Learning

**Goal:** Given a training set $T^m \sim p^m$, find a strategy $h: \mathcal{X} \rightarrow \mathcal{Y}$ with minimizing the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$. 
Learning

Goal: Given a training set $T_m \sim p^m$, find a strategy $h: \mathcal{X} \rightarrow \mathcal{Y}$ with minimizing the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p}[\ell(y, h(x))]$.

Hypothesis class (space): fixed before learning based on prior knowledge

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

- **Goal:** Given a training set $T^m \sim p^m$, find a strategy $h: \mathcal{X} \rightarrow \mathcal{Y}$ with minimizing the generalization error $R(h) = \mathbb{E}_{(x,y) \sim p} [\ell(y, h(x))]$.

- **Hypothesis class (space):** fixed before learning based on prior knowledge

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

- **Learning algorithm:** a function

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

returns a strategy $h_m = A(T^m)$ from $\mathcal{H}$ based on a training set $T^m$
Empirical Risk Minimization learning

- The generalization error $R(h)$ is approximated by the empirical risk $R_{T^m}(h)$ computed on the training examples $T^m \sim \mathcal{P}^m$:

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

- The ERM based learning algorithm returns $h_m$ such that

$$h_m \in \underset{h \in \mathcal{H}}{\text{Argmin}} R_{T^m}(h) \quad (1)$$
Empirical Risk Minimization learning

- The generalization error $R(h)$ is approximated by the empirical risk $R_{\mathcal{T}^m}(h)$ computed on the training examples $\mathcal{T}^m \sim p^m$:

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

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

- The generalization error $R(h)$ is approximated by the empirical risk $R_{\mathcal{T}_m}(h)$ computed on the training examples $\mathcal{T}_m \sim p^m$:

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

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

---

![Graph showing empirical risk and hypothesis set](image-url)
Empirical Risk Minimization learning

- The generalization error $R(h)$ is approximated by the empirical risk $R_{\mathcal{T}^m}(h)$ computed on the training examples $\mathcal{T}^m \sim p^m$:

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

- The ERM based learning algorithm returns $h_m$ such that

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

(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, ...
ERM can fail due to overfitting

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 $p(y = +1) = 0.8$. 
ERM can fail due to overfitting

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 $R^* = 0.2$. 
ERM can fail due to overfitting

- 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 $R^* = 0.2$.

- Learning algorithm “lookup table”: given training set $\mathcal{T}^m$ it returns

$$h_m(x) = \begin{cases} 
  y^j & \text{if } x = x^j \text{ for some } j \in \{1, \ldots, m\} \\
  -1 & \text{otherwise}
\end{cases}$$
ERM can fail due to overfitting

- 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 \( R^* = 0.2 \).

- Learning algorithm “lookup table”: given training set \( \mathcal{T}^m \) it returns

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

- Implies ERM principle as \( \mathbb{P}(R_{\mathcal{T}^m}(h_m) = 0) = 1 \).
- Fails to find a good solution as \( \mathbb{P}(R(h_m) = 0.8) = 1, \forall m \in \mathbb{N} \).
ERM can fail due to overfitting

- 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 $R^* = 0.2$.

- Learning algorithm “lookup table”: given training set $\mathcal{T}^m$ it returns

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

- Implements ERM principle as $\mathbb{P}(R_{\mathcal{T}^m}(h_m) = 0) = 1$.

- Fails to find a good solution as $\mathbb{P}(R(h_m) = 0.8) = 1$, $\forall m \in \mathbb{N}$.

- **Overfitting**: the case when $h_m = A(\mathcal{T}^m)$ and the training error $R_{\mathcal{T}^k}(h_m)$ is low while the generalization error $R(h_m)$ is high.
ERM can fail due to overfitting

- 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 $R^* = 0.2$.

- Learning algorithm “lookup table”: given training set $\mathcal{T}^m$ it returns

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

  - Implements ERM principle as $\mathbb{P}(R_{\mathcal{T}^m}(h_m) = 0) = 1$.
  - Fails to find a good solution as $\mathbb{P}(R(h_m) = 0.8) = 1$, $\forall m \in \mathbb{N}$.

- **Overfitting**: the case when $h_m = A(\mathcal{T}^m)$ and the training error $R_{\mathcal{T}^k}(h_m)$ is low while the generalization error $R(h_m)$ is high.

- **Problem**: under which conditions the overfitting can be eliminated?
Why the 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 independent random variables with the expected value \(\mu\).
Why the 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 independent random variables with the expected value \(\mu\).

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

- Hoeffding inequality $\Pr(|\hat{\mu} - \mu| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(b-a)^2}}$, $\hat{\mu} = \frac{1}{m} \sum_{i=1}^{m} z^i$, requires $(z^1, \ldots, z^m)$ to be sample from independent random variables with the 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.

- We can apply Hoeffding $\Pr(|R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}$
Why the 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 independent random variables with the 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.

- We can apply Hoeffding \(\mathbb{P}(|R_{T^m}(h) - R(h)| \geq \varepsilon) \leq 2e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}}-\ell_{\text{min}})^2}}\)

Learning:

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

- We cannot apply Hoeffding to bound \(\mathbb{P}(|R_{T^m}(h_m) - R(h_m)| \geq \varepsilon)\)
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$. 
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class \( \mathcal{H} = \{ h_1, \ldots, h_K \} \).
- ERM learning: \( h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \).

The probability that the empirical risk fails can be reduced to zero if we have enough examples:
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}
\left(
\underbrace{|R(h_m) - R_{\mathcal{T}^m}(h_m)|}_{\text{ER fails}} \geq \varepsilon\right)$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}\left( \left| R(h_m) - R_{T^m}(h_m) \right| \geq \varepsilon \right)$$

**ER fails**

$$\mathcal{B}(h) = \left\{ T^m \in (\mathcal{X} \times \mathcal{Y})^m \left| R_{T^m}(h) - R(h) \right| \geq \varepsilon \right\}$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}\left( \left| R(h_m) - R_{\mathcal{T}^m}(h_m) \right| \geq \varepsilon \right)$$

$\text{ER fails}$

$\mathcal{B}(h) = \{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \}$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_T^m(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}\left( \left| R(h_m) - R_T^m(h_m) \right| \geq \varepsilon \right) \leq \mathbb{P}\left( \bigcup_{i=1}^K \left| R(h_i) - R_T^m(h_i) \right| \geq \varepsilon \right)$$

$$\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \left| R_T^m(h) - R(h) \right| \geq \varepsilon \right\}$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class \( \mathcal{H} = \{ h_1, \ldots, h_K \} \).

- ERM learning: \( h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h) \).

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

\[
\mathbb{P} \left( \left| R(h_m) - R_{\mathcal{T}^m}(h_m) \right| \geq \varepsilon \right) \leq \mathbb{P} \left( \bigcup \left( \left| R(h_1) - R_{\mathcal{T}^m}(h_1) \right| \geq \varepsilon, \ldots, \left| R(h_K) - R_{\mathcal{T}^m}(h_K) \right| \geq \varepsilon \right) \right)
\]

Mutually exclusive events: \( \mathcal{B}(h_1) \cap \mathcal{B}(h_2) \cap \mathcal{B}(h_3) = \emptyset \)

\[
\mathbb{P}(\mathcal{B}(h_1) \cup \mathcal{B}(h_2) \cup \mathcal{B}(h_3)) = \mathbb{P}(\mathcal{B}(h_1)) + \mathbb{P}(\mathcal{B}(h_2)) + \mathbb{P}(\mathcal{B}(h_3))
\]

\[
\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (X \times Y)^m \left| \left| R_{\mathcal{T}^m}(h) - R(h) \right| \geq \varepsilon \right\}
\]
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class \( \mathcal{H} = \{ h_1, \ldots, h_K \} \).
- ERM learning: \( h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h) \).

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

\[
P(\left| R(h_m) - R_{T^m}(h_m) \right| \geq \varepsilon) \leq P \left( \bigcup_{k=1}^{K} \left| R(h_1) - R_{T^m}(h_1) \right| \geq \varepsilon \right)
\]

Union bound:

\[
P(B(h_1) \cup B(h_2) \cup B(h_3)) \leq P(B(h_1)) + P(B(h_2)) + P(B(h_3))
\]

\[
B(h) = \{ T^m \in (X \times Y)^m | |R_{T^m}(h) - R(h)| \geq \varepsilon \}
\]
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}\left(\frac{|R(h_m) - R_{T^m}(h_m)|}{\text{ER fails}} \geq \varepsilon\right) \leq \mathbb{P}\left(\bigcup \left(\frac{|R(h_1) - R_{T^m}(h_1)|}{\geq \varepsilon} \cup \frac{|R(h_2) - R_{T^m}(h_2)|}{\geq \varepsilon} \cup \ldots \right)\right)$$

$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\left|h - R_{T^m}(h)\right| \geq \varepsilon\right)$$

Union bound:

$$\mathbb{P}(\mathcal{B}(h_1) \cup \mathcal{B}(h_1) \cup \mathcal{B}(h_2)) \leq \mathbb{P}(\mathcal{B}(h_1)) + \mathbb{P}(\mathcal{B}(h_2)) + \mathbb{P}(\mathcal{B}(h_3))$$

$$\mathcal{B}(h) = \{T^m \in (X \times Y)^m \left|h - R_{T^m}(h)\right| \geq \varepsilon\}$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}(\underbrace{|R(h_m) - R_{T^m}(h_m)| \geq \varepsilon}_{\text{ER fails}}) \leq \mathbb{P}\left(\bigcup_{i=1}^{K} \left| R(h_i) - R_{T^m}(h_i) \right| \geq \varepsilon \right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(\left| R(h) - R_{T^m}(h) \right| \geq \varepsilon \right)$$

$$a \geq \varepsilon \text{ or } b \geq \varepsilon \iff \max\{a, b\} \geq \varepsilon$$

$$\mathcal{B}(h) = \left\{ \mathcal{T}^m \in (\mathcal{X} \times \mathcal{Y})^m \middle| \left| R_{T^m}(h) - R(h) \right| \geq \varepsilon \right\}$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

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

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

$$a \geq \varepsilon \text{ or } b \geq \varepsilon \iff \max\{a, b\} \geq \varepsilon$$

$$\mathcal{B}(h) = \{ T^m \in (X \times Y)^m \left| R_{T^m}(h) - R(h) \right| \geq \varepsilon \}$$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_T^m(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

$$\mathbb{P}\left( |R(h_m) - R_T^m(h_m)| \geq \varepsilon \right) \leq \mathbb{P}\left( \max_{h \in \mathcal{H}} |R(h) - R_T^m(h)| \geq \varepsilon \right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left( |R(h) - R_T^m(h)| \geq \varepsilon \right)$$

Hoeffding inequality:

$$\mathbb{P}( |R(h) - R_T^m| \geq \varepsilon ) \leq 2e^{\frac{-2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}$$

$\mathcal{B}(h) = \{ T^m \in (\mathcal{X} \times \mathcal{Y})^m | |R_T^m(h) - R(h)| \geq \varepsilon \}$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

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

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

$$\leq 2 |\mathcal{H}| e^{-\frac{2m \varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}$$

Hoeffding inequality:

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

$\mathcal{B}(h) = \{ T^m \in (\mathcal{X} \times \mathcal{Y})^m \left| R_{T^m}(h) - R(h) \right| \geq \varepsilon \}$
The overfitting can be eliminated in case of the finite hypothesis space

- Assume a finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$.
- ERM learning: $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$.

The probability that the empirical risk fails can be reduced to zero if we have enough examples:

\[
\mathbb{P}\left( \left| R(h_m) - R_{\mathcal{T}^m}(h_m) \right| \geq \varepsilon \right) \leq \mathbb{P}\left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left( \left| R(h) - R_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}}
\]

1. $\mathbb{P}(\text{ER fails for } h_m \in \mathcal{H})$ is replaced by $\mathbb{P}(\text{ER can fail for some } h \in \mathcal{H})$.
2. Union bound.
3. Hoeffding inequality.
Uniform Law of Large Numbers

We have shown for that for the finite hypothesis space, \( \mathcal{H} = \{ h_1, \ldots, h_K \} \), the Law of Large Numbers holds simultaneously (uniformly) for every \( h \in \mathcal{H} \):

\[
P \left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{F}^m}(h) \right| \geq \varepsilon \right) \leq \frac{2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(\ell_{\max}-\ell_{\min})^2}}}{\text{ER can fail}} \text{ converges to } 0 \text{ for } m \to \infty
\]
Uniform Law of Large Numbers

We have shown for the finite hypothesis space, \( \mathcal{H} = \{ h_1, \ldots, h_K \} \), the Law of Large Numbers holds simultaneously (uniformly) for every \( h \in \mathcal{H} \):

\[
\mathbb{P}\left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}m}(h) \right| \geq \varepsilon \right) \leq \frac{2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(\ell_{\max} - \ell_{\min})^2}}}{\text{ER can fail}} \text{ converges to } 0 \text{ for } m \to \infty
\]

**Definition:** We say that Uniform Law of Large Numbers applies for hypothesis space \( \mathcal{H} \subset \mathcal{Y}^\mathcal{X} \) if there exists a function \( m_\mathcal{H} : \mathbb{R}_{>0} \times (0, 1) \to \mathbb{N} \) such that for every \( \varepsilon > 0, \delta \in (0, 1) \), every distribution \( p(x,y) \) and every \( m \geq m_\mathcal{H}(\varepsilon, \delta) \) the following inequality holds

\[
\mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}m}(h) \right| \geq \varepsilon \right) \leq \delta.
\]
Uniform Law of Large Numbers

We have shown for that for the finite hypothesis space, \( \mathcal{H} = \{h_1, \ldots, h_K\} \), the Law of Large Numbers holds simultaneously (uniformly) for every \( h \in \mathcal{H} \):

\[
P \left( \max_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}}(h) \right| \geq \varepsilon \right) \leq \frac{2|\mathcal{H}| e^{-\frac{2m \varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}}{\text{converges to 0 for } m \to \infty}
\]

ER can fail

**Definition:** We say that Uniform Law of Large Numbers applies for hypothesis space \( \mathcal{H} \subset \mathcal{Y}^\mathcal{X} \) if there exists a function \( m_\mathcal{H} : \mathbb{R}_{>0} \times (0, 1) \to \mathbb{N} \) such that for every \( \varepsilon > 0, \delta \in (0, 1) \), every distribution \( p(x, y) \) and every \( m \geq m_\mathcal{H}(\varepsilon, \delta) \) the following inequality holds

\[
P \left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}}(h) \right| \geq \varepsilon \right) \leq \delta .
\]

The next lecture:
- If ULLN applies then ERM learning is guaranteed to succeed.
- VC dimension as a tool to recognize that ULLN applies for given \( \mathcal{H} \).
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

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

holds for all $h \in \mathcal{H}$ simultaneously.
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

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

holds for all $h \in \mathcal{H}$ simultaneously.

- To decrease the complexity term: increase $m$ or decrease $|\mathcal{H}|$. 
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

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

holds for all \( h \in \mathcal{H} \) simultaneously.

- To decreases the complexity term: increase \( m \) or decrease \( |\mathcal{H}| \).
- The generalization bound holds for any learning algorithm not just ERM.
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

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

holds for all $h \in \mathcal{H}$ simultaneously.

- To decrease the complexity term: increase $m$ or decrease $|\mathcal{H}|$.
- The generalization bound holds for any learning algorithm not just ERM.
- Recommendations for learning:
  1. Minimize the empirical risk.
  2. Use as much training examples $m$ as you can.
  3. Limit the size of the hypothesis space $|\mathcal{H}|$, i.e. use prior knowledge.
Generalization bound for finite hypothesis class: the proof

We have shown that ULLN holds for finite hypothesis class $\mathcal{H}$:

$$\mathbb{P}\left(\max_{h \in \mathcal{H}} |R_{\mathcal{T}m}(h) - R(h)| \geq \varepsilon \right) \leq 2|\mathcal{H}|e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}$$
Generalization bound for finite hypothesis class: the proof

- We have shown that ULLN holds for finite hypothesis class $\mathcal{H}$:

$$
\mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{Tm}(h) - R(h)| \geq \varepsilon \right) \leq 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}}
$$

- Prob. $R_{Tm}(h)$ is a good proxy of $R(h)$ for all $h \in \mathcal{H}$ simultaneously:

$$
\mathbb{P}\left( |R_{Tm}(h) - R(h)| < \varepsilon, \forall h \in \mathcal{H} \right) = \mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{Tm}(h) - R(h)| < \varepsilon \right) = 1 - \mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{Tm}(h) - R(h)| \geq \varepsilon \right) \geq 1 - 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}} = 1 - \delta
$$
Generalization bound for finite hypothesis class: the proof

- We have shown that ULLN holds for finite hypothesis class $\mathcal{H}$:

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

- Prob. $R_{\mathcal{T}}^m(h)$ is a good proxy of $R(h)$ for all $h \in \mathcal{H}$ simultaneously:

$$
\mathbb{P}\left( |R_{\mathcal{T}}^m(h) - R(h)| < \varepsilon, \forall h \in \mathcal{H} \right) = \mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}}^m(h) - R(h)| < \varepsilon \right)
= 1 - \mathbb{P}\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}}^m(h) - R(h)| \geq \varepsilon \right)
\geq 1 - 2|\mathcal{H}| e^{-\frac{2m\varepsilon^2}{(\ell_{\text{max}} - \ell_{\text{min}})^2}} = 1 - \delta
$$

- Solving the last equality for $\varepsilon$ yields $\varepsilon = L \sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}$ so that:

$$
\mathbb{P}\left( |R_{\mathcal{T}}^m(h) - R(h)| < L \sqrt{\frac{\log 2|\mathcal{H}| + \log \frac{1}{\delta}}{2m}}, \forall h \in \mathcal{H} \right) \geq 1 - \delta
$$
Summary

◆ Learning algorithm: the definition.

◆ Empirical Risk Minimization.

◆ Unrestricted hypothesis space: the ERM can overfit regardless the number of training examples.

◆ Finite hypothesis space: the chance of overfitting can be always eliminated.

◆ Uniform Law of Large Numbers.

◆ Generalization bound for finite hypothesis space.
$\beta(h_1)$
$\mathcal{B}(h_1)$

$\mathcal{B}(h_2)$

$\mathcal{B}(h_3)$
$\mathcal{B}(h_1)$
$\mathcal{B}(h_2)$
$\mathcal{B}(h_3)$
\[ \mathcal{B}(h_1) \]
\[ \mathcal{B}(h_2) \]
\[ \mathcal{B}(h_3) \]
\[ \mathcal{B}(h_1) \]
\[ \mathcal{B}(h_2) \]
\[ \mathcal{B}(h_3) \]
$\mathcal{B}(h_1)$

$\mathcal{B}(h_2)$

$\mathcal{B}(h_3)$