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

Czech Technical University in Prague
V. Franc

BE4M33SSU – Statistical Machine Learning, Winter 2021
Learning

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

\[
\mathcal{T}^m = \{(x^i, y^i) \in (\mathcal{X} \times \mathcal{Y}) \mid i = 1, \ldots, m\}
\]
drawn from i.i.d. according to unknown \( p(x, y) \).

- **Hypothesis class (space):**

\[
\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}} = \{ h: \mathcal{X} \to \mathcal{Y}\}
\]

- **Learning algorithm:** a function

\[
A: \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H}
\]

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 $\mathcal{T}^m$,

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

- The ERM based algorithm returns $h_m$ such that

$$h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$$  \hspace{1cm} (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 $\mathcal{T}^m$,

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

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

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

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

![Graph showing empirical risk and function $h_m(x)$]
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

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

- 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 $m$. 
Generalization error

ERM may fail when $R_{\mathcal{T}m}(h_m)$ is not a good proxy of $R(h_m)$, because $R_{\mathcal{T}m}(h)$ is used as a guidance to select $h_m$. 
Generalization error

ERM may fail when \( R_{\mathcal{T}m}(h_m) \) is not a good proxy of \( R(h_m) \), because \( R_{\mathcal{T}m}(h) \) is used as a guidance to select \( h_m \).

We need the generalization error, i.e., the discrepancy between \( R(h) \) and \( R_{\mathcal{T}m}(h) \), to become small when the number of examples \( m \) grows:

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

where \( h_m = A(\mathcal{T}_m) \) is learned by \( A: \bigcup_{m=1}^{\infty} (\mathcal{X} \times \mathcal{Y})^m \to \mathcal{H} \).
**Generalization error**

- ERM may fail when $R_T^m(h_m)$ is not a good proxy of $R(h_m)$, because $R_T^m(h)$ is used as a guidance to select $h_m$.

- We need the generalization error, i.e., the discrepancy between $R(h)$ and $R_T^m(h)$, to become small when the number of examples $m$ grows:

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

  where $h_m = A(T_m)$ is learned by $A: \bigcup_{m=1}^{\infty} (X \times Y)^m \to \mathcal{H}$.

Plan for this lecture:

- Conditions on $\mathcal{H}$ which guarantee that the generalization error converges to zero with growing number of examples $m$. 

What’s wrong with Hoeffding?

- 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 \( \mu \).
What’s wrong with Hoeffding?

- 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 $p(x, y)$. 
What’s wrong with Hoeffding?

- 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 \( \mu \).

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

Evaluation:

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

- Therefore \( \forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}(|R_{T^m}(h) - R(h)| \geq \varepsilon) = 0 \)
What’s wrong with Hoeffding?

- 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 $\forall \varepsilon > 0: \lim_{m\to\infty} \mathbb{P}(|R_{\mathcal{T}^m}(h_m) - R(h_m)| \geq \varepsilon) = 0$
Uniform Law of Large Numbers

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

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

high generalization error
Uniform Law of Large Numbers

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

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

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

\[
\forall \varepsilon > 0: \lim_{m \to \infty} \mathbb{P}\left( \begin{array}{l}
|R(h_1) - R_{T^m}(h_1)| \geq \varepsilon \\
|R(h_2) - R_{T^m}(h_2)| \geq \varepsilon \\
\vdots \\
|R(h_{|\mathcal{H}|}) - R_{T^m}(h_{|\mathcal{H}|})| \geq \varepsilon 
\end{array} \right) = 0
\]

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

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

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

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 seeing \( T^m \) we have

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

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

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

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

- **ULLN provides guarantees for all** \( h \in \mathcal{H} \) including \( h_m = A(T_m) \):

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

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

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

Single strategy

$$P\left(|R_{\mathcal{T}^m(h_1)} - R(h_1)| \geq \varepsilon \right) = P\left(\mathcal{T}^m \in \mathcal{B}(h_1) \right) \leq 2e^{-\frac{2m \varepsilon^2}{(b-a)^2}}$$
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

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

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

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

$$\mathbb{P}\left( \max_{h \in \{h_1, h_2, h_3\}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \right) =$$

$$\mathbb{P}\left( \mathcal{T}^m \in \mathcal{B}(h_1) \text{ or } \mathcal{T}^m \in \mathcal{B}(h_2) \text{ or } \mathcal{T}^m \in \mathcal{B}(h_3) \right) \leq$$

$$\mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_1)) + \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_2)) + \mathbb{P}(\mathcal{T}^m \in \mathcal{B}(h_3))$$

Three strategies
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

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

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

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

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

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

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

- ULLN applies for finite hypothesis class

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

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

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

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

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

- Find an upper bound $\varepsilon$ on the generalization error which holds uniformly for all $h \in \mathcal{H}$ with probability $1 - \delta$ at least:

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

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

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

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

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

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 bound:

- Minimize the empirical risk.
- More training examples the better.
- Select appropriate trade-off between $|\mathcal{H}|$ and $m$: 
Structural Risk Minimization

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

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

where \( \epsilon(m,|\mathcal{H}|,\delta) \) is the generalization bound term.
Structural Risk Minimization

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

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

- 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

$$R(h) \leq R_{\mathcal{T}m}(h) + (b - a) \sqrt{\log \frac{2|\mathcal{H}| + \log \frac{1}{\delta}}{2m} \epsilon(m, |\mathcal{H}|, \delta)}$$
\( \mathcal{B}(h_1) \)

\( \mathcal{B}(h_2) \)

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

$\mathcal{B}(h_2)$

$\mathcal{B}(h_3)$