Equivalence between weak and strong learning

- We consider a modified version of AdaBoost
- Instead of weighted sample (D_t, S) we draw an (unweighted) i.i.d. sample S_t from D_t
- We feed S_t to the weak learner, which produces $h_t = WL(S_t)$
- Note: In practical applications, if you have an algorithm

that does support sample weights, you can use the same trick.

- The sizes of the samples $S_1, S_2, \ldots$ will be m_0

- If WL is γ-weak learner

$$err_{D_t}(WL(S_t)) \leq \frac{1}{2} - \gamma$$

- So if $T > \frac{\ln |S|}{2\gamma^2}$ then AdaBoost outputs h such that $\widehat{err_S}(h) = 0$.

- W/o want prove that

- $err_D(h)$ is small without relying on the assumption that $h_1, h_2, \ldots, h_T$ belong to a class with finite VC dimension.

- We use compression schemes.

Definition: A learning algorithm $A: (X \times Y)^* \to Y^X$ is called k-compressing if there exist a mapping $\phi: \bigcup_{t=0}^{k} (X \times Y)^t \to Y^X$ and for any labeled sample $S = ((x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m))$

there exist $r \in \{0, 1, \ldots, k\}$ and indices $i_1, i_2, \ldots i_r \in \{1, 2, \ldots, m\}$ such that

$$A(S) = \phi((x_{i_1}, y_{i_1}), \ldots, (x_{i_r}, y_{i_r})).$$

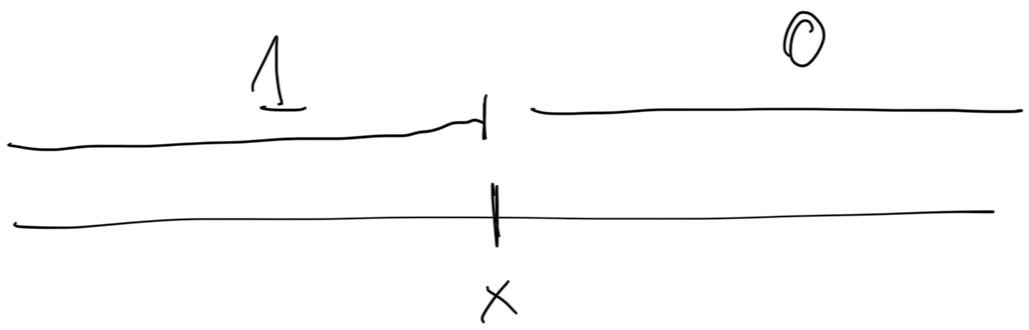
- ϕ is called decompressor for A

Example:

- Let $X = \mathbb{R}$ and $Y = \{0, 1\}$
- $\phi((x, y)) = h_x \in Y^X$

where

$$h_x(x') = \mathbb{1}[x' \leq x]$$



and

$$\phi(\emptyset) = h_0$$

where $h_0(x) = 0$ for all $x \in X$.

- Consider the class of thresholds

$$H = \{ h_x : x \in \mathbb{R} \}$$

- There exists a particular version of ERM that is 1-compressing:

Given $S = (x_1, y_1) \dots (x_m, y_m)$

let $H_S = \{h_0, h_{x_1}, h_{x_2}, \ldots, h_{x_m}\}$

The algorithm outputs

$$\hat{h} = \underset{h \in H_S}{\operatorname{argmin}} \ \widehat{err}_S(h)$$

Theorem:

Let D be arbitrary probability distribution over $X \times Y$.

Let A be a k-compressing algorithm.

Let S be an i.i.d. sample from D of size $m > k$.

Then

$$\Pr\left[\widehat{err}_S(A(S)) = 0 \text{ and } err_D(A(S)) > \varepsilon\right]$$

$$\leq (1-\varepsilon)^{m-k} \sum_{r=0}^{k} m^r$$

Proof:

- Let $I = (i_1, i_2, \ldots, i_r)$ be a tuple of $r \leq k$ indices $i_1, i_2, \ldots, i_r \in \{1, 2, \ldots, m\}$
- Let $S = ((x_1, y_1), \ldots, (x_m, y_m))$
- Let $S_I = ((x_{i_1}, y_{i_1}), \ldots (x_{i_r}, y_{i_r}))$
- Let ϕ be the decompressor

for A.

- Then

$$\Pr\left[\widehat{err}_S(\phi(S_I)) \text{ and } err_D(\phi(S_I)) > \varepsilon\right]$$
$$\leq (1-\varepsilon)^{m=|I|}$$
$$\leq (1-\varepsilon)^{m-k}$$

- Union bound over all choices of I of size $|I| \leq k$

$$\Pr\left[\exists I : \widehat{err}_S(\phi(S_I)) \text{ and } err_D(\phi(S_I)) > \varepsilon\right]$$
$$\leq (1-\varepsilon)^{m-k} \sum_{r=1}^{k} m^r$$

- $A(S) = \phi(S_I)$ for some I

of size $|I| \leq k$.

Corrolary: Let D, A, k, S, m be as before. With probality at least $1-\delta$, $\widehat{err}_S(A(S)) = 0$ then

$$err_D(A(S)) \leq \frac{(k+1)\ln m + \ln(1/\delta)}{m-k}$$

Proof:

$$\sum_{r=0}^{k} m^r \leq m^{(k+1)}$$

$$(1-\varepsilon)^{m-k} \leq e^{-\varepsilon(m-k)}$$

Solve

$$m^{k+1} \cdot e^{-\varepsilon(m-k)} = \delta$$

for ε.

- The sampling version of AdaBoost is a $T(m)$ - compressing algorithm.

  T: number of rounds; m: #samples for weak learners

- Once the samples $S_1, S_2, \ldots, S_T$ are selected $h_1, h_2, \ldots, h_T$ are determined. In turn, these determine $\alpha_1, \alpha_2, \ldots, \alpha_T$ and

$$h = \sum_{t=1}^{T} \alpha_t h_t$$

Claim:

Let $\varepsilon, \delta \in (0,1)$. Let $\gamma \in (0, \frac{1}{2})$.

Let m_0 be a positive integer.

There exist positive integers T, m such that

$$T > \frac{\ln m}{2\gamma^2}$$

$$\frac{(Tm_0+1)\ln m + \ln(T/\delta)}{m - Tm_0} < \varepsilon.$$

Proof:

- For any m, we consider

$$T := T(m) = \left\lceil \frac{\ln m}{\quad} \right\rceil + 1$$

$$\frac{1}{} \quad \frac{1}{(\cdot)} \quad \frac{1}{2g^2} \frac{1}{}$$

- Clearly the first inequality is satisfied.
- The second inequality becomes

$$\frac{\left(T(m)\, m_0 + 1\right) \ln(m) + \ln(2/\delta)}{m - T(m)\, m_0} < \varepsilon$$

- Clearly for large enough m the inequality holds.

Theorem: Let $H \subseteq \{+1, -1\}^X$ be a class of predictors.
Let $g \in (0, \frac{1}{2})$.

H is γ-weakly PAC learnable if and only if H is PAC learnable.

Proof:

- $\Leftarrow$ is trivial
- $\Rightarrow$: Suppose A is a γ-weak learner for H.
- Let m_0 be sample size such such that for any distribution Q and any i.i.d. sample S' from Q labeled by any $h \in H$

$$err_{Q,h}(A(S')) \leq \gamma.$$

with probability at least $1/2$.

- Let T, m be positive integers

such that

$$T > \frac{\ln m}{2\gamma^2}$$

$$\frac{(Tm_0+1)\ln m + \ln(2/\delta)}{m - Tm_0} < \varepsilon$$

(see previous lemma.)

- Let $L = \lceil \log_2(2T/\delta) \rceil$
- We run AdaBoost with subsampling and check

$$err_{D_t}(A(S_t)) \leq \gamma$$

- If $err_{D_t}(A(S_t)) > \gamma$, we resample S_t afresh and repeat.

- We repeat up to L times if necessary.
- If we fail after L attempts, AdaBoost fails (e.g. output all-zero classifier).
- The failure probality is at most

$$T \cdot 2^{-L} \leq T \cdot 2^{-\log_2(2T/\delta)}$$

$$= T \cdot \frac{\delta}{2T}$$

$$= \frac{\delta}{2}$$

- If the algorithm does not fail, it outputs

$$h = \sum_{t=1}^{T} \alpha_t h_t$$

with $\widehat{err}_S(h) = 0$.

- With probality at least $1 - \frac{\delta}{2}$ the generalization error of h satisfies

$$err_D(h) \leq \frac{(Tm_0+1)\ln m + \ln(3/\delta)}{m - Tm_0}$$

$$\leq \varepsilon$$

Compression theorem.

- By union bound, the algorithm fails to output h with $\widehat{err}_D(h) \leq \varepsilon$ with probability at most

$$\frac{\delta}{2} + \frac{\delta}{2} = \delta$$