**Linear classifier minimizing classification error**

- $\mathcal{X}$ is a set of observations and $\mathcal{Y} = \{+1, -1\}$ a set of hidden labels
- $\phi: \mathcal{X} \rightarrow \mathbb{R}^n$ is fixed feature map embedding $\mathcal{X}$ to $\mathbb{R}^n$
- **Task:** find linear classification strategy $h: \mathcal{X} \rightarrow \mathcal{Y}$

$$h(x; \mathbf{w}, b) = \text{sign}(\langle \mathbf{w}, \phi(x) \rangle + b) = \begin{cases} +1 & \text{if } \langle \mathbf{w}, \phi(x) \rangle + b \geq 0 \\ -1 & \text{if } \langle \mathbf{w}, \phi(x) \rangle + b < 0 \end{cases}$$

with minimal expected risk

$$R^{0/1}(h) = \mathbb{E}_{(x,y) \sim p} \left( \ell^{0/1}(y, h(x)) \right) \quad \text{where} \quad \ell^{0/1}(y, y') = [y \neq y']$$

- We are given a set of training examples

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

drawn from i.i.d. with the distribution $p(x, y)$. 
ERM learning for linear classifiers

- The Empirical Risk Minimization principle leads to solving

\[
(w^*, b^*) \in \operatorname{Argmin}_{(w,b) \in (\mathbb{R}^n \times \mathbb{R})} R^{0/1}_{\mathcal{T}}(h(\cdot ; w, b))
\]

where the empirical risk is

\[
R^{0/1}_{\mathcal{T}}(h(\cdot ; w, b)) = \frac{1}{m} \sum_{i=1}^{m} [y^i \neq h(x^i ; w, b)]
\]

- Algorithmic issues (next lecture): in general, there is no known algorithm solving the task (1) in time polynomial in \(m\).

- The uniform bound on the generalization error (this lecture):

\[
P\left( \sup_{h \in \mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}}(h) \right| \geq \varepsilon \right) \leq B(m, \mathcal{H}, \varepsilon)
\]
Vapnik-Chervonenkis (VC) dimension

Definition: Let $\mathcal{H} \subseteq \{-1, +1\}^X$ and $\{x^1, \ldots, x^m\} \in X^m$ be a set of $m$ input observations. The set $\{x^1, \ldots, x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $y \in \{+1, -1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i) = y^i$, $i \in \{1, \ldots, m\}$.

Definition: Let $\mathcal{H} \subseteq \{-1, +1\}^X$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $X$ which can be shattered by $\mathcal{H}$.
Theorem: The VC-dimension of the hypothesis class of all two-class linear classifiers operating in $n$-dimensional feature space
$$\mathcal{H} = \{h(x; \mathbf{w}, b) = \text{sign}(\langle \mathbf{w}, \phi(x) \rangle + b) \mid (\mathbf{w}, b) \in (\mathbb{R}^n \times \mathbb{R})\}$$ is $n + 1$.

Example for $n = 2$-dimensional feature class
ULLN for two class predictors and 0/1-loss

**Theorem:** Let $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$ and $\mathcal{T}^m = \{(x^1, y^1), \ldots, (x^m, y^m)\} \in (\mathcal{X} \times \mathcal{Y})^m$ a training set drawn from i.i.d. random variables with distribution $p(x, y)$. Then, for any $\varepsilon > 0$ it holds

$$
P\left(\sup_{h \in \mathcal{H}} \left| R^{0/1}_{\mathcal{H}}(h) - R^{0/1}_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 4 \left(\frac{2 e m}{d}\right)^d e^{-\frac{m \varepsilon^2}{8}}$$

**Corollary:** Let $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$. Then ULLN applies.
Summary: uniform bounds on the generalization error

- We learned how to bound the generalization error uniformly for:
  - Finite hypothesis class \( \mathcal{H} = \{h_1, \ldots, h_K\} \):
    \[
    P\left( \max_{h \in \mathcal{H}} |R_{\mathcal{T}m}(h) - R(h)| \geq \varepsilon \right) \leq 2|\mathcal{H}|e^{\frac{-2m\varepsilon^2}{(b-a)^2}}
    \]
  - Two-class classifiers \( \mathcal{H} \subseteq \{+1, -1\}^X \) a finite VC-dimensions \( d \):
    \[
    P\left( \sup_{h \in \mathcal{H}} \left| R^{0/1}_{\mathcal{T}m}(h) - R^{0/1}_m(h) \right| \geq \varepsilon \right) \leq 4 \left( \frac{2em}{d} \right)^d e^{-\frac{me\varepsilon^2}{8}}
    \]

In both cases the bound goes to zero, i.e., ULLN applies.

- Does ERM algorithm \( h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}m}(h) \) finds strategy with the minimal risk \( R(h) \)?
Statistically consistent learning algorithm

- $h_H \in \text{Argmin}_{h \in H} R(h)$ the best strategy in $H$ has the risk $R(h_H)$
- $h_m = A(T_m)$ strategy learned from $T_m$ with has risk $R(h_m)$
- $R(h_m) - R(h_H)$ is the estimation error
- The statistically consistent algorithm can make the estimation error arbitrarily small if it has enough examples.

**Definition:** The algorithm $A$: $\bigcup_{m=1}^{\infty} (X \times Y)^m \rightarrow H$ is statistically consistent in $H \subseteq Y^X$ if for any $p(x, y)$ it holds that

$$\forall \varepsilon > 0: \lim_{m \rightarrow \infty} \mathbb{P}\left( \underbrace{R(h_m) - R(h_H)}_{\text{high estimation error}} \geq \varepsilon \right) = 0$$

where $h_m = A(T^m)$ is learned by $A$ for $T^m$ generated from $p(x, y)$.
Example: generalization error and estimation error

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

- Highest generalization error

\[ P \left( R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon \right) \leq P \left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{\mathcal{T}m}(h) \right| \geq \frac{\varepsilon}{2} \right) \]

- Estimation error

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

\[ R(h) \quad R_{\mathcal{T}m}(h) \]

\[ R(h_m) \quad R_{\mathcal{T}m}(h_m) \]
Theorem: ULLN implies consistency of ERM

For fixed $T^m$ and $h_m \in \text{Argmin}_{h \in \mathcal{H}} R_{T^m}(h)$ we have:

$$R(h_m) - R(h_\mathcal{H}) = \left( R(h_m) - R_{T^m}(h_m) \right) + \left( R_{T^m}(h_m) - R(h_\mathcal{H}) \right)$$

$$\leq \left( R(h_m) - R_{T^m}(h_m) \right) + \left( R_{T^m}(h_\mathcal{H}) - R(h_\mathcal{H}) \right)$$

$$\leq 2 \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right|$$

Therefore $\varepsilon \leq R(h_m) - R(h_\mathcal{H})$ implies $\frac{\varepsilon}{2} \leq \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right|$ and

$$\mathbb{P}\left( R(h_m) - R(h_\mathcal{H}) \geq \varepsilon \right) \leq \mathbb{P}\left( \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right| \geq \frac{\varepsilon}{2} \right)$$

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).
Finite sample bound on the estimation error

Let $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ be the hypothesis class with the best predictor

$$h_\mathcal{H} \in \text{Argmin}_{h \in \mathcal{H}} R(h)$$

We learn $h_m$ from $\mathcal{T}^m \sim p(x, y)$ with the ERM algorithm

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

Assume we have for $\mathcal{H}$ the uniform bound on the generalization error

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

Then, for any $\varepsilon > 0$ the inequality

$$R(h_m) \leq R(h_\mathcal{H}) + \varepsilon$$

holds with the probability $1 - B(m, \mathcal{H}, \varepsilon / 2)$ at least.
Excess error $= \text{Estimation error} + \text{Approximation errors}$

The characters of the play:

- $R^* = \inf_{h \in \mathcal{Y} \times \mathcal{X}} R(h)$ best attainable true risk
- $R(h_H)$ best risk in $\mathcal{H}$ where $h_H \in \text{Argmin}_{h \in \mathcal{H}} R(h)$
- $R(h_m)$ risk of $h_m = A(T_m)$ learned from $T^m$

**Excess error:** the quantity we want to minimize

$$
\underbrace{R(h_m) - R^*}_{\text{excess error}} = \underbrace{R(h_m) - R(h_H)}_{\text{estimation error}} + \underbrace{R(h_H) - R^*}_{\text{approximation error}}
$$

Questions:

- What causes individual errors?
- How do the errors depend on $\mathcal{H}$ and $m$?