Instructor: Rocco A. Servedio

Computer Science 4252: Introduction to Computational Learning Theory
Problem Set #4 Fall 2025

Due 11:59pm Tuesday, November 25, 2025

See the course Web page for instructions on how to submit homework.
Important: To make life easier for the TAs, **please start each problem on a new page.**

Remember to strive for both clarity and concision in your solutions;
solutions which are excessively long may be penalized.

Problem 1 In this problem you'll explore how the AdaBoost algorithm (which, as we saw in class, works over a fixed sample of data points) can be used to efficiently PAC learn certain linear threshold functions.

(i) (easy) Suppose that h and f are both functions which take values in $\{-1, 1\}$. Show that for any distribution $\mathcal{D}$, h is a weak hypothesis for f with advantage γ if and only if $\mathbf{E}_{x\sim\mathcal{D}}[h(x)f(x)] \geq 2\gamma$.

(ii) Suppose that $f(x_1, \ldots, x_n) : \{-1, 1\}^n \to \{-1, 1\}$ is a linear threshold function $f(x) = \text{sign}(w \cdot x)$ where

- **[1]** each x_i takes values in $\{-1, 1\}$;
- **[2]** $w = (w_1, \ldots, w_n)$ where each w_i is an integer value and $W = \sum_{i=1}^n |w_i|$;
- **[3]** for all $x \in \{-1, 1\}^n$, we have $w_1x_1 + \cdots + w_nx_n \neq 0$.

Show that for any distribution $\mathcal{D}$ over $\{-1, 1\}^n$, there must be some x_i such that $|\mathbf{E}_{x\sim\mathcal{D}}[f(x) \cdot x_i]| \geq \frac{1}{W}$. (Hint: Use (and justify) the fact that $1 \leq \mathbf{E}_{x\sim\mathcal{D}}[|w \cdot x|]$.)

(iii) Fix a polynomial $p(n)$ and let $\mathcal{C}$ be the concept class of all linear threshold functions $f(x) = \text{sign}(w \cdot x)$ over $\{-1, 1\}^n$ as in (ii) where $\sum_{i=1}^n |w_i| \leq p(n)$. Show how AdaBoost can be used as a PAC learning algorithm for $\mathcal{C}$. Analyze the running time and sample complexity of your algorithm. (Hint: Use AdaBoost as a consistent hypothesis finder.)

Problem 2 Suppose that concept class $\mathcal{C}$ is efficiently learnable in the SQ model by an algorithm that uses only queries with tolerance τ, where $\frac{1}{\tau} \leq r = r(1/\varepsilon, n, \text{size}(c))$. Show that then $\mathcal{C}$ is efficiently learnable in the malicious noise model if the malicious noise rate η is at most $\frac{1}{2r}$.

Problem 3 Our definition of efficient PAC learning in the presence of random classification noise at rate $\eta < 1/2$ requires that the algorithm run in time $\text{poly}(\frac{1}{1-2\eta})$ (ignore all the other parameters for simplicity for this problem). This is intuitively plausible, since (i) if $\eta = 0$ (no noise) the function $\frac{1}{1-2\eta}$ equals 1, and (ii) as η approaches 1/2 (and learning becomes impossible) the function $\frac{1}{1-2\eta}$ approaches infinity. But why is $\frac{1}{1-2\eta}$ the "right" function, as opposed to some other function such as $1 + \log(\frac{1}{1-2\eta})$ or $\exp(\frac{1}{1-2\eta} - 1)$, that satisfies (i) and (ii)?

Argue as clearly and convincingly as you can that any PAC learning algorithm for learning in the presence of random classification noise at rate η must have runtime which grows as $\Omega(\frac{1}{1-2\eta})$. (Hint: One way to show this is to show that the sample complexity must grow as $\Omega(\frac{1}{1-2\eta})$.)

Problem 4 Consider the uniform distribution $\mathcal{U}$ over $[N] = \{1, \ldots, N\}$. A single draw is guaranteed to return an element i that has $\mathcal{U}(i) = 1/N$, which is "typical" for draws from this distribution (since every element has weight $1/N$).

Now consider the distribution $\mathcal{D}_i$ over $[N]$ which puts all of its weight on the point i. A single draw is guaranteed to return the element i that has $\mathcal{D}(i) = 1$. Once again this is "typical" for draws from this distribution, since every draw from $\mathcal{D}_i$ will return an element (the same element) whose weight is 1.

In this problem you'll show that the above examples are special cases of a general phenomenon: for any distribution, a small number of samples will "cover" most of the "typical" probability weights that the distribution assigns to elements.

Let $\mathcal{D}$ be a probability distribution on the set $[N] = \{1, \ldots, N\}$. Given a value $\varepsilon > 0$ and a point $i \in [N]$, we say that a set $R = \{r_1, \ldots, r_t\} \subseteq [N]$ *ε-covers* i if there is some $r_j \in R$ such that

$$\mathcal{D}(i) \in \left[\frac{1}{1+\varepsilon} \cdot \mathcal{D}(r_j), (1+\varepsilon) \cdot \mathcal{D}(r_j)\right].$$

(Here "$\mathcal{D}(i)$" denotes the amount of probability weight that distribution $\mathcal{D}$ puts on point i.)

Let R be a sample of m points drawn from $\mathcal{D}$. Let U (for "uncovered") be the set of all points $i \in [N]$ such that R does *not* ε-cover i. Show that for a suitable choice of $m = \text{poly}(\log N, 1/\varepsilon)$, with probability at least 99/100 it is the case that $\mathcal{D}(U) \leq \varepsilon$.