Statistical Machine Learning (BE4M33SSU)
Lecture 4: Empirical Risk Minimization II

Czech Technical University in Prague
V. Franc

Recap of the previous lecture

- We bounded the probability that empirical risk $R_{\mathcal{T}^m}(h_m)$ is not a good proxy of true risk $R(h_m)$ where $h_m = A(\mathcal{T}^m)$ is a learned from $\mathcal{T}^m$:

$$\mathbb{P}\Big(\Big|R(h_m) - R_{\mathcal{T}^m}(h_m)\Big| \geq \varepsilon\Big) \underbrace{\leq}_{\text{uniform bound}} \mathbb{P}\Big(\sup_{h\in\mathcal{H}}\Big|R(h) - R_{\mathcal{T}^m}(h)\Big| \geq \varepsilon\Big)$$

$$\underbrace{\leq}_{\text{union bound}} \sum_{h\in\mathcal{H}} \mathbb{P}\Big(\Big|R(h) - R_{\mathcal{T}^m}(h)\Big| \geq \varepsilon\Big) \underbrace{\leq}_{\text{Hoeffding inequality}} 2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = B(m, |\mathcal{H}|, \varepsilon)$$

- We derived a generalization bound:

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

- This lecture answers the following quations:
  - How to deal with infinite hypothesis space $\mathcal{H}$?
  - How to define a good learning algorithm? Is ERM good?

Linear classifier minimizing classification error

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

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

with minimal expected risk

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

- 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

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

$$(\boldsymbol{w}^*, b^*) \in \operatorname*{Argmin}_{h \in \mathcal{H}} R^{0/1}_{\mathcal{T}^m}(h) = \operatorname*{Argmin}_{(\boldsymbol{w},b)\in(\mathbb{R}^n \times \mathbb{R})} R^{0/1}_{\mathcal{T}^m}(h(\cdot; \boldsymbol{w}, b)) \qquad (1)$$

  where the empirical risk is

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

- 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? Recall that ULLN $\forall\, \varepsilon > 0$: $\mathbb{P}\big(\sup_{h \in \mathcal{H}} \big|R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h)\big| \geq \varepsilon\,\big) = 0$

Vapnik-Chervonenkis (VC) dimension

- VC dimension is a concept to measure complexity of an infinite hypothesis space $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$.

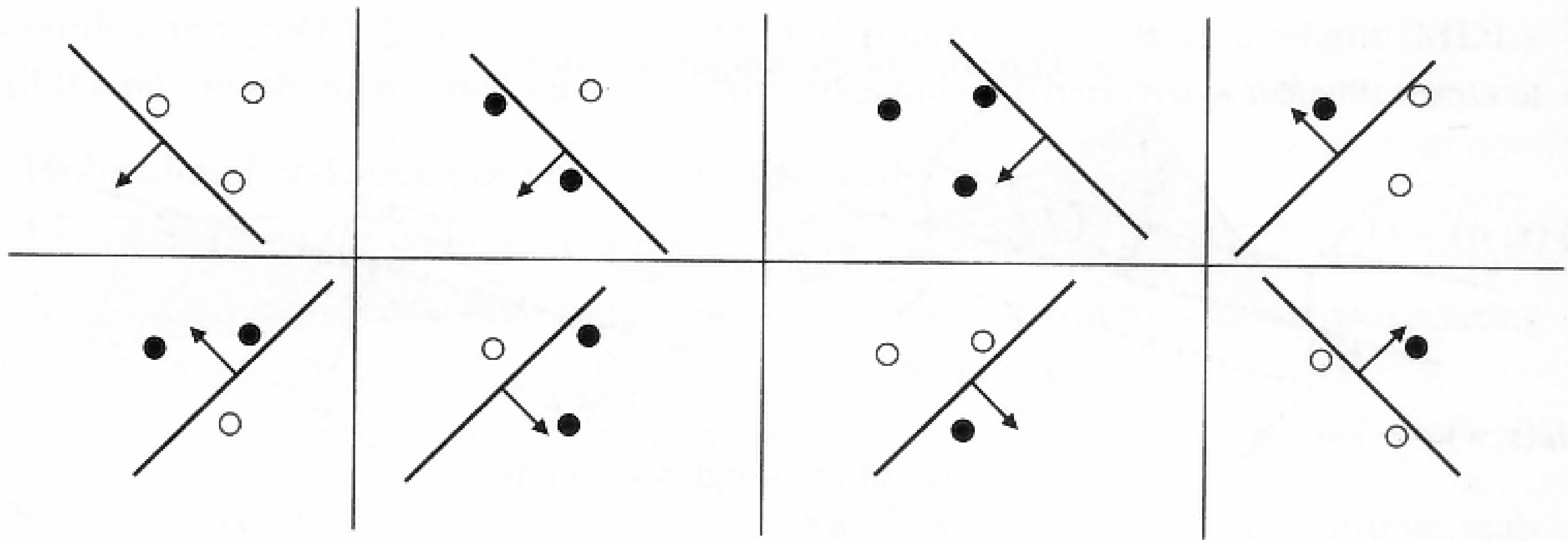
Definition: Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ and $\{x^1,\dots,x^m\} \in \mathcal{X}^m$ be a set of m input observations. The set $\{x^1,\dots,x^m\}$ is said to be shattered by $\mathcal{H}$ if for all $\boldsymbol{y} \in \{+1,-1\}^m$ there exists $h \in \mathcal{H}$ such that $h(x^i)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

Definition: Let $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$. The Vapnik-Chervonenkis dimension of $\mathcal{H}$ is the cardinality of the largest set of points from $\mathcal{X}$ which can be shattered by $\mathcal{H}$.

VC dimension of class of two-class linear classifiers

Theorem: The VC-dimension of the hypothesis class of all two-class linear classifiers operating in n-dimensional feature space
$\mathcal{H} = \{h(x; \boldsymbol{w}, b) = \operatorname{sign}(\langle \boldsymbol{w}, \boldsymbol{\phi}(x) \rangle + b) \mid (\boldsymbol{w}, b) \in (\mathbb{R}^n \times \mathbb{R})\}$ is $n + 1$.

Example for $n = 2$-dimensional feature space



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 draw from i.i.d. rand vars with distribution $p(x,y)$. Then

$$\forall\, \varepsilon > 0 \colon \mathbb{P}\biggl(\sup_{h\in\mathcal{H}} \Bigl| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h) \Bigr| \geq \varepsilon \biggr) \leq 4 \biggl(\frac{2\, e\, m}{d}\biggr)^{d} e^{-\frac{m\,\varepsilon^2}{8}}$$

Corollary: Let $\mathcal{H} \subseteq \{+1,-1\}^{\mathcal{X}}$ be a hypothesis class with VC dimension $d < \infty$. Then ULLN applies.

Summary: uniform law of large numbers

- We learned how to bound deviation between the empirical and the true risk uniformly for:

  - Finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$:

$$\mathbb{P}\Big(\max_{h \in \mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon \Big) \leq 2|\mathcal{H}| e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = B_1(m, |\mathcal{H}|, \varepsilon)$$

  - Two-class classifiers $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ a finite VC-dimensions d:

$$\mathbb{P}\left(\sup_{h \in \mathcal{H}} \left| R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h) \right| \geq \varepsilon \right) \leq 4 \left(\frac{2\,e\,m}{d} \right)^d e^{-\frac{m\,\varepsilon^2}{8}} = B_2(m, d, \varepsilon)$$

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

- Does ERM algorithm $h_m \in \operatorname*{Argmin}_{h \in \mathcal{H}} R_{\mathcal{T}^m}(h)$ finds strategy with the minimal risk $R(h)$?

Excess error = Estimation error + Approximation error

The characters of the play:

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

Excess error: the quantity we want to minimize

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

Note that:

- The approximation error depends on $\mathcal{H}$.
- The estimation error is random and depends on $\mathcal{H}$, m and A.

Universally statistically consistent learning algorithm

- A good algorihtm $h_m = A(\mathcal{T}^m)$ for $\mathcal{H}$ can make the estimation error $R(h_m) - R(h_\mathcal{H})$ arbitrarily small if it has enough examples m.

Definition: Let $\mathcal{H} \subseteq \mathcal{Y}^\mathcal{X}$ be a hypothesis space and $h_\mathcal{H} \in \operatorname{Argmin}_{h\in\mathcal{H}} R(h)$ the best strategy in $\mathcal{H}$. The algorithm $A\colon \cup_{m=1}^\infty (\mathcal{X}\times\mathcal{Y})^m \to \mathcal{H}$ is universally statistically consistent in $\mathcal{H}$ if there exists a function $m_\mathcal{H}\colon (0,1)^2 \to \mathbb{N}$ such that, for every $\varepsilon, \delta \in (0,1)$, if $m \geq m_\mathcal{H}(\varepsilon,\delta)$ then with probability $1-\delta$ it holds that

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

where $h_m = A(\mathcal{T}^m)$ is learned on $\mathcal{T}^m$ generated i.i.d. from $p(x,y)$.

- Equivalently we can say that algorithm is univ. stat. consistent in $\mathcal{H}$ iff

$$\forall\, \varepsilon > 0\colon \lim_{m\to\infty} \mathbb{P}\Big(R(h_m) - R(h_\mathcal{H}) \geq \varepsilon\Big) = 0$$

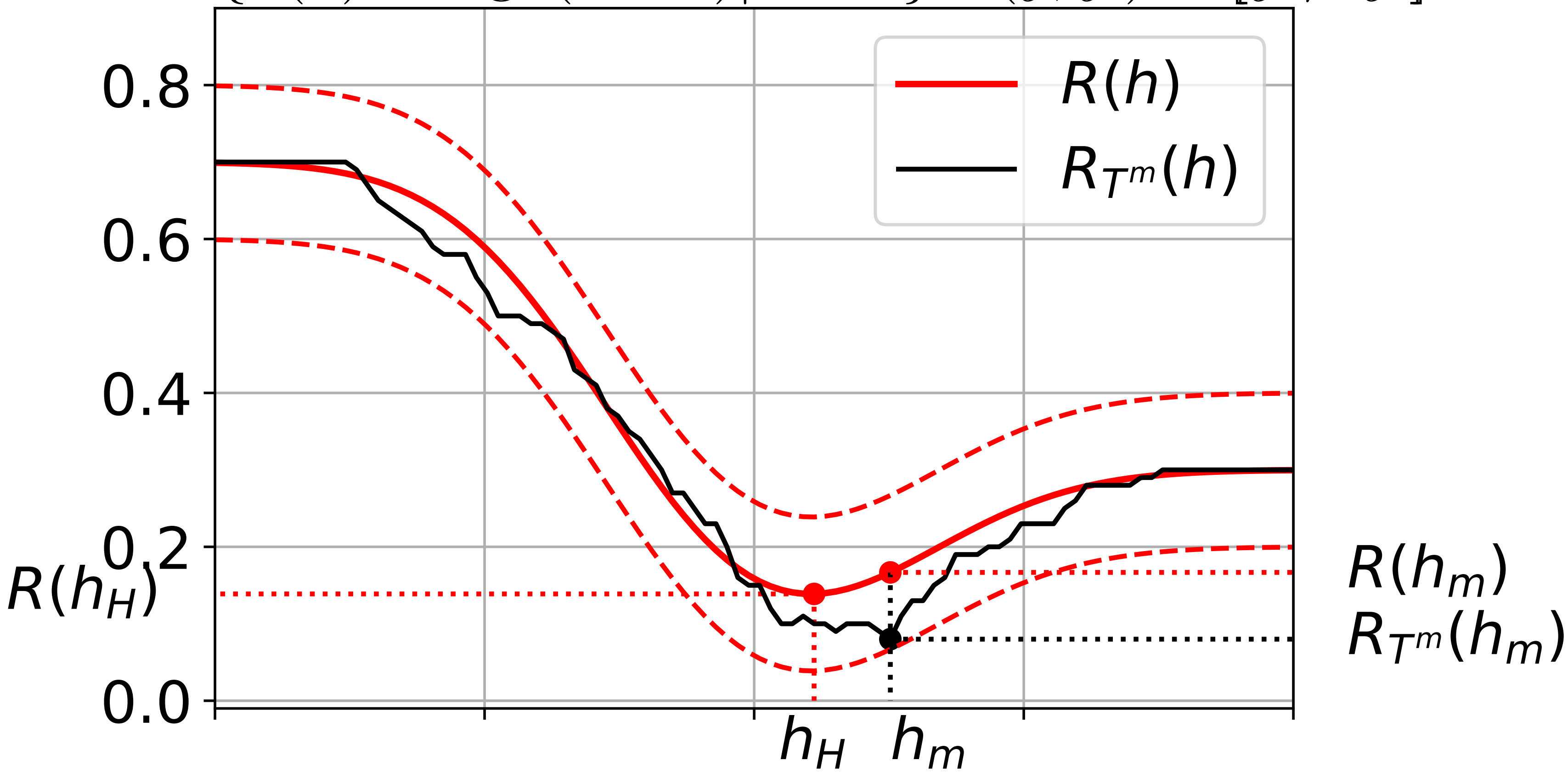
- When is ERM based algorithm universally statistically consistent?

ULLN implies universal consistency of ERM

$$\mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \varepsilon}_{\text{empirical risk fails for some h}\in\mathcal{H}} \Big) \leq B(m,\mathcal{H},\varepsilon)$$

$$\mathbb{P}\Big(\underbrace{R(h_m) - R(h_{\mathcal{H}})}_{\text{estimation error}} \geq \varepsilon \Big) \leq \mathbb{P}\Big(\underbrace{\sup_{h\in\mathcal{H}} \big|R(h) - R_{\mathcal{T}^m}(h)\big| \geq \tfrac{\varepsilon}{2}}_{\text{empirical risk fails for some h}\in\mathcal{H}} \Big) \leq B(m,H,\tfrac{\varepsilon}{2})$$

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



Theorem: ULLN implies universal consistency of ERM

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

$$
\begin{aligned}
R(h_m) - R(h_{\mathcal{H}}) &= \Big(R(h_m) - R_{\mathcal{T}^m}(h_m) \Big) + \Big(R_{\mathcal{T}^m}(h_m) - R(h_{\mathcal{H}}) \Big) \\
&\leq \Big(R(h_m) - R_{\mathcal{T}^m}(h_m) \Big) + \Big(R_{\mathcal{T}^m}(h_{\mathcal{H}}) - R(h_{\mathcal{H}}) \Big) \\
&\leq 2 \sup_{h \in \mathcal{H}} \Big| R(h) - R_{\mathcal{T}^m}(h) \Big|
\end{aligned}
$$

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

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

so if converges the RHS to zero (ULLN) so does the LHS (estimation error).

Universal consistency for ERM algorithms

- We have shown relation between the estimation error and the uniform bound on the empirical risk:

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

- We have shown ULLN for:
  - Finite hypothesis class $\mathcal{H} = \{h_1, \ldots, h_K\}$:

$$\mathbb{P}\Big(\max_{h\in\mathcal{H}} |R_{\mathcal{T}^m}(h) - R(h)| \geq \varepsilon\Big) \leq 2|\mathcal{H}|e^{-\frac{2m\,\varepsilon^2}{(b-a)^2}} = B_1(m, |\mathcal{H}|, \varepsilon)$$

  - Two-class classifiers $\mathcal{H} \subseteq \{+1, -1\}^{\mathcal{X}}$ a finite VC-dimensions d:

$$\mathbb{P}\Bigg(\sup_{h\in\mathcal{H}} \Big|R^{0/1}(h) - R^{0/1}_{\mathcal{T}^m}(h)\Big| \geq \varepsilon\Bigg) \leq 4\left(\frac{2\,e\,m}{d}\right)^d e^{-\frac{m\,\varepsilon^2}{8}} = B_2(m, d, \varepsilon)$$

Corollary: If $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is finite or $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ has finite VC-dimension, then ERM algorithm is universally consistent in $\mathcal{H}$.

Bound on the number of training examples for finite hypothesis space

- For finite hypothesis space we derived that

$$\mathbb{P}\Big(R(h_m) - R(h_{\mathcal{H}}) \geq \varepsilon\Big) \leq \mathbb{P}\Big(\sup_{h\in\mathcal{H}} \Big|R(h) - R_{\mathcal{T}^m}(h)\Big| \geq \frac{\varepsilon}{2}\Big) \leq 2|\mathcal{H}|e^{-\frac{m\,\varepsilon^2}{2(b-a)^2}}$$

Corollary: Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a finite hypothesis space and $h_{\mathcal{H}} \in \operatorname{Argmin}_{h\in\mathcal{H}} R(h)$ the best strategy in $\mathcal{H}$. For very $\varepsilon, \delta \in (0,1)$, let us define $m_{\mathcal{H}} \colon (0,1)^2 \to \mathbb{N}$ such that

$$m_{\mathcal{H}}(\varepsilon,\delta) = \frac{2(\log 2|\mathcal{H}| - \log\delta)}{\varepsilon^2}(\ell_{max} - \ell_{min})^2\,.$$

Let $h_m \in \operatorname{Argmin}_{h\in\mathcal{H}} R_{\mathcal{T}^m}(h)$ be a strategy learned by ERM algorithm from $m \geq m_{\mathcal{H}}(\varepsilon,\delta)$ training examples $\mathcal{T}^m$ generated i.i.d. from some $p(x,y)$. Then, with probability $1-\delta$ at least it holds that

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