strategy, hypothesis, classifier) \(h : \mathcal{X} \rightarrow \mathcal{Y}\)

- **Single prediction evaluated by loss function** \(\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\)
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 distribution \(p(x, y)\).

  observation \(x \in \mathcal{X}\)  
  hidden state \(y \in \mathcal{Y}\)

  \(p(x, y)\)

  \(\xrightarrow{\text{i.i.d. samples}} (x^1, y^1), (x^2, y^2), \ldots\)

- We want to find a **predictor** (strategy, hypothesis, classifier) \(h: \mathcal{X} \rightarrow \mathcal{Y}\).

- Single prediction evaluated by **loss function** \(\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}\).

- The performance of \(h\) is evaluated by **generalization error** (expected risk)

\[
R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) \, p(x, y) \, dx = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]
\]
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 distribution \(p(x, y)\).

  - observation \(x \in X\) hidden state \(y \in Y\)

  World \(p(x, y)\) \(\mapsto (x^1, y^1), (x^2, y^2), \ldots\) i.i.d. samples

- We want to find a predictor (strategy, hypothesis, classifier) \(h: \mathcal{X} \to \mathcal{Y}\)

- Single prediction evaluated by loss function \(\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}\)

- The performance of \(h\) is evaluated by generalization error (expected risk)

  \[ R(h) = \int \sum_{y \in \mathcal{Y}} \ell(y, h(x)) \ p(x, y) \ dx = \mathbb{E}_{(x,y) \sim p} \left[ \ell(y, h(x)) \right] \]

- 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:

  \[ X = \mathbb{R}, \quad \mathcal{Y} = \{+1, -1\}, \quad \ell(y, y') = \begin{cases} 
  0 & \text{if } y = y' \\
  1 & \text{if } y \neq y'
  \end{cases} \]

  \[ p(x, y) = p(y) \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \quad y \in \mathcal{Y}. \]
Example of a prediction problem

- The statistical model is known:
  
  \[ \mathcal{X} = \mathbb{R}, \quad \mathcal{Y} = \{+1, -1\}, \quad \ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases} \]

  \[ p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu y)^2}, \quad y \in \mathcal{Y}. \]

- The optimal strategy (assuming \( \mu_- < \mu_+ \)):

  \[ h(x) = \arg\max_{y \in \mathcal{Y}} p(y \mid x) = \text{sign}(x - \theta) \]
Example of a prediction problem

- The statistical model is known:

  - \( \mathcal{X} = \mathbb{R}, \; \mathcal{Y} = \{+1, -1\}, \; \ell(y, y') = \begin{cases} 0 & \text{if } y = y' \\ 1 & \text{if } y \neq y' \end{cases} \)

  - \( p(x, y) = p(y) \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \; y \in \mathcal{Y} \).

- The optimal strategy (assuming \( \mu_- < \mu_+ \)):

  \[ h(x) = \arg\max_{y \in \mathcal{Y}} p(y \mid x) = \text{sign}(x - \theta) \]

- The value of the true risk:

  \[ R(h) = \int_{-\infty}^{\theta} p(x, +1)dx + \int_{\theta}^{\infty} p(x, -1)dx \]
Example of a prediction problem

- The statistical model is known:
  
  - \( \mathcal{X} = \mathbb{R} \), \( \mathcal{Y} = \{+1, -1\} \), \( \ell(y, y') = \begin{cases} 
    0 & \text{if } y = y' \\
    1 & \text{if } y \neq y' 
  \end{cases} \)
  
  - \( p(x, y) = p(y) \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu_y)^2}, \ y \in \mathcal{Y} \).
**Setup:** we have only samples i.i.d drawn from an unknown $p(x, y)$.

![Diagram]
Machine learning: Learning and evaluation based on data

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

  ![Diagram](https://via.placeholder.com/150)

  World $p(x, y)$ $\mapsto (x^1, y^1), \ldots, (x^m, y^m), (x^{m+1}, y^{m+1}), \ldots, (x^{m+l}, y^{m+l}), \ldots$

  - training set
  - test set

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

  $$\mathcal{T}^m = ((x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m) \text{ drawn i.i.d. from } p(x, y)$$
Machine learning: Learning and evaluation based on data

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

  ![Diagram](image)

  $\text{World } p(x, y) \rightarrow \{(x^1, y^1), \ldots, (x^m, y^m), (x^{m+1}, y^{m+1}), \ldots, (x^{m+l}, y^{m+l})\}, \ldots$

  - **training set**
  - **test set**

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

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

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

  $$\mathcal{S}^l = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, l\} \text{ drawn i.i.d. from } p(x, y)$$
Evaluation: estimation of the generalization error

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

\[
R_{S^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))
\]
Evaluation: estimation of the generalization error

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

$$R_{Sl}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$
Evaluation: estimation of the generalization error

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

$$R_{S^l}(h) = \frac{1}{l} (\ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$
Evaluation: estimation of the generalization error

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

$$R_{S^l}(h) = \frac{1}{l} \left( \ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) \right) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$
Evaluation: estimation of the generalization error

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

$$R_{S^l}(h) = \frac{1}{l} \left( \ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) \right) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$
Evaluation: estimation of the generalization error

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

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

Is the test error $R_{Sl}(h)$ a good estimate of $R(h)$?

- $R_{Sl}(h)$ is a random number with an unknown distribution.
- $R_{Sl}(h)$ is an unbiased estimate of $R(h)$. 
Evaluation: estimation of the generalization error

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

$$R_{Sl}(h) = \frac{1}{l} \left( \ell(y^1, h(x^1)) + \cdots + \ell(y^l, h(x^l)) \right) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$$

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

**Problem:** With only knowledge of $S^l$, can we confidently assess the difference between $R_{Sl}(h)$ 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.
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

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

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

\[
\begin{align*}
z^1 &= 3 \\
z^2 &= 1 \\
z^3 &= 5 \\
z^l &= 2
\end{align*}
\]
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

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

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

$$
z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad z^l = 2
$$
**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

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

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

\[
z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad z^l = 2
\]

[Diagram showing rolling a die: 2 experiments]
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

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

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

\[ z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \cdots \quad z^l = 2 \]
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

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

\[
\hat{\mu}_l = \frac{1}{l} \sum_{i=1}^{l} z^i
\]
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?

sample size $l = 50$, deviation $\varepsilon = 0.5$
Given a finite sample size \( l \), how effectively does the sample mean \( \hat{\mu}_l \) estimate the expected value \( \mu \)?

Rolling a die: 5 experiments

sample size \( l = 50 \), deviation \( \varepsilon = 0.5 \)

\[
\frac{\#(|\hat{\mu}_l - \mu| \geq \varepsilon)}{\#\text{experiments}} = \frac{1}{5} = 0.2
\]
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?

Rolling a die: 100 experiments

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

$$\frac{\#(|\hat{\mu}_l - \mu| \geq \varepsilon)}{\#\text{experiments}} = \frac{5}{100} = 0.05$$
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?

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

\[
\frac{\# \left( |\hat{\mu}_l - \mu| \geq \varepsilon \right)}{\# \text{experiments}} = \frac{5}{100} = 0.05 \quad \rightarrow \quad \mathbb{P} \left( |\hat{\mu}_l - \mu| \geq \varepsilon \right)
\]
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?

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

Hoeffding inequality

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

$a = 1, b = 6$
Given a finite sample size $l$, how effectively does the sample mean $\hat{\mu}_l$ estimate the expected value $\mu$?

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

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

$$a = 1, \quad b = 6$$
Hoeffding inequality

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

\[
P\left( |\hat{\mu}_l - \mu| \geq \varepsilon \right) \leq 2e^{-\frac{2l \varepsilon^2}{(b-a)^2}}
\]
Hoeffding inequality

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

\[
P\left(|\hat{\mu}_l - \mu| \geq \varepsilon\right) \leq 2e^{-\frac{2l \varepsilon^2}{(b-a)^2}}
\]

**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 $h: \mathcal{X} \to \mathcal{Y}$ and test set $S^l \sim p^l$, estimate the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$ by test error $R_{S^l}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(x^i))$. 
Confidence interval for the generalization error

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

We set \( z^i = \ell(y^i, h(x^i)) \) and apply the Hoeffding inequality:

\[
\mathbb{P} \left( |R_{S^l}(h) - R(h)| \geq \varepsilon \right) \leq 2e^{-\frac{2l \varepsilon^2}{(\ell_{\text{max}}-\ell_{\text{min}})^2}} \quad \forall \varepsilon > 0
\]
Confidence interval for the generalization error

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

- We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$\mathbb{P}\left( |R_{S^l}(h) - R(h)| \geq \varepsilon \right) \leq 2e^{-\frac{2\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}} \quad \forall \varepsilon > 0$$

- We use Hoeffding inequality to construct the confidence interval:

$$R(h) \in (R_{S^l}(h) - \varepsilon, R_{S^l}(h) + \varepsilon)$$ holds with prob. $1 - \delta$ at least.
Confidence interval for the generalization error

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

We set \( z^i = \ell(y^i, h(x^i)) \) and apply the Hoeffding inequality:

\[
\mathbb{P}\left( |R_{S^l}(h) - R(h)| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}} \quad \forall \varepsilon > 0
\]

We use Hoeffding inequality to construct the confidence interval:

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

For fixed \( l \) and \( \delta \in [0, 1] \), compute \( \varepsilon = (\ell_{\text{max}} - \ell_{\text{min}}) \sqrt{\frac{\log(2) - \log(\delta)}{2l}} \)
Confidence interval for the generalization error

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

- We set $z^i = \ell(y^i, h(x^i))$ and apply the Hoeffding inequality:

$$
\mathbb{P}\left( |R_{S^l}(h) - R(h)| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}} \quad \forall \varepsilon > 0
$$

- We use Hoeffding inequality to construct the confidence interval:

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

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

- For fixed $\varepsilon$ and $\delta \in [0, 1]$, compute $l = \frac{\log(2) - \log(\delta)}{2\varepsilon^2} (\ell_{\text{max}} - \ell_{\text{min}})^2$
Confidence interval for the generalization error

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

\[
\mathbb{P}\left( |R_{S^l}(h) - R(h)| \geq \varepsilon \right) \leq 2e^{-2l\varepsilon^2/(\ell_{\text{max}} - \ell_{\text{min}})^2} \quad \forall \varepsilon > 0
\]

- We use Hoeffding inequality to construct the confidence interval:

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

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

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

- Formulation of the prediction problem.
- Evaluation vs learning.
- Law of Large numbers.
- Hoeffding inequality.
- Confidence intervals to estimate the generalization error.
$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \ldots \quad z^l = 2$
$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad z^l = 2$
Rolling a die
$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \cdots \quad z^l = 2$
Rolling a die: 2 experiments
\[ z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad z^l = 2 \]
Rolling a die: 3 experiments
$z^1 = 3 \quad z^2 = 1 \quad z^3 = 5 \quad \ldots \quad z^l = 2$
Rolling a die: 5 experiments
Rolling a die: 5 experiments
Rolling a die: 5 experiments

\[ \hat{\mu} \]

\[ \mu + \varepsilon \]

\[ \mu - \varepsilon \]

\[ \mu \]
Rolling a die: 5 experiments

\[ \mu \pm \varepsilon \]

\[ \mu \]

\[ \mu - \varepsilon \]
Rolling a die: 100 experiments
Rolling a die: 100 experiments
Rolling a die: 100 experiments
\[ P(\left| \hat{\mu} - \mu \right| > \varepsilon) \leq 2 \exp\left( -2 \varepsilon^2 / (b - a)^2 \right) \]

**empirical**