Machine Learning Theory (CSC 482A/581B)

Problem Set 3 **Due on Wednesday, March 13th, 7pm**

Instructions:

- You must write up your solutions individually.
- You may have high-level discussions with 1 other student registered in the course. If you discuss problems with another student, include at the top of your submission: their name, V#, and the problems discussed.
- Your must type up your solutions and are encouraged to use LaTeX to do this. For any problems where you only have a partial solution, be clear about any parts of your solution for which you have low confidence.
- Please submit your solutions via conneX by the due date of Wednesday, March 13th, 7pm. This is a hard deadline.

Questions:

1. Let $\mathcal{A}$ be a learning algorithm (for a concept class $\mathcal{C}$) satisfying the following property: for any $\varepsilon > 0$, when $\mathcal{A}$ receives as input a training sample of size $n(\varepsilon)$ (distributed according to P and labeled according to some $c \in \mathcal{C}$), it outputs a hypothesis $\hat{f}$ which, with probability at least $\frac{1}{2}$ over the training sample, satisfies the following risk guarantee:

$$\Pr_{X \sim P}\left(\hat{f}(X) \neq c(X)\right) \leq \varepsilon.$$

   Now, let $\delta \in (0, 1/2)$. Devise a learning algorithm that, using a training sample of size $p(n(\varepsilon), \varepsilon, \delta)$ (distributed according to P and labeled according to some $c \in \mathcal{C}$), returns a hypothesis which, with probability at least $1 - \delta$ over the training sample, has risk at most ε.

   Your algorithm may call $\mathcal{A}$ as a sub-procedure. The function $p(n(\varepsilon), \varepsilon, \delta)$ should be linear in $n(\varepsilon)$, polynomial in $\frac{1}{\varepsilon}$, and polynomial in $\log \frac{1}{\delta}$.

   For extra credit: In addition, ensure that p only grows linearly in $\frac{1}{\varepsilon}$.

2. In AdaBoost, distribution D_t is updated to distribution D_{t+1} via

$$D_{t+1}(j) = \frac{D_t(j) e^{-\alpha_t y_j h_t(x_j)}}{Z_t} \qquad \text{for } j \in [n],$$

   with $\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$ and $\varepsilon_t = \Pr_{j \sim D_t}\left(h_t(X_j) \neq Y_j\right)$.

   This update increases the weight of examples on which h_t made a mistake. We thus should expect h_t to perform poorly under distribution D_{t+1}. Show that

$$\Pr_{j \sim D_{t+1}}\left(h_t(X_j) \neq Y_j\right) = \frac{1}{2}.$$

3. Let $\mathcal{H}$ be a set of classifiers, and consider the set of hypotheses $\mathcal{F}$ used by AdaBoost when run for T rounds with a weak learner that outputs hypotheses in $\mathcal{H}$:

$$\mathcal{F} := \left\{ x \mapsto \operatorname{sgn}\left(\sum_{t=1}^{T} \alpha_t h_t(x) \right) : \alpha \in \mathbb{R}^T, h_t \in \mathcal{H}, t \in [T] \right\};$$

technically, AdaBoost only uses nonnegative weights, but we will consider the simpler case above where each $\alpha_t \in \mathbb{R}$.

Recall that the growth function of $\mathcal{F}$ is defined as

$$\Pi_{\mathcal{F}}(n) = \sup_{(x_1,\ldots,x_n)\in\mathcal{X}^n} \left| \{ (f(x_1), \ldots, f(x_n)) : f \in \mathcal{F} \} \right|$$

(a) Suppose that $|\mathcal{H}|$ is finite and $\mathcal{F}$ is defined in terms of this $\mathcal{H}$. Prove that

$$\Pi_{\mathcal{F}}(n) \leq |\mathcal{H}|^T \left(\frac{en}{T} \right)^T.$$

(b) Suppose instead that $\operatorname{VCdim}(\mathcal{H}) = V$ and $\mathcal{F}$ is defined in terms of this $\mathcal{H}$. Prove that

$$\Pi_{\mathcal{F}}(n) \leq \left(\frac{en}{V} \right)^{TV} \left(\frac{en}{T} \right)^T.$$

(c) We now consider an implication of the above result. Using the above and an argument similar to (but much simpler than) the compression bound-based argument from class, it is possible to show the following bound:

> Given as input a training sample of size n, if AdaBoost is run for T rounds with a weak learner that outputs hypotheses in $\mathcal{H}$ with $\operatorname{VCdim}(\mathcal{H}) = V$, then it returns a hypothesis $\hat{f}$ which with probability at least $1 - \delta$ satisfies
>
> $$\Pr_{X\sim P}\left(\hat{f}(X) \neq c(X) \right) \leq \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\left[\hat{f}(X_j) \neq c(X_j) \right] + C\sqrt{\frac{TV \log n + \log \frac{1}{\delta}}{n}},$$
>
> for a universal constant $C > 0$.

Suppose that this bound is also tight, so that the inequality is sometimes also an equality. Explain in words how this bound demonstrates the usual tradeoff between model complexity and estimation error, and comment on whether or not AdaBoost can overfit. Also, comment on whether or not the above bound can be tight if the weak learner always obtains an edge of at least $\gamma = \frac{1}{4}$ (and if not, explain why not).