Error decomposition

- $R^* = \inf_{h \in Y \times X} R(h)$ best attainable risk
- $R(h_{\mathcal{H}})$ best risk in the class where $h_{\mathcal{H}} \in \text{Argmin}_{h \in \mathcal{H}} R(h)$
- $R(h_m)$ generalization error of $h_m = A(\mathcal{T}_m)$ learned from data $\mathcal{T}^m$

Error decomposition:

$$R(h_m) = \underbrace{(R(h_m) - R(h_{\mathcal{H}}))}_{\text{estimation error}} + \underbrace{(R(h_{\mathcal{H}}) - R^*)}_{\text{approximation error}} + R^*$$

- The approximation error: depends on $\mathcal{H}$ chosen prior to learning.
- The estimation error: depends on $\mathcal{H}$, data $\mathcal{T}$ and the algorithm $A$. 
Probably Approximately Correct (PAC) learning

Successful PAC learning algorithm

- Given a hypothesis space $\mathcal{H}$ and the loss $\ell$, the algorithm with high probability learns a predictor that has low estimation error.

- The following can be arbitrary: desired estimation error $\varepsilon > 0$, probability of failure $\delta \in (0, 1)$, and data distribution $p(x, y)$.

Definition. Algorithm is a successful PAC learner for hypothesis space $\mathcal{H}$ w.r.t. loss $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ if there exists a function (called sample complexity) $m^{\text{pac}}_{\mathcal{H}} : \mathbb{R}^+ \times (0, 1) \rightarrow \mathbb{N}$ such that: For every $\varepsilon > 0$, $\delta \in (0, 1)$, and every distribution $p(x, y)$, when running the algorithm on $m \geq m^{\text{pac}}_{\mathcal{H}}(\varepsilon, \delta)$ examples $\mathcal{T}_m$ i.i.d. drawn from $p(x, y)$, then the algorithm returns $h_m = A(\mathcal{T}_m)$ such that

$$
\mathbb{P}\left( R(h_m) - R(h_{\mathcal{H}}) \leq \varepsilon \right) \geq 1 - \delta .
$$
ULLN implies that ERM is successful PAC learner

ULLN applies for $\mathcal{H} \subset \mathcal{Y}^X$: there exists $m_{\mathcal{H}}^{ul}: \mathbb{R}_{>0} \times (0, 1) \rightarrow \mathbb{N}$ such that for every $\varepsilon > 0, \delta \in (0, 1)$, every distribution $p(x, y)$ and every $m \geq m_{\mathcal{H}}^{ul}(\varepsilon, \delta)$ it holds that

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

ER can fail

Successful PAC learner for $\mathcal{H} \subset \mathcal{Y}^X$: there exists $m_{\mathcal{H}}^{\text{pac}}: \mathbb{R}_{>0} \times (0, 1) \rightarrow \mathbb{N}$ such that when running the algorithm on $m \geq m_{\mathcal{H}}^{\text{pac}}(\varepsilon, \delta)$ examples $\mathcal{T}^m \sim p^m$ then it returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}\left( R(h_m) - R(h_\mathcal{H}) \leq \varepsilon \right) \geq 1 - \delta .$$

low estimation error

**Theorem:** If ULLN applies for $\mathcal{H} \subset \mathcal{Y}^X$ with a function $m_{\mathcal{H}}^{ul}$ then ERM is a successful PAC learner for $\mathcal{H}$ with the sample complexity $m_{\mathcal{H}}^{\text{pac}}(\varepsilon, \delta) = m_{\mathcal{H}}^{ul}(\varepsilon/2, \delta)$. 
ULLN implies that ERM is successful PAC learner: proof (1)

ULLN:  
\[ m \geq m^ul_H(\varepsilon, \delta) \Rightarrow \mathbb{P}\left(\sup_{h \in H} |R(h) - R_{Tm}(h)| > \varepsilon\right) \leq \delta \]

ER can fail

\[ R(h_m) - R(h_H) \leq 2 \sup_{h \in H} |R(h) - R_{Tm}(h)| \]

estimation error

\[ R(h_m) - R(h_H) > \bar{\varepsilon} \Rightarrow \sup_{h \in H} |R(h) - R_{Tm}(h)| > \frac{\bar{\varepsilon}}{2} \]

\[ \mathbb{P}\left(R(h_m) - R(h_H) > \bar{\varepsilon}\right) \leq \mathbb{P}\left(\sup_{h \in H} |R(h) - R_{Tm}(h)| > \frac{\bar{\varepsilon}}{2}\right) \]

\[ m \geq m^ul_H(\frac{\bar{\varepsilon}}{2}, \delta) \Rightarrow \mathbb{P}\left(R(h_m) - R(h_H) > \bar{\varepsilon}\right) \leq \delta \]

\[ \mathbb{P}\left(R(h_m) - R(h_H) \leq \bar{\varepsilon}\right) = 1 - \mathbb{P}\left(R(h_m) - R(h_H) > \bar{\varepsilon}\right) \geq 1 - \delta \]

Successful PAC:  
\[ m \geq m^\text{pac}_H(\bar{\varepsilon}, \delta) \Rightarrow \mathbb{P}\left(R(h_m) - R(h_H) \leq \bar{\varepsilon}\right) \leq 1 - \delta \]

where \( m^\text{pac}_H(\bar{\varepsilon}, \delta) = m^ul_H(\frac{\bar{\varepsilon}}{2}, \delta) \)
ULLN implies that ERM is successful PAC learner: proof (2)

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

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

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

$$\leq 2 \sup_{h \in \mathcal{H}} \left| R(h) - R_{T^m}(h) \right|$$
ERM is successful PAC learner for finite hypothesis space

- We showed that for finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$ it holds

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

and hence ULLN applies with $m^{\text{ul}}_{\mathcal{H}}(\varepsilon, \delta) = \frac{\log 2|\mathcal{H}| - \log \delta}{2\varepsilon^2}(\ell_{\text{max}} - \ell_{\text{min}})^2$.

- Therefore ERM is successful PAC learner for $\mathcal{H}$ with sample complexity

$$m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon}, \delta) = 2\frac{\log 2|\mathcal{H}| - \log \delta}{\bar{\varepsilon}^2}(\ell_{\text{max}} - \ell_{\text{min}})^2,$$

that is, when running ERM on $\mathcal{T}^m$ with $m \geq m^{\text{pac}}_{\mathcal{H}}(\bar{\varepsilon}, \delta)$ then it returns $h_m = A(\mathcal{T}^m)$ such that

$$\mathbb{P}\left( R(h_m) - R(h_{\mathcal{H}}) \leq \bar{\varepsilon} \right) \geq 1 - \delta.$$
Linear classifier minimizing classification error

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

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

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

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

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

- ERM for $\mathcal{H} = \{ h(x; w, b) = \text{sign}(\langle w, \phi(x) \rangle + b) \mid (w, b) \in \mathbb{R}^{n+1} \}$ leads to

$$
(w^*, b^*) \in \text{Argmin} \ R_{T_m}^{0/1}(h) = \text{Argmin} \ R_{T_m}^{0/1}(h(\cdot; w, b))
$$

where the empirical risk is

$$
R_{T_m}^{0/1}(h(\cdot; w, b)) = \frac{1}{m} \sum_{i=1}^{m} [y^i \neq h(x^i; w, b)]
$$

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

- Does ULLN applies for the class of two-class linear classifiers? If yes then ERM is PAC successful learner.
VC dimension is a concept to measure complexity of an infinite hypothesis space \( \mathcal{H} \subseteq \{-1, +1\}^X \).

**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 $n + 1$.
ULLN for two class predictors and 0/1-loss

**Theorem:** Let $\mathcal{H} \subset \{+1, -1\}^\mathcal{X}$ 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 draw from i.i.d. rand vars with distribution $p(x, y)$. Then for any $\varepsilon > 0$ it holds

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

**Corollary:** Let $\mathcal{H} \subset \{+1, -1\}^\mathcal{X}$ be a hypothesis class with a finite VC dimension $d < \infty$. Then, ULLN applies for $\mathcal{H}$ and there exists a constant $C$ such that

$$m_{\mathcal{H}}^{\text{pac}}(\varepsilon, \delta) \leq C \frac{d - \log \delta}{\varepsilon^2}
$$

that is, ERM is PAC successful learner.

**Remark:** Recall that in case of finite hypothesis space $\mathcal{H} = \{h_1, \ldots, h_K\}$ and 0/1-loss we have the sample complexity $m_{\mathcal{H}}^{\text{pac}}(\varepsilon, \delta) = 2^{\log 2 |\mathcal{H}| - \log \delta} / \varepsilon^2$. 
Summary

- Error decomposition: Generalization error = estimation error + approximation error + Bayes risk.
- Probably Approximately Correct (PAC) learning.
- ULLN implies that ERM is a successful PAC learner.
- VC dimension: hypothesis space complexity of two-class classifier.
- VC dimension of linear hypothesis space.
- Finite VC dimension implies that ERM is a successful PAC learner.