Statistical Machine Learning (BE4M33SSU)
Lecture 3: Empirical Risk Minimization

Czech Technical University in Prague
V. Franc

BE4M33SSU – Statistical Machine Learning, Winter 2022
Learning

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

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

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

- **Hypothesis class (space):**

$$\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{h: \mathcal{X} \rightarrow \mathcal{Y}\}$$

- **Learning algorithm:** a function

$$A: \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \rightarrow \mathcal{H}$$

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 $\mathcal{T}^m$,

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

- The ERM based algorithm returns $h_m$ such that

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

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

$\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R} \}$, $\ell(y, y') = [y \neq y']$

![Graph showing the relationship between $x$ and $m = 10$]
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 $\mathcal{T}^m$,

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

$$\mathcal{H} = \{h(x) = \text{sign}(x - \theta) \mid \theta \in \mathbb{R}\}, \ell(y, y') = [y \neq y']$$
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 $\mathcal{T}^m$,

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

- The ERM based algorithm returns $h_m$ such that

$$h_m \in \text{Argmin } R_{\mathcal{T}^m}(h)$$

- Depending on the choice of $\mathcal{H}$ and $\ell$ 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, ...
Example of ERM failure

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 $p(y = +1) = 0.8$. 
Example of ERM failure

- 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 $R^* = 0.2$. 
Example of ERM failure

- 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 memorizing strategy

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

- 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 memorizing 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 $m$. 
Wrap up of the previous lecture

- We use the empirical risk $R_{Sl}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(y^i))$ as a proxy of the true risk $R(h) = \mathbb{E}_{x, y \sim p}[\ell(y, h(x))]$. 
Wrap up of the previous lecture

- We use the empirical risk $R_{Sl}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(y^i))$ as a proxy of the true risk $R(h) = \mathbb{E}_{x,y \sim p}[\ell(y, h(x))]$.

- In case of evaluation, $h$ is fixed and due to the law of large numbers, $R_{Sl}(h)$ gets close to $R(h)$ if we have enough examples:

$$
\mathbb{P}\left( \left| R_{Sl}(h) - R(h) \right| \geq \varepsilon \right) \leq 2e^{-\frac{2l\varepsilon^2}{(\ell_{max}-\ell_{min})^2}}
$$
Wrap up of the previous lecture

- We use the empirical risk $R_{Sl}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(y^i))$ as a proxy of the true risk $R(h) = \mathbb{E}_{x,y \sim p}[\ell(y, h(x))]$.

- In case of evaluation, $h$ is fixed and due to the law of large numbers, $R_{Sl}(h)$ gets close to $R(h)$ if we have enough examples:

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

We say that $R_{Sl}(h)$ converges in probability to $R(h)$, i.e.

$$
\forall \varepsilon > 0: \lim_{l \to \infty} \mathbb{P}\left( |R_{Sl}(h) - R(h)| \geq \varepsilon \right) = 0
$$
Wrap up of the previous lecture

- We use the empirical risk $R_{Sl}(h) = \frac{1}{l} \sum_{i=1}^{l} \ell(y^i, h(y^i))$ as a proxy of the true risk $R(h) = \mathbb{E}_{x,y \sim p}[\ell(y, h(x))]$.

- In case of evaluation, $h$ is fixed and due to the law of large numbers, $R_{Sl}(h)$ gets close to $R(h)$ if we have enough examples:

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

  We say that $R_{Sl}(h)$ converges in probability to $R(h)$, i.e.

  $$\forall \varepsilon > 0: \lim_{l \to \infty} \mathbb{P}\left( |R_{Sl}(h) - R(h)| \geq \varepsilon \right) = 0$$

- In case of learning, $h_m = A(T_m)$ is learned from $T^m$ then $R_{T^m}(h)$ does not have to get close to $R(h)$ even if we have enough examples:

  $$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( |R_{T^m}(h_m) - R(h_m)| \geq \varepsilon \right) \neq 0$$
Why law of large numbers does not apply for learning?

- Hoeffding inequality $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{\frac{-2m\varepsilon^2}{(b-a)^2}}$, $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i$, requires $\{z^1, \ldots, z^m\}$ to be sample from i.i.d. rv. with expected value $\mu$. 
Why law of large numbers does not apply for learning?

- Hoeffding inequality $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$, $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i$, requires $\{z^1, \ldots, z^m\}$ to be sample from i.i.d. rv. with expected value $\mu$.

- $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ is drawn from i.i.d. rv. with $p(x, y)$. 
Why law of large numbers does not apply for learning?

- Hoeffding inequality \( \mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}} \), \( \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i \), requires \( \{z^1, \ldots, z^m\} \) to be sample from i.i.d. rv. with expected value \( \mu \).

- \( \mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \) is drawn from i.i.d. rv. with \( p(x, y) \).

Evaluation:

- \( h \) fixed independently on \( \mathcal{T}^m \), \( z^i = \ell(y^i, h(x^i)) \) and \( \{z^1, \ldots, z^m\} \) is i.i.d.

- Therefore \( \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) = 0 \)
Why law of large numbers does not apply for learning?

- Hoeffding inequality $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$, $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i$, requires $\{z^1, \ldots, z^m\}$ to be sample from i.i.d. rv. with expeted value $\mu$.

- $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\}$ is drawn from i.i.d. rv. with $p(x, y)$.

Evaluation:

- $h$ fixed independently on $\mathcal{T}^m$, $z^i = \ell(y^i, h(x^i))$ and $\{z^1, \ldots, z^m\}$ is i.i.d.

- Therefore $\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) = 0$

Learning:

- $h_m = A(\mathcal{T}^m)$, $z^i = \ell(y^i, h_m(x^i))$ and thus $\{z^1, \ldots, z^m\}$ is not i.i.d.

- No guarantee that $\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{\mathcal{T}^m}(h_m) - R(h_m)| \geq \varepsilon) = 0$
Why law of large numbers does not apply for learning?

- **Hoeffding inequality** \( \mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}} \), \( \hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i \), requires \( \{z^1, \ldots, z^m\} \) to be sample from i.i.d. rv. with expected value \( \mu \).

- \( \mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \) is drawn from i.i.d. rv. with \( p(x, y) \).

**Evaluation:**

- \( h \) fixed independently on \( \mathcal{T}^m \), \( z^i = \ell(y^i, h(x^i)) \) and \( \{z^1, \ldots, z^m\} \) is i.i.d.
- Therefore \( \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) = 0 \)

**Learning:**

- \( h_m = A(\mathcal{T}^m), z^i = \ell(y^i, h_m(x^i)) \) and thus \( \{z^1, \ldots, z^m\} \) is not i.i.d.
- No guarantee that \( \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{\mathcal{T}^m}(h_m) - R(h_m)| \geq \varepsilon) = 0 \)
- The task for the rest of the lecture is to show how to fix it.
To fix the problem we need uniform law of large numbers

$$\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \quad \ell(y, y') = [y \neq y']$$
To fix the problem we need uniform law of large numbers

\[ \mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \quad \ell(y, y') = [y \neq y'] \]
To fix the problem we need uniform law of large numbers

\[ \mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \ell(y, y') = [y \neq y'] \]
To fix the problem we need uniform law of large numbers

For learning we need: the empirical risk $R_{T^m}(h_m)$ of the learned strategy $h_m = A(T^m)$ converges to the true risk $R(h_m)$:

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \left| R(h_m) - R_{T^m}(h_m) \right| \geq \varepsilon \right) = 0$$

$$H = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \ell(y, y') = [y \neq y']$$
To fix the problem we need uniform law of large numbers

\[ \mathbb{P}\left(\left|R(h_m) - R_{T^m}(h_m)\right| \geq \varepsilon \right) \leq \mathbb{P}\left(\begin{array}{l}
\left|R(h_1) - R_{T^m}(h_1)\right| \geq \varepsilon \\
\left|R(h_2) - R_{T^m}(h_2)\right| \geq \varepsilon \\
\vdots \\
\left|R(h_{|\mathcal{H}|}) - R_{T^m}(h_{|\mathcal{H}|})\right| \geq \varepsilon
\end{array}\right) \]

\[ \mathcal{H} = \{h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R}\}, \ \ell(y, y') = [y \neq y'] \]
To fix the problem we need uniform law of large numbers

\[
\mathbb{P}\left( \left| R(h_m) - R_{Tm}(h_m) \right| \geq \varepsilon \right) \leq \mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{Tm}(h) \right| \geq \varepsilon \right)
\]

\(\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \ell(y, y') = [y \neq y']\)
To fix the problem we need uniform law of large numbers

\[
\mathbb{P}\left(\left|R(h_m) - R_{T^m}(h_m)\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} |R(h) - R_{T^m}(h)| \geq \varepsilon\right) \leq B(m, \mathcal{H}, \varepsilon)
\]

\[\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \quad \ell(y, y') = [y \neq y']\]
To fix the problem we need uniform law of large numbers

\[
\mathbb{P}\left( \left| R(h_m) - R_{T^m}(h_m) \right| \geq \varepsilon \right) \leq \mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right| \geq \varepsilon \right) \leq B(m, \mathcal{H}, \varepsilon)
\]

\[\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \ \ell(y, y') = [y \neq y']\]
To fix the problem we need uniform law of large numbers

$$\mathbb{P}\left( |R(h_m) - R_{T^m}(h_m)| \geq \varepsilon \right) \leq \mathbb{P}\left( \sup_{h \in \mathcal{H}} |R(h) - R_{T^m}(h)| \geq \varepsilon \right) \leq B(m, \mathcal{H}, \varepsilon)$$

$$\mathcal{H} = \{ h(x) = \text{sign}(x - \theta) | \theta \in \mathbb{R} \}, \ell(y, y') = [y \neq y']$$

![Graph illustrating the function $R(h)$ and $R_{T^m}(h)$ for $m=1000$]
Uniform Law of Large Numbers

Law of Large Numbers: for any $p(x, y)$ generating $\mathcal{T}^m$, and $h \in \mathcal{H}$ fixed without using $\mathcal{T}^m$ we have

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left(\left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) = 0$$

empirical risk fails for $h$
Uniform Law of Large Numbers

- **Law of Large Numbers:** for any \( p(x, y) \) generating \( \mathcal{T}^m \), and \( h \in \mathcal{H} \) fixed without using \( \mathcal{T}^m \) we have

\[
\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) = 0
\]

- **Uniform Law of Large Numbers:** if for any \( p(x, y) \) generating \( \mathcal{T}^m \) it holds that

\[
\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \left| R(h_1) - R_{\mathcal{T}^m}(h_1) \right| \geq \varepsilon \text{ or } \left| R(h_2) - R_{\mathcal{T}^m}(h_2) \right| \geq \varepsilon \text{ or } \ldots \right) = 0
\]

we say that ULLN applies for \( \mathcal{H} \).
Uniform Law of Large Numbers

- **Law of Large Numbers**: for any $p(x, y)$ generating $T^m$, and $h \in \mathcal{H}$ fixed without using $T^m$ we have

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \left| R(h) - R_{T^m}(h) \right| \geq \varepsilon \right) = 0$$

empirical risk fails for $h$

- **Uniform Law of Large Numbers**: if for any $p(x, y)$ generating $T^m$ it holds that

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right| \geq \varepsilon \right) = 0$$

empirical risk fails for some $h \in \mathcal{H}$

we say that ULLN applies for $\mathcal{H}$.
Uniform Law of Large Numbers

- **Law of Large Numbers**: for any $p(x, y)$ generating $\mathcal{T}^m$, and $h \in \mathcal{H}$ fixed without using $\mathcal{T}^m$ we have

  \[ \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) = 0 \]

  empirical risk fails for $h$

- **Uniform Law of Large Numbers**: if for any $p(x, y)$ generating $\mathcal{T}^m$ it holds that

  \[ \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) = 0 \]

  empirical risk fails for some $h \in \mathcal{H}$

  we say that ULLN applies for $\mathcal{H}$.

- Alternatively we say: the empirical risk converges uniformly to the true risk, or that the hypothesis class $\mathcal{H}$ has the uniform convergence property.
ULLN applies for finite hypothesis class

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

$$B(h) = \left\{ T^m \in (\mathcal{X} \times \mathcal{Y})^m \left| R_{T^m}(h) - R(h) \geq \varepsilon \right. \right\}$$
ULLN applies for finite hypothesis class

- Assume a finite hypothesis class \( \mathcal{H} = \{h_1, \ldots, h_K\} \).
- Define the set of all “bad” training sets for a strategy \( h \in \mathcal{H} \) as

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

\[
P\left(|R_{\mathcal{T}^m}(h_1) - R(h_1)| \geq \varepsilon \right) = P\left(\mathcal{T}^m \in \mathcal{B}(h_1) \right) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}
\]
**ULLN applies for finite hypothesis class**

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

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

Three strategies

Events $\mathcal{T}^m \in \mathcal{B}(h), h \in \mathcal{H}$ mutually exclusive

$$
\mathbb{P}\left( \max_{h \in \{h_1, h_2, h_3\}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \right) = 
\mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_1) \text{ or } \mathcal{T}^m \in \mathcal{B}(h_2) \text{ or } \mathcal{T}^m \in \mathcal{B}(h_3)) = 
\mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_1)) + \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_2)) + \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_3))
$$
ULLN applies for finite hypothesis class

- Assume a finite hypothesis class \( \mathcal{H} = \{h_1, \ldots, h_K\} \).
- Define the set of all “bad” training sets for a strategy \( h \in \mathcal{H} \) as

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

\[
\mathbb{P}\left( \max_{h \in \{h_1, h_2, h_3\}} |R_{T^m}(h) - R(h)| \geq \varepsilon \right) = \mathbb{P}\left( T^m \in \mathcal{B}(h_1) \text{ or } T^m \in \mathcal{B}(h_2) \text{ or } T^m \in \mathcal{B}(h_3) \right) \leq \mathbb{P}(T^m \in \mathcal{B}(h_1)) + \mathbb{P}(T^m \in \mathcal{B}(h_2)) + \mathbb{P}(T^m \in \mathcal{B}(h_3))
\]
ULLN applies for finite hypothesis class

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

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

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

$$
\mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h)) = 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{(b-a)^2}}
$$
ULLN applies for finite hypothesis class

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

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

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

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

- ULLN applies for finite hypothesis class

$$\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \max_{h \in \mathcal{H}} \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \right) = 0$$
Generalization bound for finite hypothesis class

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

$$
P\left( \max_{h \in \mathcal{H}} |R_{m}(h) - R(h)| \geq \varepsilon \right) \leq 2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$$
Generalization bound for finite hypothesis class

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

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

Find an upper bound $\varepsilon$ on the discrepancy between $R_{\mathcal{T}}m(h)$ and $R(h)$ which holds uniformly for all $h \in \mathcal{H}$ with probability $1 - \delta$ at least:

$$\mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}}m(h) - R(h)| < \varepsilon \right) = 1 - \mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}}m(h) - R(h)| \geq \varepsilon \right) \geq 1 - 2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(b-a)^2}} = 1 - \delta$$

and solving the last equality for $\varepsilon$ yields

$$\varepsilon = (b - a)\sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$$
Generalization bound for finite hypothesis class

**Theorem:** Let \( T^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m \) be draw from i.i.d. rv. with p.d.f. \( p(x, y) \) and let \( \mathcal{H} \) be a finite hypothesis class. Then, for any \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \) the inequality

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

holds for all \( h \in \mathcal{H} \) simultaneously and any loss function \( \ell: \mathcal{Y} \times \mathcal{Y} \to [a, b] \).
Generalization bound for finite hypothesis class

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

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

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

- Recommendations that follow from the generalization bound:
  1. Minimize the empirical risk.
  2. Use as much training examples as possible.
  3. Limit the size of the hypothesis space $|\mathcal{H}|$.

Note that 1) and 3) are conflicting recommendations.
Generalization bound for finite hypothesis class

**Theorem:** Let $T^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ be drawn from i.i.d. rv. with p.d.f. $p(x, y)$ and let $\mathcal{H}$ be a finite hypothesis class. Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$ the inequality

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

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

- **Recommendations that follow from the generalization bound:**
  1. Minimize the empirical risk.
  2. Use as much training examples as possible.
  3. Limit the size of the hypothesis space $|\mathcal{H}|$.

  Note that 1) and 3) are conflicting recommendations.

- **The generalization bound holds for any learning algorithm not just ERM.**
Structural Risk Minimization

- Learn \( h : \mathcal{X} \to \mathcal{Y} \) by minimizing the generalization bound

\[
R(h) \leq R_{\mathcal{T}}^m(h) + (b - a) \sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}} \epsilon(m,|\mathcal{H}|,\delta)
\]
Structural Risk Minimization

- Learn \( h : \mathcal{X} \rightarrow \mathcal{Y} \) by minimizing the generalization bound

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

- Design a nested sequence of hypothesis classes

\[
\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_K
\]
Structural Risk Minimization

Learn $h: \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing the generalization bound

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

Design a nested sequence of hypothesis classes

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_K$$

Minimize the generalization bound:

1. $h_i = \arg\min_{h \in \mathcal{H}_i} R_{T_m}(h)$, $\forall i \in \{1, \ldots, K\}$

2. $i^* = \arg\min_{i=1,\ldots,K} \left(R_{T_m}(h_i) + \epsilon(m, |\mathcal{H}_i|, \delta)\right)$

3. Output $h_{i^*}$
Structural Risk Minimization

- Learn $h: \mathcal{X} \rightarrow \mathcal{Y}$ by minimizing the generalization bound

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

$$\epsilon(m, |\mathcal{H}|, \delta)$$
\[ R(h_m) \]

\[ R_{\tau m}(h) \]

\[ m = 20 \]
$R(h) + R(h) + R(h_m)$

$m = 20$

$R(h_m)$

$R(h) + \epsilon$

$R(h) - \epsilon$

$R_{TM}(h_m)$
$B(h_1)$
$B(h_1)$

$B(h_2)$

$B(h_3)$
$\mathcal{B}(h_1)$

$\mathcal{B}(h_2)$

$\mathcal{B}(h_3)$