# Assignment 2 - Advance Machine Learning

[redacted]

**Problem 1.** Consider $\mathcal{H} = \{h_{\theta_1}: \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \mathbf{1}_{[\theta_1,\infty)}, \theta_1 \in \mathbb{R}\} \cup \{h_{\theta_2}: \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \{\mathbf{1}_{(-\infty,\theta_2)}, \theta_2 \in \mathbb{R}\}$.

a) Compute the shattering coefficient $\tau_{\mathcal{H}}(m)$ of the growth function for $m \geq 0$.
b) Compare your result with the general upper bound for the growth functions.
c) Does there exist a hypothesis class $\mathcal{H}$ for which $\tau_{\mathcal{H}}(m)$ is equal to the general upper bound (over $\mathbb{R}$ or another domain $X$)? If your answer is yes, please provide an example, if your answer is no please provide a justification.

**Solution problem 1.a:**

$$\mathcal{H} = \left\{h_{\theta_1}: \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \begin{cases} 1, & x \geq \theta_1 \\ 0, & otherwise \end{cases}, \theta_1 \in \mathbb{R}\right\} \cup \left\{h_{\theta_2}: \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \begin{cases} 1, & x < \theta_2 \\ 0, & otherwise \end{cases}, \theta_2 \in \mathbb{R}\right\}$$

$$= \overbrace{\left\{h_{\theta_1}: \mathbb{R} \to \{0,1\} \mid h_{\theta_1}(x) = \begin{cases} 1, & x \geq \theta_1 \\ 0, & otherwise \end{cases}, \theta_1 \in \mathbb{R}\right\}}^{\mathcal{H}_1} \cup \overbrace{\left\{h_{\theta_2}: \mathbb{R} \to \{0,1\} \mid h_{\theta_2}(x) = \begin{cases} 1, & x < \theta_2 \\ 0, & otherwise \end{cases}, \theta_2 \in \mathbb{R}\right\}}^{\mathcal{H}_2}$$

$$\tau_{\mathcal{H}}: \mathbb{N} \to \mathbb{N}, \qquad \tau_{\mathcal{H}}(m) = \max_{\substack{C \subseteq X \\ |C|=m}} |\mathcal{H}_C| = \max_{\substack{C \subseteq X \\ |C|=m}} |\{h: C \subseteq X \to \{0,1\} | h \in \mathcal{H}\}|$$

We have already demonstrated that $VCdim(\mathcal{H}) = 2$ in problem 1 of assignment 1. So obviously, for $m \leq 2$ we have $\tau_{\mathcal{H}}(m) = 2^m \Rightarrow \begin{cases} \tau_{\mathcal{H}}(0) = 1 \\ \tau_{\mathcal{H}}(1) = 2 \\ \tau_{\mathcal{H}}(2) = 4 \end{cases}$ (1).

If $C = \emptyset \Rightarrow \mathcal{H}_C = \{\emptyset\} \Rightarrow |\mathcal{H}_C| = 1$ $(convention, does\ not\ restrict\ the\ generality)$.

If $C = \{a\} \Rightarrow \mathcal{H}_C = \underbrace{\mathcal{H}_{1_C}}_{\{(0),(1)\}} \cup \underbrace{\mathcal{H}_{2_C}}_{\{(0),(1)\}} = \{(0),(1)\} \Rightarrow |\mathcal{H}_C| = 2, \quad a \in \mathbb{R}$.

If $C = \{a, b\} \Rightarrow \mathcal{H}_C = \underbrace{\mathcal{H}_{1_C}}_{\{(0,0),(0,1),(1,1)\}} \cup \underbrace{\mathcal{H}_{2_C}}_{\{(0,0),(1,0),(1,1)\}} = \{(0,0),(0,1),(1,0),(1,1)\} \Rightarrow |\mathcal{H}_C| = 4, \quad a, b \in \mathbb{R}$.

If $m > 2$ let us say $m = 3 \Rightarrow C = \{a, b, c\}$, $a, b, c \in \mathbb{R}$ then $\mathcal{H}_C = \underbrace{\mathcal{H}_{1_C}}_{\{(0,0,0),(0,0,1),(0,1,1),(1,1,1)\}} \cup$

$\underbrace{\mathcal{H}_{2_C}}_{\{(0,0),(1,0,0),(1,1,0),(1,1,1)\}} = \{(0,0,0),(0,0,1),(0,1,1),(1,0,0),(1,1,0),(1,1,1)\} \Rightarrow |\mathcal{H}_C| = 6 < 8 = 2^3$ , $(0,1,0)$ and $(1,0,1)$ are missing.

Let $m \in \mathbb{N} > 2 \Rightarrow C = \{a_i \in \mathbb{R} \mid i \in \overline{1,m}\} \subseteq \mathbb{R}, |C| = m \Rightarrow \forall (i \neq j \Rightarrow a_i \neq a_j)$ then $\mathcal{H}_C = \mathcal{H}_{1_C} \cup \mathcal{H}_{2_C}$, where

$$\mathcal{H}_{1_C} = \{(0,0,\dots,0),(0,0,\dots,0,1),\dots,(0,0,\dots,0,1,\dots,1),\dots,(0,1,\dots,1),(1,1,\dots,1)\}$$
$$= \left\{\left(0,0,\dots,0,\underbrace{1}_{i},\dots,1\right) | i \in \overline{1,m}\right\} \cup \{(0,0,\dots,0)\}$$

and

$$\mathcal{H}_{2_C} = \{(0,0,\dots,0),(1,0,\dots,0),\dots,(1,1,\dots,1,0,\dots,0),\dots,(1,1,\dots,1,0),(1,1,\dots,1)\}$$
$$= \left\{\left(1,1,\dots,1,\underbrace{0}_{i},\dots,0\right) | i \in \overline{1,m}\right\} \cup \{(1,1,\dots,1)\}$$

Using inclusion–exclusion principle, we have

$$|\mathcal{H}_C| = |\mathcal{H}_{1_C}| + |\mathcal{H}_{2_C}| - |\mathcal{H}_{1_C} \cap \mathcal{H}_{2_C}|$$
$$= \left|\left\{\left(0,0,\dots,0,\underbrace{1}_{i},\dots,1\right) | i \in \overline{1,m}\right\} \cup \{(0,0,\dots,0)\}\right|$$
$$+ \left|\left\{\left(1,1,\dots,1,\underbrace{0}_{i},\dots,0\right) | i \in \overline{1,m}\right\} \cup \{(1,1,\dots,1)\}\right| - |\{(0,0,\dots,0),(1,1,\dots,1)\}|$$
$$= m + 1 + m + 1 - 2 = 2m \; \forall\, C = \{a_i \in \mathbb{R} \mid i \in \overline{1,m}\} \subseteq \mathbb{R}, |C| = m > 2$$

The shattering coefficient $\tau_{\mathcal{H}}(m)$ of the growth function is

$$\tau_{\mathcal{H}}(m) = \max_{\substack{C \subseteq X \\ |C|=m}} |\mathcal{H}_C| = \max_{\substack{C \subseteq X \\ |C|=m>2}} 2m = 2m \quad (2)$$

$$\left.\begin{matrix}(1)\\(2)\end{matrix}\right| \Rightarrow \boxed{\tau_{\mathcal{H}}(m) = 2m \; \forall m \in \mathbb{N}^*}$$

**Solution problem 1.b:**

**Lemma (Sauer – Shelah – Perles)**: Let $\mathcal{H}$ be a hypothesis class with $VCdim(\mathcal{H}) \leq d < \infty$. Then, for all $m$, we have that:

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^i = \sum_{i=0}^{d} \binom{m}{i}$$

Using this lemma, for $d = VCdim(\mathcal{H}) = 2$, we get that:

$$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \sum_{i=0}^{2} \binom{m}{i} = \binom{m}{0} + \binom{m}{1} + \binom{m}{2} = \frac{m!}{m! \cdot 0!} + \frac{m!}{(m-1)! \cdot 1!} + \frac{m!}{(m-2)! \cdot 2!}$$
$$= 1 + m + \frac{m(m-1)}{2} = \frac{m^2 + m + 2}{2} = O(m^2) = O(m^d) \quad (3)$$

At the previous point, we got that $\tau_{\mathcal{H}}(m) = 2m$ (4)

$$\left.\begin{matrix}(3)\\(4)\end{matrix}\right| \Rightarrow \boxed{\underbrace{2m}_{O(m)} = \tau_{\mathcal{H}}(m) \leq \frac{m^2 + m + 2}{2} = O(m^2)} \Leftrightarrow 2m \leq \frac{m^2 + m + 2}{2}$$

We obtained that the growth function is a polynomial of first degree but the Sauer's Lemma says that it is increased by a polynomial of degree 2, which is a correct limit but too high.

**Solution problem 1.c:**

Let $\mathcal{H}_{thresholds} = \left\{h_a: \mathbb{R} \to \{0,1\} \mid h_a(x) = \begin{cases} 1, & x < a \\ 0, & otherwise \end{cases}\right\}$. We know that $VCdim(\mathcal{H}_{thresholds}) = 1$.

Using Sauer's Lemma, for $d = VCdim(\mathcal{H}_{thresholds}) = 1$, we get that:

$$\tau_{\mathcal{H}_{thresholds}}(m) \leq \sum_{i=0}^{d} \binom{m}{i} = \sum_{i=0}^{1} \binom{m}{i} = \binom{m}{0} + \binom{m}{1} = \frac{m!}{m! \cdot 0!} + \frac{m!}{(m-1)! \cdot 1!} = 1 + m \quad (5)$$

$\tau_{\mathcal{H}_{thresholds}}(1) = 2^1 = 2$ $(\mathcal{H}_{thresholds\,C} = \{(0),(1)\} \Rightarrow \left|\mathcal{H}_{thresholds\,C}\right| = 2 \;\forall\, C = \{a\}, a \in \mathbb{R})$

If $m > 1$ let us say $m = 2 \Rightarrow \mathrm{C} = \{\mathrm{a}, \mathrm{b}\},\ a, b \in \mathbb{R}$ then $\mathcal{H}_{thresholds\,C} = \{(0,0),(1,0),(1,1)\}$
$\Rightarrow \left|\mathcal{H}_{thresholds\,\mathrm{C}}\right| = 3 < 4 = 2^2$, (0,1) is missing.

Let $m \in \mathbb{N} > 1 \Rightarrow C = \left\{a_i \in \mathbb{R} \mid i \in \overline{1,m}\right\} \subseteq \mathbb{R}, |C| = m \Rightarrow\ \forall (i \neq j \Rightarrow a_i \neq a_j)$ then $\mathcal{H}_{thresholds\,C} =$
$\{(0,0,\dots,0),(1,0,\dots,0),\dots,(1,1,\dots,1,0,\dots,0),\dots,(1,1,\dots,1,0),(1,1,\dots,1)\} = \left\{\left(1,1,\dots,1,\underbrace{0}_{i},\dots,0\right) \mid i \in \overline{1,m}\right\} \cup \{(1,1,\dots,1)\} \Rightarrow \left|\mathcal{H}_{thresholds\,C}\right| = m + 1\ \forall\ C = \left\{a_i \in \mathbb{R} \mid i \in \overline{1,m}\right\} \subseteq \mathbb{R}, |C| = m\ \Rightarrow \tau_{\mathcal{H}_{thresholds}}(m) =$
$\max_{\substack{C \subseteq X \\ |C|=m}} \left|\mathcal{H}_{thresholds\,\mathrm{C}}\right| = \max_{\substack{C \subseteq X \\ |C|=m}} (\mathrm{m}+1) = \mathrm{m} + 1 \quad (6)$

$$\left.\begin{matrix}(5)\\(6)\end{matrix}\right| \Rightarrow \mathrm{m} + 1 = \tau_{\mathcal{H}}(m) \leq m + 1 \Rightarrow \boxed{\tau_{\mathcal{H}_{thresholds}}(m)\ is\ equal\ to\ the\ general\ upper\ bound}$$

**Problem 2.** Let $\Sigma$ be a finite alphabet and let $X = \Sigma^m$ be a sample space of all strings of length $m$ over $\Sigma$. Let $\mathcal{H}$ be a hypothesis space over $X$, where

$$\mathcal{H} = \left\{ h_w : \Sigma^m \to \{0,1\}, w \in \Sigma^*, 0 < |w| \le m, \quad s.t. \; h_w(x) = \begin{cases} 1, & w \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases} \right\}$$

a) Give an upper bound (any upper bound that you can come up) of the VCdimension of H in terms of $|\Sigma|$ and $m$.

b) Give an efficient algorithm for finding a hypothesis $h_w$ consistent with a training set in the realizable case. What is the complexity of your algorithm?

Example: let $\Sigma = \{a, b, c\}$, $m = 4$, and the training set $S = \{(aabc, 1), (baca, 0), (bcac, 0), (abba, 1)\}$. The output of the algorithm should be $h_{ab}$.

**Solution problem 2.a:**

We know that $VCdim(\mathcal{H}) \le \log_2|\mathcal{H}|$ (1), so all we have to do is calculate $|\mathcal{H}|$.

If $\Sigma = \{a\} \Rightarrow X = \{a\}^m = \left\{ \underbrace{aaa \dots a}_{m \text{ times}} \right\} \Rightarrow |X| = 1$

$$\mathcal{H} = \left\{ \begin{matrix} h_w : \{a\}^m \to \{0,1\} | \; w \in \{a\}^*, 0 < |w| \le m, \\ h_w(x) = \begin{cases} 1, & w \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases} \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} h_w : \{a\}^m \to \{0,1\} | \; w \in \left\{ a, aa, aaa, \dots, \underbrace{aaa \dots a}_{m \text{ times}} \right\}, \\ h_w(x) = \begin{cases} 1, & w \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases} \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} h_a, h_{aa}, h_{aaa}, \dots, h_{\underbrace{aaa \dots a}_{m \text{ times}}} | \\ h_{\underbrace{aaa \dots a}_{q \text{ times}}}(x) = \begin{cases} 1, & \underbrace{aaa \dots a}_{q \text{ times}} \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases}, q \in \overline{1, m} \end{matrix} \right\} \Rightarrow$$

$$|\mathcal{H}| = \left| \left\{ a, aa, aaa, \dots, \underbrace{aaa \dots a}_{m \text{ times}} \right\} \right| = m$$

If $\Sigma = \{a, b\} \Rightarrow X = \{a, b\}^m = \left\{ \underbrace{aaa \dots a}_{m \text{ times}}, \underbrace{aaa \dots a}_{m-1 \text{ times}} b, \underbrace{aaa \dots a}_{m-2 \text{ times}} ba, \dots, \underbrace{bbb \dots b}_{m \text{ times}} \right\} \Rightarrow |X| = 2^m$

$$\mathcal{H} = \left\{ \begin{matrix} h_w : \{a, b\}^m \to \{0,1\} | \; w \in \{a, b\}^*, 0 < |w| \le m, \\ h_w(x) = \begin{cases} 1, & w \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases} \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} h_w : \{a, b\}^m \to \{0,1\} | \; w \in \left\{ a, b, aa, ab, ba, bb, aaa, \dots, \underbrace{bbb \dots b}_{m \text{ times}} \right\}, \\ h_w(x) = \begin{cases} 1, & w \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases} \end{matrix} \right\}$$

$$= \left\{ \begin{matrix} h_a, h_b, h_{aa}, h_{ab}, h_{ba}, h_{bb}, h_{aaa}, \dots, h_{\underbrace{bbb \dots b}_{m \text{ times}}} | \\ h_{(e_i)_{i \in \overline{1,q}}}(x) = \begin{cases} 1, & (e_i)_{i \in \overline{1,q}} \text{ is a substring of } x \\ 0, & \text{otherwise} \end{cases}, q \in \overline{1, m}, e_i \in \Sigma, i \in \overline{1, q} \end{matrix} \right\} \Rightarrow$$

$$|\mathcal{H}| = \left| \left\{ a, b, aa, ab, ba, bb, aaa, \dots, \underbrace{bbb \dots b}_{m \text{ times}} \right\} \right| = 2 + 4 + 8 + \dots + 2^m = \sum_{i=1}^{m} 2^i = 2^{m+1} - 2$$

Let $|\Sigma| = s \in \mathbb{N}^*, \Sigma = \{a_i | i \in \overline{1,s}\} \Rightarrow X = \{a_i | i \in \overline{1,s}\}^m =$

$$\left\{\underbrace{a_1a_1a_1 \dots a_1}_{m\ times}, \underbrace{a_1a_1a_1 \dots a_1}_{m-1\ times}a_2, \underbrace{a_1a_1a_1 \dots a_1}_{m-2\ times}a_2a_1, \dots, \underbrace{a_sa_sa_s \dots a_s}_{m\ times}\right\} \Rightarrow |X| = s^m = |\Sigma|^m$$

$$\mathcal{H} = \left\{\begin{matrix} h_w: \{a_i | i \in \overline{1,s}\}^m \to \{0,1\} | \ w \in \{a_i | i \in \overline{1,s}\}^*, 0 < |w| \leq m, \\ h_w(x) = \begin{cases} 1, & w\ is\ a\ substring\ of\ x \\ 0, & otherwise \end{cases} \end{matrix}\right\}$$

$$= \left\{\begin{matrix} h_w: \{a_i | i \in \overline{1,s}\}^m \to \{0,1\} | \ w \in \{(e_i)_{i \in \overline{1,q}} | \ q \in \overline{1,m}, e_i \in \Sigma, i \in \overline{1,q}\}, \\ h_w(x) = \begin{cases} 1, & w\ is\ a\ substring\ of\ x \\ 0, & otherwise \end{cases} \end{matrix}\right\}$$

$$= \left\{\begin{matrix} h_{(e_i)_{i \in \overline{1,q}}} | \\ h_{(e_i)_{i \in \overline{1,q}}}(x) = \begin{cases} 1, & (e_i)_{i \in \overline{1,q}}\ is\ a\ substring\ of\ x \\ 0, & otherwise \end{cases} \end{matrix}, q \in \overline{1,m}, e_i \in \Sigma, i \in \overline{1,q}\right\} \Rightarrow$$

$$|\mathcal{H}| = \left|\{(e_i)_{i \in \overline{1,q}} | \ q \in \overline{1,m}, e_i \in \Sigma, i \in \overline{1,q}\}\right| = \left|\left\{a_1, a_2, \dots, a_s, a_1a_1, a_1a_2, \dots a_sa_s, \dots, \underbrace{a_sa_sa_s \dots a_s}_{m\ times}\right\}\right|$$

$$= s + s^2 + s^3 + \cdots + s^m = \sum_{i=1}^{m} s^i = S\ (2)$$

$$\begin{matrix} S = s + s^2 + s^3 + \cdots + s^m | \cdot s \\ sS = s^2 + s^3 + \cdots + s^m + s^{m+1} \end{matrix} \Big| - \quad \Rightarrow$$

$$S(s-1) = s^{m+1} - s$$

$$S = \frac{s^{m+1} - s}{s-1} \quad (3)$$

$$\begin{matrix} (2) \\ (3) \end{matrix} \Big| \Rightarrow \boxed{|\mathcal{H}| = \frac{s^{m+1} - s}{s-1} = \frac{|\Sigma|^{m+1} - |\Sigma|}{|\Sigma| - 1}}\ (4)$$

$$\begin{matrix} (1) \\ (4) \end{matrix} \Big| \Rightarrow \boxed{VCdim(\mathcal{H}) \leq \log_2|\mathcal{H}| = \log_2 \frac{|\Sigma|^{m+1} - |\Sigma|}{|\Sigma| - 1}}\ (5)\ \blacksquare$$

**Solution problem 2.b:**

We are in the realizable case $\Rightarrow \mathcal{H}$ is PAC learnable $\Rightarrow$

$$C_1 \frac{d + ln\frac{1}{\delta}}{\varepsilon} \leq m_{\mathcal{H}}(\varepsilon, \delta) \leq C_2 \frac{d \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon}, \qquad where\ d = VCdim(\mathcal{H}), C_1, C_2 \in \mathbb{R}_+\ (6)$$

$$(5) \Rightarrow d = VCdim(\mathcal{H}) \leq \log_2|\mathcal{H}| = \log_2 \frac{|\Sigma|^{m+1} - |\Sigma|}{|\Sigma| - 1} < \log_2(|\Sigma|^{m+1} - |\Sigma|)$$

$$< \log_2|\Sigma|^{m+1} = (m+1)\log_2|\Sigma| \Rightarrow d < (m+1)\log_2|\Sigma|\ (7)$$

$$\begin{matrix} (6) \\ (7) \end{matrix} \Big| \Rightarrow \boxed{m_{\mathcal{H}}(\varepsilon, \delta) < C_2 \frac{(m+1)\log_2|\Sigma| \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon}}\ (8) \Rightarrow$$

$m_{\mathcal{H}}(\varepsilon, \delta)$ is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}, \underbrace{d}_{<(m+1)\log_2|\Sigma|}$ (measures the complexity of the hypothesis class $\mathcal{H}$).

Consider the following algorithm $\mathcal{A}$:

0. $Input: S = \{(w_i, l_i) \mid w_i \in \Sigma^m, l_i \in \{0,1\}\}_{i\in\overline{1,T}}, |S| = T < C_2 \dfrac{(m+1)\log_2|\Sigma| \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon}$
1. $Poz = \{w| (w,1) \in S\}, \quad |Poz| = p$

   $Neg = \{w| (w,0) \in S\}, \quad |Neg| = n = T - p$
2. $If\ p = 0\ then\ Return\ h_w, |w| = m, w \notin Neg$
3. $\bowtie: \Sigma^* \to \mathcal{P}(\Sigma^*), \bowtie(s) = \{t \mid t\ is\ substring\ of\ s, 0 < |t| \leq |s|\}$

   $\mathcal{U}_i = \bowtie(Poz[i])_{i\in\overline{1,p}}$

   $\mathcal{V} = \bigcup_{w\in Neg} \bowtie(w)$
4. $\mathcal{W}_i = \mathcal{U}_i/\mathcal{V}, i \in \overline{1,p}$
5. $\mathcal{R} = \bigcap_{i=1}^{p} \mathcal{W}_i$
6. $Return\ h_{\mathcal{R}[1]} \stackrel{\text{def}}{=} h_{w_{\mathcal{A},S}}$

Let us run this algorithm on our example:

0. $Input: S = \{(aabc,1),(baca,0),(bcac,0),(abba,1)\}, \quad |S| = T = 4$
1. $Poz = \{w| (w,1) \in S\} = \{aabc, abba\} \quad |Poz| = p = 2$

   $Neg = \{w| (w,0) \in S\} = \{baca, bcac\} \quad |Neg| = n = T - p = 2$
2. $p = 2 \neq 0 \ \to 3.$
3. $\bowtie: \Sigma^* \to \mathcal{P}(\Sigma^*), \bowtie(s) = \{t \mid t\ is\ substring\ of\ s, 0 < |t| \leq |s|\}$

   $\mathcal{U}_1 = \bowtie(Poz[1]) = \bowtie(aabc) = \{a,b,c,aa,ab,bc,aab,abc,aabc\}$

   $\mathcal{U}_2 = \bowtie(Poz[2]) = \bowtie(abba) = \{a,b,ab,bb,ba,abb,bba,abba\}$

   $\mathcal{V} = \bigcup_{w\in Neg} \bowtie(w) = \bowtie(baca) \cup \bowtie(bcac)$
   $= \{b,a,c,ba,ac,ca,bac,aca,baca\} \cup \{b,c,a,bc,ca,ac,bca,cac,bcac\}$
   $= \{a,b,c,ac,ba,bc,ca,aca,bac,bca,cac,baca,bcac\}$
4. $\mathcal{W}_1 = \mathcal{U}_1/\mathcal{V} = \{aa,ab,aab,abc,aabc\}$

   $\mathcal{W}_2 = \mathcal{U}_2/\mathcal{V} = \{ab,bb,abb,bba,abba\}$
5. $\mathcal{R} = \bigcap_{i=1}^{p} \mathcal{W}_i = \{aa,ab,aab,abc,aabc\} \cap \{ab,bb,abb,bba,abba\} = \{ab\}$
6. $Return\ h_{\mathcal{R}[1]} = h_{ab}$

Furthermore, $h_{w_{\mathcal{A},S}}$ is ERM ($L_S\left(h_{w_{\mathcal{A},S}}\right) = 0$, because we are in the realizable case). We will demonstrate this below.

**Case1:** $\forall\ (w,1) \in S$ we have to prove that $w_{\mathcal{A},S}\ is\ substring\ of\ w$. Let $(w,1) \in S$.

$$w_{\mathcal{A},S} = \mathcal{R}[1] \Rightarrow w_{\mathcal{A},S} \in \mathcal{R} \Rightarrow w_{\mathcal{A},S} \in \mathcal{W}_i = \mathcal{U}_i/\mathcal{V}\,, \forall i \in \overline{1,p} \Rightarrow \begin{cases} w_{\mathcal{A},S} \in \mathcal{U}_i = \bowtie (Poz[i]) \;\; \forall i \in \overline{1,p} \\ w_{\mathcal{A},S} \notin \mathcal{V} \;\; = \bigcup_{w \in Neg} \bowtie (w) \end{cases} \Rightarrow$$

$$\Rightarrow \left.\begin{cases} w_{\mathcal{A},S}\ is\ substring\ of\ Poz[i]\ \forall i \in \overline{1,p} \\ w_{\mathcal{A},S}\ is\ not\ substring\ of\ Neg[i]\ \forall i \in \overline{1,n} \\ (w,1) \in S \end{cases}\right| \Rightarrow w_{\mathcal{A},S}\ is\ substring\ of\ w \Rightarrow h_{w_{\mathcal{A},S}}(w) = 1 \Rightarrow$$

$$\Rightarrow \left|\left\{ i \mid i \in \overline{1,T}, h_{w_{\mathcal{A},S}}(w_i) = 0, l_i = 1\,, (w_i, l_i) \in S \right\}\right| = |\emptyset| = 0 \quad (9)$$

**Case 2:** $\forall\,(w,0) \in S$ we have to prove that $w_{\mathcal{A},S}\ is\ not\ substring\ of\ w$. Let $(w,0) \in S$.

$$w_{\mathcal{A},S} = \mathcal{R}[1] \Rightarrow w_{\mathcal{A},S} \in \mathcal{R} \Rightarrow w_{\mathcal{A},S} \in \mathcal{W}_i = \mathcal{U}_i/\mathcal{V}\,, \forall i \in \overline{1,p} \Rightarrow \begin{cases} w_{\mathcal{A},S} \in \mathcal{U}_i = \bowtie (Poz[i]) \;\; \forall i \in \overline{1,p} \\ w_{\mathcal{A},S} \notin \mathcal{V} \;\; = \bigcup_{w \in Neg} \bowtie (w) \end{cases} \Rightarrow$$

$$\Rightarrow \left.\begin{cases} w_{\mathcal{A},S}\ is\ substring\ of\ Poz[i]\ \forall i \in \overline{1,p} \\ w_{\mathcal{A},S}\ is\ not\ substring\ of\ Neg[i]\ \forall i \in \overline{1,n} \\ (w,0) \in S \end{cases}\right| \Rightarrow w_{\mathcal{A},S}\ is\ not\ substring\ of\ w \Rightarrow h_{w_{\mathcal{A},S}}(w) = 0 \Rightarrow$$

$$\Rightarrow \left|\left\{ i \mid i \in \overline{1,T}, h_{w_{\mathcal{A},S}}(w_i) = 1, l_i = 0\,, (w_i, l_i) \in S \right\}\right| = |\emptyset| = 0 \quad (10)$$

$$\left.\begin{matrix} (9) \\ (10) \end{matrix}\right| \Rightarrow L_S(h_S) = \frac{\left|\left\{ i \mid i \in \overline{1,m}, h_{w_{\mathcal{A},S}}(w_i) \neq l_i, (w_i, l_i) \in S \right\}\right|}{T} = \frac{|\emptyset| + |\emptyset|}{T} = 0 \Rightarrow h_{w_{\mathcal{A},S}}\ is\ ERM$$

The complexity of the algorithm $\mathcal{A}$:

1. $Poz = \{w|\ (w,1) \in S\} = \{aabc, abba\} \quad |Poz| = p = 2$
   $Neg = \{w|\ (w,0) \in S\} = \{baca, bcac\} \quad |Neg| = n = T - p = 2$

Dividing the train set into positive and negative examples - $O(T)$

2. $If\ p = 0\ \ then\ Return\ h_w, |w| = m, w \notin Neg$

If an artificial classifier is issued that label, everything with zero then the complexity is $O(1)$. Alternatively, you can generate this $w$ with a randomized algorithm that at one-step generates a random character from $\Sigma$ for each of the characters. The final error is $\left(\frac{T}{|\Sigma|^m}\right)^K <$ $\left(\frac{C_2 \frac{(m+1)\log_2|\Sigma| \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon}}{|\Sigma|^m}\right)^K$, where $K \in \mathbb{N}^*$ the number of steps is $O(K \cdot m \cdot T)$.

3. $\bowtie: \Sigma^* \to \mathcal{P}(\Sigma^*), \bowtie (s) = \{t \mid t\ is\ substring\ of\ s, 0 < |t| \leq |s|\ \}$
   $\mathcal{U}_i = \bowtie (Poz[i])_{i \in \overline{1,p}}$

Generating all sub strings for positive strings - $O(p \cdot m^2) = O(T \cdot m^2)$

$$\mathcal{V} = \bigcup_{w \in Neg} \bowtie (w)$$

Generating all sub strings for negative strings $O(n \cdot m^2) = O(T \cdot m^2)$ and concatenate this strings - $O\left(\sum_{w \in Neg} \bowtie (w)\right) = O(n \cdot m^2) = O(T \cdot m^2)$

4. $\mathcal{W}_i = \mathcal{U}_i / \mathcal{V}, i \in \overline{1,p}$

Eliminate the substrings of negative examples from each set of positive example substrings - $O\left(\sum_{i=1}^{p} |\mathcal{U}_i|\right) = O(p \cdot m^2) = O(T \cdot m^2)$

5. $\mathcal{R} = \bigcap_{i=1}^{p} \mathcal{W}_i$

Calculates the substrings of positive examples (which are common to all positive examples) that are not in the negative ones $O((p-1) \cdot m^2 \cdot m^2) = O(T \cdot m^4)$ (is $O(T \cdot m^2)$ in average case)

6. $Return\ h_{\mathcal{R}[1]} \stackrel{\text{def}}{=} h_{w_{\mathcal{A},S}}$

Returns one of these results. As we are in RA case it results that there will be in the worst case at least one element in $\mathcal{R}$ - $O(1)$

The complexities of the basic operations were taken from here.

In the end the complexity of the algorithm is:

$$\underbrace{O(T \cdot m^4)}_{O(m_{\mathcal{H}}(\varepsilon,\delta) \cdot m^4)} \subseteq O\left(C_2 \frac{(m+1)\log_2|\Sigma| \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon} \cdot m^4\right) = O\left(C \frac{m^5 \cdot \log_2|\Sigma| \cdot ln\frac{1}{\varepsilon} + ln\frac{1}{\delta}}{\varepsilon}\right)$$

$h_{w_{\mathcal{A},S}}$ can be implemented in time which is polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}, \underbrace{d}_{<(m+1)\log_2|\Sigma|}$ (measures the complexity of the $\mathcal{H}$) therefore we have efficient learning.

**Problem 3.** Consider the boosting algorithm described (page 4) in the article "Rapid object detection using a boosted cascade of simple features", P. Viola and M. Jones, CVPR 2001. Consider that the number of positives is equal with the number of negative examples ($l = m$).

a) Prove that the distribution $w_{t+1}$ obtained at round $t + 1$ based on the algorithm described in the article is the same with the distribution $D^{(t+1)}$ obtained based on the procedure described in lecture 11 (slides 10-12).
b) Prove that the two final classifiers (the one described in the article and the one described in the lecture) are equivalent.
c) Assume that at each iteration $t$ of AdaBoost, the weak learner returns a hypothesis $h_t$ for which the error $\varepsilon_t$ satisfies $\varepsilon_t \leq 1/2-\gamma, \gamma > 0$. What is the probability that the classifier $h_t$ (selected as the best weak learner at iteration $t$) will be selected again at iteration $t + 1$? Justify your answer.

**Solution problem 3.a:**

To prove that the two distributions are equivalent we will use the principle of mathematical induction. But first, let us consider in both situations (lecture and article) the number of examples in the train set equal to $n$.

We will first check that the two distributions are equivalent at the starting step, i.e. step 1 (verification step).

In article:

$$\begin{cases} l = m \\ l + m = n \end{cases} \Rightarrow \begin{cases} m = \frac{n}{2} \\ l = \frac{n}{2} \end{cases}$$

$$w_{1,i} = \begin{cases} \frac{1}{2m}, & y_i = 0 \\ \frac{1}{2l}, & y_i = 1 \end{cases} = \begin{cases} \frac{1}{2\frac{n}{2}}, & y_i = 0 \\ \frac{1}{2\frac{n}{2}}, & y_i = 1 \end{cases} = \begin{cases} \frac{1}{n}, & y_i = 0 \\ \frac{1}{n}, & y_i = 1 \end{cases} = \frac{1}{n} (1)$$

In lecture:

$$D^{(1)}(i) = \frac{1}{n} (2)$$

$$\left.\begin{matrix} (1) \\ (2) \end{matrix}\right| \Rightarrow w_{1,i} = D^{(1)}(i) \quad (3)$$

Next to the induction step, we will assume true $w_{t,i} = D^{(t)}(i)\ \forall i \in \overline{1,n}$ (4) and we will have to prove $w_{t+1,i} = D^{(t+1)}(i)\ \forall i \in \overline{1,n}$. Without restricting the generality, for the symmetry of the two algorithms we will consider that the normalization of the distribution in the article will take place at the end of the round (so that at the beginning of a new turn the sum of the probabilities in D will make 1) and not at the beginning of a new round. This does not change the algorithm in any way.

In article:

$$w_{t+1,j} = \begin{cases} w_{t,j} \cdot \beta_t^1 = \dfrac{w_{t,j}\dfrac{\varepsilon_t}{1-\varepsilon_t}}{S_{t+1}}, if\ h_t(x_j) = y_j \\ w_{t,j} \cdot \beta_t^0 = \dfrac{w_{t,j}}{S_{t+1}} \qquad , if\ h_t(x_j) \neq y_j \end{cases} \quad \forall j \in \overline{1,n}\ where$$

$$S_{t+1} = \sum_{i=1}^{n} w_{t,i} \left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{1-|h_t(x_i)-y_i|}$$

$$\varepsilon_t = \sum_{i=1}^{n} w_{t,i}|h_t(x_i) - y_i| = \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} w_{t,i}$$

In lecture:

$$D^{(t+1)}(j) = \begin{cases} \dfrac{D^{(t)}(j)\ \cdot e^{-w_t}}{Z_{t+1}} = \dfrac{D^{(t)}(j)\ \cdot \sqrt{\dfrac{\varepsilon_t}{1-\varepsilon_t}}}{Z_{t+1}}, if\ h_t(x_j) = y_j \\ \dfrac{D^{(t)}(j)\ \cdot e^{w_t}}{Z_{t+1}} = \dfrac{D^{(t)}(j)\ \cdot \sqrt{\dfrac{1-\varepsilon_t}{\varepsilon_t}}}{Z_{t+1}}, if\ h_t(x_j) \neq y_j \end{cases} \quad \forall j \in \overline{1,n}$$

$$Z_{t+1} = \sum_{i=1}^{n} D^{(t)}(i)\ \cdot e^{-w_t h_t(x_i) y_i}$$

$$\varepsilon_t = \Pr_{i\sim D^{(t)}}[h_t(x_i) \neq y_i] = \sum_{i=1}^{n} D^{(t)}(i) \cdot \mathbf{1}_{[h_t(x_i)\neq y_i]} = \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} D^{(t)}(i)$$

Obviously, we can see that the sum of the probabilities in each distribution (article and course) at the end of each the round is one.

$$\sum_{i=1}^{n} D^{(k)}(i) = 1 = \sum_{i=1}^{n} w_{k,i}\ \forall\, k \in \mathbb{N}^*$$

In article $\varepsilon_t = \sum^{n}_{\substack{i=1 \\ h_t(x_i)\neq y_i}} w_{t,i} = \sum^{n}_{\substack{i=1 \\ h_t(x_i)\neq y_i}} D^{(t)}(i) = \varepsilon_t$ from lecture because we have $w_{t,i} = D^{(t)}(i)$ from induction hypothesis (4).

$$S_{t+1} = \sum_{i=1}^{n} w_{t,i} \left(\frac{\varepsilon_t}{1-\varepsilon_t}\right)^{1-|h_t(x_i)-y_i|} = \sum_{\substack{i=1 \\ h_t(x_i)= y_i}}^{n} \left(w_{t,i}\frac{\varepsilon_t}{1-\varepsilon_t}\right) + \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} w_{t,i}$$

$$= \frac{\varepsilon_t}{1-\varepsilon_t} \sum_{\substack{i=1 \\ h_t(x_i)= y_i}}^{n} w_{t,i} + \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} w_{t,i} = \frac{\varepsilon_t}{1-\varepsilon_t}(1-\varepsilon_t) + \varepsilon_t = 2\varepsilon_t$$

$$Z_{t+1} = \sum_{i=1}^{n} D^{(t)}(i) \cdot e^{-w_t h_t(x_i) y_i} = \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{n} D^{(t)}(i)\, e^{-w_t} + \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} D^{(t)}(i)\, e^{w_t}$$

$$= \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{n} \left( D^{(t)}(i) \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \right) + \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} \left( D^{(t)}(i) \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \right)$$

$$= \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \sum_{\substack{i=1 \\ h_t(x_i)=y_i}}^{n} D^{(t)}(i) + \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \sum_{\substack{i=1 \\ h_t(x_i)\neq y_i}}^{n} D^{(t)}(i) = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}(1-\varepsilon_t) + \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}\varepsilon_t$$

$$= \sqrt{\varepsilon_t(1-\varepsilon_t)} + \sqrt{\varepsilon_t(1-\varepsilon_t)} = 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$$

So, <u>in article:</u>

$$w_{t+1,j} = \begin{cases} w_{t,j} \cdot \beta_t^1 = \dfrac{w_{t,j}\frac{\varepsilon_t}{1-\varepsilon_t}}{S_{t+1}} = \dfrac{w_{t,j}\frac{\varepsilon_t}{1-\varepsilon_t}}{2\varepsilon_t} = w_{t,j}\dfrac{1}{2(1-\varepsilon_t)}, if\ h_t(x_j) = y_j \\ w_{t,j} \cdot \beta_t^0 = \dfrac{w_{t,j}}{S_{t+1}} = w_{t,j}\dfrac{1}{2\varepsilon_t} \qquad\qquad , if\ h_t(x_j) \neq y_j \end{cases} \quad \forall j \in \overline{1,n} \ (5)$$

in addition, <u>in lecture:</u>

$$D^{(t+1)}(j) = \begin{cases} \dfrac{D^{(t)}(j) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{Z_{t+1}} = \dfrac{D^{(t)}(j) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = D^{(t)}(j)\dfrac{1}{2(1-\varepsilon_t)}, if\ h_t(x_j) = y_j \\ \dfrac{D^{(t)}(j) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{Z_{t+1}} = \dfrac{D^{(t)}(j) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = D^{(t)}(j)\dfrac{1}{2\varepsilon_t}, \qquad if\ h_t(x_j) \neq y_j \end{cases} \quad \forall j \in \overline{1,n} \ (6)$$

$$\left.\begin{matrix}(5)\\(6)\end{matrix}\right| \Rightarrow w_{t+1,j} = \begin{cases} w_{t,j}\dfrac{1}{2(1-\varepsilon_t)}, if\ h_t(x_j) = y_j \\ w_{t,j}\dfrac{1}{2\varepsilon_t} \qquad if\ h_t(x_j) \neq y_j \end{cases} (4) = \begin{cases} D^{(t)}(j)\dfrac{1}{2(1-\varepsilon_t)}, if\ h_t(x_j) = y_j \\ D^{(t)}(j)\dfrac{1}{2\varepsilon_t}, \qquad if\ h_t(x_j) \neq y_j \end{cases} = D^{(t+1)}(j)\ \forall j \in \overline{1,n}$$

Which is exactly what we wanted to prove. So according to the principle of mathematics we have that $w_{t,i} = D^{(t)}(i)\ \forall i \in \overline{1,n}\ \forall t \in \mathbb{N}^* \Leftrightarrow w_t = D^{(t)}\ \ \forall t \in \mathbb{N}^*$ ■

**Solution problem 3.b:**

In article:

$$h(x) = \begin{cases} 1, & \sum_{t=1}^{T} \alpha_t h_t(x) \geq \frac{1}{2}\sum_{t=1}^{T} \alpha_t \\ 0, & otherwise \end{cases}$$

$$= \begin{cases} 1, & \sum_{t=1}^{T} \alpha_t h_t(x) - \frac{1}{2}\sum_{t=1}^{T} \alpha_t \geq 0 \\ 0, & otherwise \end{cases} = \begin{cases} 1, & \sum_{t=1}^{T} \left(\alpha_t h_t(x) - \frac{1}{2}\alpha_t\right) \geq 0 \\ 0, & otherwise \end{cases}$$

$$= \begin{cases} 1, & \sum_{t=1}^{T} \alpha_t \left(h_t(x) - \frac{1}{2}\right) \geq 0 \\ 0, & otherwise \end{cases} = \begin{cases} 1, & \frac{1}{2}\sum_{t=1}^{T} \alpha_t (2h_t(x) - 1) \geq 0 \\ 0, & otherwise \end{cases}$$

$$= \begin{cases} 1, & \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)(2h_t(x) - 1) \geq 0 \\ 0, & otherwise \end{cases} =$$

$$= \begin{cases} 1, & \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\Delta_t(x) \geq 0 \\ 0, & otherwise \end{cases}, where\ \Delta_t(x) = (2h_t(x) - 1)\ (7)$$

In lecture:

$$h(x) = sign\left(\sum_{t=1}^{T} w_t h_t(x)\right) = \begin{cases} 1, & \sum_{t=1}^{T} w_t h_t(x) \geq 0 \\ -1, & otherwise \end{cases} = \begin{cases} 1, & \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) h_t(x) \geq 0 \\ -1, & otherwise \end{cases}$$

$$= \begin{cases} 1, & \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\nabla_t(x) \geq 0 \\ -1, & otherwise \end{cases}, where\ \nabla_t(x) = h_t(x)\ (8)$$

Next, it remains to prove that $\Delta_t(x) = \nabla_t(x)\ \forall t \in \overline{1,T}, \forall x$.

As an observation, the classifiers in the article have as labels the numbers 0 for negative examples and 1 for positive examples. While the classifiers in the course have as labels -1 and 1 for negative and positive examples, respectively.

$$\Delta_t(x) = (2h_t(x) - 1) = \begin{cases} 2 - 1 = 1, & h_t(x) = 1\ (x\ is\ classified\ as\ a\ positive\ example) \\ 0 - 1 = -1, & h_t(x) = 0\ (x\ is\ classified\ as\ a\ negative\ example) \end{cases} (9)$$

$$\nabla_t(x) = h_t(x) = \begin{cases} 1, & h_t(x) = 1\ (x\ is\ classified\ as\ a\ positive\ example) \\ -1, & h_t(x) = -1\ (x\ is\ classified\ as\ a\ negative\ example) \end{cases} (10)$$

$$\left.\begin{matrix}(7)\\(8)\\(9)\\(10)\end{matrix}\right| \Rightarrow \Delta_t(x) = \nabla_t(x)\ \forall t \in \overline{1,T}, \forall x \Rightarrow \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\Delta_t(x) = \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\nabla_t(x) \Rightarrow$$

$$\left[\frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\Delta_t(x) \geq 0 \Leftrightarrow \frac{1}{2}\sum_{t=1}^{T} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)\nabla_t(x) \geq 0\right] \Rightarrow \begin{cases} h_{article}(x) = 1 \Leftrightarrow h_{lecture}(x) = 1 \\ h_{article}(x) = 0 \Leftrightarrow h_{lecture}(x) = -1 \end{cases}$$

$$\Rightarrow h_{article} \Leftrightarrow h_{lecture}\ \blacksquare$$

**Solution problem 3.c:**

For this point, we will use version of lecture#11 of the AdaBoost algorithm.

We know that error of $h_k$ is $\varepsilon_k \leq \frac{1}{2} - \gamma, \gamma > 0, \forall k \in \overline{1,T} \Rightarrow \begin{cases} \varepsilon_t \leq \frac{1}{2} - \gamma \\ \varepsilon_{t+1} \leq \frac{1}{2} - \gamma \end{cases}, \gamma > 0.$

If $\varepsilon_t = 0 \Rightarrow \Pr_{i \sim D^{(t)}}[h_t(x_i) \neq y_i] = \sum_{i=1}^{n} D^{(t)}(i) \cdot \mathbf{1}_{[h_t(x_i) \neq y_i]} = \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{n} D^{(t)}(i) = 0 \Rightarrow \{i | h_t(x_i) \neq y_i\} = \emptyset$

$\forall i \in \overline{1,n}$ $x_i$ is correctly classified then $h_t(x_i) = y_i$ so at the next iteration $t+1$ its importance (probability distribution) will be:

$$D^{(t+1)}(i) = \frac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{\underbrace{Z_{t+1}}_{\sum_{i=1}^{n} D^{(t)}(i) \cdot e^{-w_t h_t(x_i) y_i} = \sum_{i=1}^{n} D^{(t)}(i) \cdot e^{-w_t} = \sum_{i=1}^{n} \left( D^{(t)}(i) \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \right) = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \sum_{i=1}^{n} D^{(t)} = \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}} = \frac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{\sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}$$

$$= D^{(t)}(i) \ \ \forall i \in \overline{1,n}$$

$$\varepsilon_{t+1} = \sum_{i=1}^{n} D^{(t+1)}(i) \cdot \mathbf{1}_{[h_{t+1}(x_i) \neq y_i]} = \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t+1)}(i) = \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t)}(i) \xrightarrow{h_{t+1} = h_t}$$

$$\sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t)}(i) = \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{n} D^{(t)}(i) = \ \varepsilon_t = 0$$

therefore, the probabilities of the next step do not change $\Rightarrow Pr[h_{t+1} = h_t | \ \varepsilon_t = 0] = 100\%$ (14).

If

$$\varepsilon_t \to \frac{1}{2} \Rightarrow \Pr_{i \sim D^{(t)}}[h_{t+1}(x_i) \neq y_i] = \sum_{i=1}^{n} D^{(t)}(i) \cdot \mathbf{1}_{[h_{t+1}(x_i) \neq y_i]} = \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t)}(i) \to \frac{1}{2} \Rightarrow \sum_{\substack{i=1 \\ h_t(x_i) = y_i}}^{n} D^{(t)}(i) \to \frac{1}{2}$$

that is, the classifier $h_t$ is equivalent to tossing the coin, so at the next iteration $t+1$ the importance (probability distribution) of an example will be:

$$D^{(t+1)}(i) = \frac{D^{(t)}(i) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{\underbrace{Z_{t+1}}_{\sum_{i=1}^{n} D^{(t)}(i) \cdot e^{-w_t h_t(x_i) y_i} = \sum_{i=1}^{n} D^{(t)}(i) \cdot e^{-\frac{1}{2}\ln(1)} = \sum_{i=1}^{n} \left( D^{(t)}(i) e^{0} \right) = \sum_{i=1}^{n} D^{(t)} = 1}} \to \frac{D^{(t)}(i) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{1}$$

$$\to D^{(t)}(i) \cdot \sqrt{\frac{1-\frac{1}{2}}{\frac{1}{2}}} = D^{(t)}(i) \ \ \forall i \ s.t. h_t(x_i) \neq y_i$$

$$D^{(t+1)}(i) = \frac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{\underbrace{Z_{t+1}}_{\sum_{i=1}^{n} D^{(t)}(i) \cdot e^{w_t h_t(x_i) y_i} = \sum_{i=1}^{n} D^{(t)}(i) \cdot e^{\frac{1}{2}\ln(1)} = \sum_{i=1}^{n} \left(D^{(t)}(i) e^{0}\right) = \sum_{i=1}^{n} D^{(t)} = 1}} \rightarrow \frac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{1}$$

$$\rightarrow D^{(t)}(i) \cdot \sqrt{\frac{\frac{1}{2}}{1-\frac{1}{2}}} = D^{(t)}(i) \;\; \forall i \; s.t. \; h_t(x_i) = y_i$$

If the same classifier $h_t$ is chosen at step $t+1$, we have

$$\varepsilon_{t+1} = \sum_{i=1}^{n} D^{(t+1)}(i) \cdot \mathbf{1}_{[h_{t+1}(x_i) \neq y_i]} = \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t+1)}(i) \rightarrow \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{n} D^{(t)}(i) = \varepsilon_t = \frac{1}{2}$$

which means that if the best one-feature classifier yields 50% accuracy then all single-feature classifiers have 50% accuracy. So the classifier $h_t$ was chosen randomly from the existing ones $\Rightarrow Pr\left[h_{t+1} = h_t \middle| \; \varepsilon_t \rightarrow \frac{1}{2}\right] = \frac{1}{\#features}$ (the minimum possible) (11).

But in the hypothesis it is specified that $\exists h_t \; s.t. \; \varepsilon_t \leq \frac{1}{2} - \gamma \; \forall t \in \mathbb{N}$ (12)

$$\left.\begin{matrix}(11) \\ (12)\end{matrix}\right| \Rightarrow Pr\left[h_{t+1} = h_t \middle| \; \varepsilon_t \rightarrow \frac{1}{2}\right] = 0\%$$

In the general case, if $\varepsilon_t \epsilon \left(0, \frac{1}{2}\right) \Rightarrow \gamma \epsilon \left(0, \frac{1}{2}\right)$

$$\begin{cases} D^{(t+1)}(i) = \dfrac{D^{(t)}(i) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{\underbrace{Z_{t+1}}_{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}} = \dfrac{D^{(t)}(i) \cdot \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = D^{(t)}(i)\dfrac{1}{2\varepsilon_t} & \forall i \; s.t. \; h_t(x_i) \neq y_i \\ D^{(t+1)}(i) = \dfrac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{\underbrace{Z_{t+1}}_{2\sqrt{\varepsilon_t(1-\varepsilon_t)}}} = \dfrac{D^{(t)}(i) \cdot \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}}}{2\sqrt{\varepsilon_t(1-\varepsilon_t)}} = D^{(t)}(i)\dfrac{1}{2(1-\varepsilon_t)} & \forall i \; s.t. \; h_t(x_i) = y_i \end{cases}$$

$$\begin{aligned} \varepsilon_{t+1} &= \sum_{i=1}^{n} D^{(t+1)}(i) \cdot \mathbf{1}_{[h_{t+1}(x_i) \neq y_i]} \\ &= \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t+1)}(i) = \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t)}(i)\frac{1}{2\varepsilon_t} = \frac{1}{2\varepsilon_t} \sum_{\substack{i=1 \\ h_{t+1}(x_i) \neq y_i}}^{n} D^{(t)}(i) \overset{h_{t+1}=h_t}{\Longrightarrow} \varepsilon_{t+1} \\ &= \frac{1}{2\varepsilon_t} \sum_{\substack{i=1 \\ h_t(x_i) \neq y_i}}^{n} D^{(t)}(i) = \frac{1}{2\varepsilon_t}\varepsilon_t = \frac{1}{2} \; (13) \end{aligned}$$

Because the examples misclassified by $h_t$ are also misclassified by $h_{t+1}$ but in round $t+1$, these examples have a higher weight than in the last round.

$$\left.\begin{matrix}(11)\\(13)\end{matrix}\right| \Rightarrow Pr[h_{t+1} = h_t] = 0\% \ (15)$$

$$\left.\begin{matrix}(14)\\(15)\end{matrix}\right| \Rightarrow \boxed{Pr[h_{t+1} = h_t] = \begin{cases}100\% & ,\varepsilon_t = 0\\ 0\% & ,otherwise\end{cases}}$$

**Bonus Problem**. Consider $\mathcal{H}_{2DNF}^{d}$ the class of 2-term disjunctive normal form formulae consisting of hypothesis of the form $h: \{0,1\}^d \rightarrow \{0,1\}$,

$$h(x) = A_1(x) \vee A_2(x),$$

where $A_i(x)$ is a Boolean conjunction of literals (in$\mathcal{H}_{conj}^{d}$). It is known that the class $\mathcal{H}_{2DNF}^{d}$ is not efficient properly learnable but can be learned improperly considering the class $\mathcal{H}_{2CNF}^{d}$. Give an $\gamma - weak - learner$ algorithm for learning the class $\mathcal{H}_{2DNF}^{d}$, which is not a stronger PAC learning algorithm for $\mathcal{H}_{2DNF}^{d}$ (like the one considering $\mathcal{H}_{2CNF}^{d}$). Prove that this algorithm is an $\gamma - weak - learner$ algorithm for $\mathcal{H}_{2DNF}^{d}$.

*Hint: Find an algorithm that returns* $h(x) = 0$ *or the disjunction of two literals.*

**Solution Bonus Problem:**

Pass