AdaBoost (Adaptive Boosting)

Fix training samples \( S = \{(x^1, c(x^1)), \ldots, (x^m, c(x^m))\} \) (independent samples from \( \text{EX}(c, \mathcal{D}) \))

Fix current distribution \( \mathcal{D}_t \) over \( S \)

Suppose current hypothesis \( h_t \) has error \( \varepsilon \leq \frac{1}{2} - \gamma \) under \( \mathcal{D}_t \)

Question: What should updated distribution \( \mathcal{D}_{t+1} \) be?

\( \mathcal{D}_{t+1} \) should force weak learner \( A \) to output hypothesis \( h_{t+1} \) to reveal information not available in \( h_t \)

**Key idea:** Make old hypothesis \( h_t \) have error exactly \( 1/2 \) under \( \mathcal{D}_{t+1} \)

Since \( A \) outputs hypothesis with advantage \( \gamma > 0 \) under any distribution, including \( \mathcal{D}_{t+1} \)

\( h_{t+1} \) is guaranteed to carry new information

Since \( h_t \) errs on \( \varepsilon \) prob. mass and is correct on \( 1 - \varepsilon \) prob. mass under \( \mathcal{D}_t \),

- Multiply weight of every sample \( h_t \) errs by \( \sqrt{\frac{1 - \varepsilon}{\varepsilon}}/Z \) (raised)
- Multiply weight of every sample \( h_t \) is correct by \( \sqrt{\frac{\varepsilon}{1 - \varepsilon}}/Z \) (reduced)

\( Z \) = normalization constant to keep total mass of new \( \mathcal{D}_{t+1} \) at 1

Total mass that \( h_t \) errs on under \( \mathcal{D}_{t+1} \) = \( \varepsilon \sqrt{\frac{1 - \varepsilon}{\varepsilon}}/Z = \sqrt{\varepsilon(1 - \varepsilon)}/Z \)

Total mass that \( h_t \) is correct on under \( \mathcal{D}_{t+1} \) = \( (1 - \varepsilon) \sqrt{\frac{\varepsilon}{1 - \varepsilon}}/Z \) (same!)

Hence \( \sqrt{\varepsilon(1 - \varepsilon)}/Z = 1/2 \iff Z = 2\sqrt{\varepsilon(1 - \varepsilon)} \)

**Multiplicative weight update algorithm, like Weighted Majority**

- Raise weight of samples \( x^i \) that current hypothesis errs on
- Reduce weight of samples \( x^i \) that current hypothesis already good at

<table>
<thead>
<tr>
<th>Weighted Majority</th>
<th>AdaBoost</th>
</tr>
</thead>
<tbody>
<tr>
<td>i-th expert, 1 ≤ i ≤ m</td>
<td>i-th sample, 1 ≤ i ≤ m</td>
</tr>
<tr>
<td>t-th round</td>
<td>t-th run of weak PAC algorithm ( A )</td>
</tr>
<tr>
<td>prediction of i-th expert in round t</td>
<td>( h_t(x^i) )</td>
</tr>
<tr>
<td>weight of i-th expert in round t</td>
<td>( \mathcal{D}_t(x^i) )</td>
</tr>
</tbody>
</table>

**Question:** How to combine \( h_1, \ldots, h_R \) into final hypothesis \( h \) ?

(Wighted) majority vote!

To simplify calculations, suppose \( h_t : X \to \{-1, +1\} \) (as opposed to \( \{0, 1\} \))

Also assume labels \( y^i \in \{-1, +1\} \) (as opposed to \( \{0, 1\} \))

Define \( \text{sign} : \mathbb{R} \to \{-1, 1\} \) as \( \text{sign}(z) = 1 \) if \( z \geq 0 \) and \( \text{sign}(z) = -1 \) if \( z < 0 \)

Output hypothesis \( h(x) = \text{sign}(\sum_{1 \leq i \leq R} \alpha_i h_t(x)) \) for some positive weights \( \alpha_i > 0 \)

Let \( f(x) = \sum_{1 \leq i \leq R} \alpha_i h_t(x) \) so that \( h(x) = \text{sign}(f(x)) \)

**AdaBoost**

- Draw independent training samples \( S = \{(x^1, y^1), \ldots, (x^m, y^m)\} \) from \( \text{EX}(c, \mathcal{D}) \)
- Initially set \( \mathcal{D}_1 = \text{uniform distribution over } S \)
- Repeat \( t = 1, \ldots, R \) times:
  - Run \( A \) on samples from \( \text{EX}(c, \mathcal{D}_t) \) to get hypothesis \( h_t \)
  - Compute \( \varepsilon_t = \text{err}_{\mathcal{D}_t}(h, c) \) (empirical error under \( \mathcal{D}_t \))
  - Set \( \alpha_t = \frac{1}{2} \ln \frac{1 - \varepsilon_t}{\varepsilon_t} \) and \( Z_t = 2\sqrt{\varepsilon_t(1 - \varepsilon_t)} \)
  - Update \( \mathcal{D}_{t+1}(x^i) = \mathcal{D}_t(x^i) \cdot \exp(-\alpha_t h_t(x^i)y^i)/Z_t \)
  - Set \( f(x) = \sum_{1 \leq i \leq k} \alpha_i h_t(x) \) and output hypothesis \( h(x) = \text{sign}(f(x)) \)

If \( h_t(x^i) = y^i \) (correct), then \( h_t(x^i)y^i = 1, \exp(-\alpha_t h_t(x^i)y^i) = \exp(-\alpha) = \sqrt{\frac{\varepsilon_t}{1 - \varepsilon_t}} \) (reduced)

If \( h_t(x^i) \neq y^i \) (mistake), then \( h_t(x^i)y^i = -1, \exp(-\alpha_t h_t(x^i)y^i) = \exp(\alpha) = \frac{1 - \varepsilon_t}{\varepsilon_t} \) (raised)

**Claim:** \( \frac{1}{m}\sum_{1 \leq i \leq m} |h_t(x^i) \neq y^i| = \frac{1}{m}\sum_{1 \leq i \leq m} 1(y^i f(x^i) \leq 0) \leq \frac{1}{m}\sum_{1 \leq i \leq m} \exp(-y^i f(x^i)) \)

Reason: \( 1(z \leq 0) \leq \exp(-z) \) for any \( z \in \mathbb{R} \)
Claim: \( \frac{1}{m} \sum_{1 \leq i \leq m} \exp(-y^i f(x^i)) = Z_1 Z_2 \cdots Z_R \)

Reason: \( D_{R+1}(x^i) = \exp(-\alpha_R h_R(x^i)y^i) \frac{Z_R}{D_R(x^i)} = (\text{keep expanding } D_R, \ldots, D_2) \)
\[
= \frac{\exp(-\alpha_R h_R(x^i)y^i)}{Z_R} \cdots \frac{\exp(-\alpha_1 h_1(x^i)y^i)}{Z_1} D_1(x^i)
\]

Sum over all \( x^i \), using \( D_1(x^i) = \frac{1}{m} \) and \( D_{R+1} \) has total mass 1,
\[
1 = \frac{1}{m} \sum_{1 \leq i \leq m} \frac{\exp(-\alpha_R h_R(x^i)y^i)}{Z_R} \cdots \frac{\exp(-\alpha_1 h_1(x^i)y^i)}{Z_1}
\]
\[
Z_1 \cdots Z_R = \frac{1}{m} \sum_{1 \leq i \leq m} \exp(-y^i (\alpha_1 h_1(x^i) + \cdots + \alpha_R h_R(x^i)))
\]

Claim: \( Z_1 \cdots Z_R = \sqrt{1 - 4\gamma_1^2} \cdots \sqrt{1 - 4\gamma_R^2} \) where \( \gamma_t \equiv \frac{1}{2} - \epsilon_t \geq \gamma \)

Reason: \( Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)} = 2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{(1-2\gamma_t)(1+2\gamma_t)} = \sqrt{1 - 4\gamma_t^2} \)

Previous three Claims imply that training error of \( h \) on \( S \) is
\[
\frac{1}{m} \left| \{1 \leq i \leq m \mid h(x^i) \neq y^i \} \right| \leq \left( \sqrt{1 - 4\gamma^2} \right)^R < (e^{-4\gamma^2})^{R/2} \leq \epsilon \quad \text{if } R \geq \frac{1}{2\gamma^2} \ln \frac{1}{\epsilon}
\]
e.g. If \( \epsilon = \frac{1}{m} \), then \( h \) is correct on all of \( S \)
But our goal is to get hypothesis with small (true) error, not training error!
By Theorem in Notes13, suffices to show the following hypothesis class \( \mathcal{H}_R \) has small VC dimension
\[
\mathcal{H}_R = \left\{ \text{sign} \left( \sum_{1 \leq t \leq R} \alpha_t h_t \right) \mid \alpha_t \in \mathbb{R}, h_t \in \mathcal{H} \text{ for } 1 \leq t \leq R \right\}
\]
Here \( \mathcal{H} \) denotes the hypothesis class of weak learner \( A \)
Functions in \( \mathcal{H}_R \) are (\pm 1 version of) centered linear threshold functions of at most \( R \) hypotheses of \( A \)

**Proposition 1.** If \( \text{VCDim}(\mathcal{H}) \leq d \), then \( \text{VCDim}(\mathcal{H}_R) \leq O(Rd \log R) \)

This proposition can be proved by considering growth function (next lecture)