1. *[30 points]*

For each of the following statements, determine if it holds true for every concept classes H and H'. If it does, prove it, if not, give a counter example.

Q1a 5

(a) *[5 points]* If $|H| \leq |H'|$ then $\text{VCdim}(H) \leq \text{VCdim}(H')$.

Consider $H = \{ h_{a,b,s} : a \leq b, \ s \in \{1,-1\} \}$

$$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a,b] \\ -s & x \notin [a,b] \end{cases}$$

We have shown in previous assignment that $VCdim(H) = 3$.

$H' :=$ be the decision stumps on natural number

We already seen in class that $VCdim(H') = 2$.

Notice $|H| = |H'| = \infty$, but $VCdim(H) > VCdim(H')$

Therefore, the statement is false.

(b) *[5 points]* If $H \subseteq H'$ then $\mathrm{VCdim}(H) \leq \mathrm{VCdim}(H')$.

Q1b 5

let A be any set shattered by H. Since $H \subseteq H'$, A is also shattered by H' and $VCdim(H')$ is not less than $VCdim(H)$.

Therefore $VCdim(H) \leq VCdim(H')$. and statement is true.

Q1c 10

(c) *[10 points]* Let $f : X \to X$. Given a class H, define a class H_f as $\{h_f : h \in H\}$, where for every $h : X \to \{0,1\}$ and $x \in X$, $h_f(x) = h(f(x))$. Then, for every such H and f, $\text{VCdim}(H_f) \le \text{VCdim}(H)$.

Consider some subset $A \subseteq X$ shattered by H_f.

claim: function f is a one-to-one function on subset A.

Assume towards contradiction that f is not one-to-one on subset A. Since subset A is shattered by H_f and $h_f(x) = h(f(x))$, then for any $x_1 \neq x_2$, it has to be $f(x_1) \neq f(x_2)$, because otherwise $h_f(x_1) = h_f(x_2)$ and H_f does not contain function for x_1 and x_2 labeled differently. Notice from previous statement, we conclude that $x_1, x_2 \in A$, $x_1 \neq x_2$ then $f(x_1) \neq f(x_2)$, which contradicts our assumption. Therefore f is one-to-one on subset A.

Then since f is one-to-one on A, we can construct B where each element of $x_B = f(x_A)$ for some $x_A \in A$. $|A| = |B|$.

Since A is shattered by H_f and any $h_f \in H_f$, $h_f(x) = h(f(x))$ then B is shattered by H. Therefore, for any subset A shattered by H_f, we can find B that is shattered by H, and $\text{VCdim}(H) \geq \text{VCdim}(H_f)$.

(d) ***[10 points]*** For every H, f, H_f, as above, $\mathrm{VCdim}(H_f) \geq \mathrm{VCdim}(H)$.

Q1d 10

let H be the class of tree stumps, we know from class $VCdim(H) = 2$.
let $f(x) = 1$ for all $x \in X$.

Since $h_f \in H_f$ will output same label for every $x \in X$. Therefore it cannot shatter any set of size greater than 1.

Therefore $VCdim(H_f) < VCdim(H) = 2$. and the statement is false.

Q2a 10

2. *[30 points]* Let $m_H(\epsilon, \delta)$ denote the agnostic sampe complexity of H. Namely, minimum over all learners A of the training sample size needed to guarantee that for every probability distribution P over $X \times \{0,1\}$, on samples of that size generated i.i.d. by P, A outputs a predictor function $h : X \to \{0,1\}$ such that with probability (over the samples) greater than $(1-\delta)$, $L_P(h) \leq L_P(H)+\epsilon$. (where $L_P(H)$ is the minimum over all $h \in H$ of $L_P(h)$). Also, let $t_H(\epsilon, \delta)$ be the minimum computation time (over all learning algorithms A) to output such a predictor (that can be evaluated on every $x \in X$ within that time bound).

   (a) *[10 points]* Prove that if a class H is agnostic PAC learnable the function $m_H(\epsilon, \delta)$ is polynomial in $1/\epsilon$ and $\log(1/\delta)$.

Since H is agnostic PAC learnable, VCdim (H) is finite and is a constant value for function $m_H(\varepsilon, \delta)$.

By the fundamental theorem of PAC learning.

$$m_H(\varepsilon, \delta) = C \frac{VCdim(H) + \log(\frac{1}{\delta})}{\varepsilon^2} \quad \text{for some constant } c.$$

which is polynomial in $1/\varepsilon$ and $\log(1/\delta)$.

(b) *[**5 points**]* Prove that whenever $H \subseteq H'$, then for all ϵ, δ, $m_H(\epsilon,\delta) \leq m_{H'}(\epsilon,\delta)$.

Q2b 5

Assume that both H and H' be agnostic PAC learnable

Let learner A be the best learner that PAC learns H' with $m_{H'}(\varepsilon, \delta)$ samples. Then by definition, with $m \geq m_{H'}(\varepsilon,\delta)$, A is able to output predictor h' with probability of $(1-\delta)$ that.

$$Pr\left[L_D(h) \leq L_D(h^*) + \varepsilon \right] \geq 1-\delta, \quad h^* \in H' \text{ has minimum } L_D(h), h \in H'$$

$$\Longrightarrow \quad Pr\left[L_D(h) \leq L_D(h') + \varepsilon \right] \geq 1-\delta, \quad h' \in H \text{ has minimum } L_D(h), h \in H$$

The last inequality is from the fact that $H \subseteq H'$. Also learner for H does not needs to output hypothesis from H. Therefore A is also a learner for class H. We have $m_H(\varepsilon,\delta) \leq m_{H'}(\varepsilon,\delta)$.

(c) ***[5 points]*** Prove that whenever $H \subseteq H'$, then for all ϵ, δ, $t_H(\epsilon, \delta) \leq t_{H'}(\epsilon, \delta)$.

Q2c 5

Assume that both H and H' be agnostic PAC learnable, otherwise there's no point in analyzing the computation time.

Let learner A be the best learner that PAC learns H' in $t_{H'}(\varepsilon, \delta)$ time. Then by definition, $\exists\, m_{H'}(\varepsilon, \delta)$ that is able to output predictor h' with probability of $(1-\delta)$ that.

$$Pr\left[L_D(h) \leq L_D(h^*) + \varepsilon \right] \geq 1-\delta, \quad h^* \in H' \text{ has minimum } L_D(h), h \in H'$$

$$\Longrightarrow \quad Pr\left[L_D(h) \leq L_D(h') + \varepsilon \right] \geq 1-\delta, \quad h' \in H \text{ has minimum } L_D(h), h \in H$$

The last inequality is from the fact that $H \subseteq H'$. Also learner for H does not needs to output hypothesis from H. Therefore A is also a learner for class H. We have $t_H(\varepsilon, \delta) \leq t_{H'}(\varepsilon, \delta)$.

(d) *[10 points]* Which of the above two statements remains valid if we require that the learning is *proper* (that is, a learner for any class must output a classifier that belongs to that class)? Explain your answers.

Q2d 10

The statement in (c) is no longer true. As we have seen in class, there is an efficient algorithm for 3-term DNF, but it is not a proper learner since the output is from class of CNF formulas. The problem is NP-hard and does not exist proper learning algorithm.

The statement in (b) is still true. By the quantitative version of the fundamental theorem of statistical learning.

Q3 20

3. *[20 points]* Let H_{Outlier} be the class of all $\{0,1\}$- valued functions over the set of natural numbers that assign the same label to all domain instances except for at most a single x. Describe an agnostic PAC learner for H_{Outlier} that runs in time $t(\epsilon, \delta)$ that is polynomial in $1/\epsilon$ and $\log(1/\delta)$. Prove your claims.

Claim $VC\dim(H_{Outlier}) = 3$.

let $A = \{1, 2, 3\}$.

If all the elements in A have y values of 1 or 0, then $H_{outlier}$ contains such labeling. Since there're only 0 and 1 labelings, there must be at least 2 elements in A have the same labeling. Thus. $H_{outlier}$ also contains labelings where exactly 2 elements have the same labeling.

Therefore $VC\dim(H_{outlier}) \geq 3$.

Consider any $|A| > 3$. We can always split into two parts with different labelings. and each part has at least two elements. Notice now $H_{outlier}$ does not contain such labeling. Therefore $H_{outlier}$ doesnot shatter any set with more than 3 elements and $VC\dim(H_{outlier}) = 3$.

Since $VC\dim(H_{outlier})$ is finite, by the fundamental theorem of statistical learning, any ERM learner is a successful agnostic PAC learner.

ERM Algorithm: Iterates each pair (x, y) in the sample and labels it with y. and label the remaining samples with label other than y. Lastly also calculates the loss when labeling all the samples as 1 and 0. Outputs the predictor with the least sample error.

Correctness: The algorithm iterates through all the possible labelings. For each (x, y) we do not calculates the sample error for labeling the (x, y) pair wrong. That is if $(x, 0)$, then we do not label x as 1. This is because if labeling x as 1 and all the other samples label of 0 yields a better error, then we can produces an even better error by labeling all the samples as 0.
Thus the algorithm tries all the possible labelings and outputs the best one. It is indeed an ERM learner.

Time: Labeling all the samples as 1 and calculates the error, Err, in $\Theta(m)$ time. At each pair (x, y) we can calculate the sample error in constant time by:

If $y=1$, the sample error $= m - \text{Err} - 1$

If $y=0$, the sample error $= \text{Err} - 1$

Also calculates the sample error when labeling all samples as 0 costs $\Theta(m)$. Thus, the total running time is $\Theta(m)$.

By the fundamental theorem of PAC learning.

$$m \leq C \frac{3 + \log(\frac{1}{\delta})}{\varepsilon^2} \quad \text{for some constant } C.$$

Therefore, the ERM Algorithm runs in time polynomial in $\frac{1}{\varepsilon}$ and $\log(\frac{1}{\delta})$.

4. *[20 +10 points]*

Q4a 10

Given a probability distribution, P over X, we say that a probability distribution, $\overline{P}$ *extends* P if, for every $x \in X$, either $\overline{P}(x,1) = P(x)$ and $\overline{P}(x,0) = 0$ or $\overline{P}(x,0) = P(x)$ and $\overline{P}(x,1) = 0$. A class of function, H, is $P-learnable$ if there exist a learning function, A, such that for every $\epsilon, \delta > 0$ there exist $m(\epsilon, \delta)$ such that for every $\overline{P}$ extending P, for every sample size $m > m(\epsilon, \delta)$,

$$Pr_{S\sim\overline{P}^m}\left[L_{\overline{P}}(A(S)) > L_{\overline{P}}(H) + \epsilon\right] < \delta$$

(Where, as before, $L_{\overline{P}}(H) = \inf\{L_{\overline{P}}(h) : h \in H\}$).

(a) ***[10 points]*** Prove that for every class $H \subseteq \{0,1\}^X$, if H is agnostic PAC learnable then, for every probability distribution, P over X, H is P-learnable.

Since by the definition of agnostic PAC learnability, for every P there exists. $m_H(\varepsilon, \delta)$ such that when $m \geq m_H(\varepsilon, \delta)$.

$$Pr_{s\sim p^m}\left[L_p(A(S)) > L_p(H) + \varepsilon\right] < \delta$$

Denote $P_x :=$ all the distribution over the domain X.

$Q :=$ all the labeling distribution for the agnostic learning above.

$P_x : X \longmapsto [0,1]$ $\quad Q : X \times \{0,1\} \longmapsto [0,1]$.

Since for any $p \in P$, distribution $q(x,0) = p(x)$, $q(x,1) = 0$ or $q(x,1) = p(x)$, $q(x,0) = 0$ for $x \in X$ is a valid member of Q. By the fact that H is agnostic PAC learnable, we can conclude that H is P-learnable.

(b) *[10 points]* Show that there exists a class H such that the VC-dimension of H is infinite and a probability distribution P over X such that H is P-learnable.

Q4b 10

H be the class of all functions on the domain X.

let $a \in X$, distribution P is defined as $P(a) = 1$ and 0 otherwise.

Then, the distribution $\bar{P}$ is either $q(a,1) = 1$, $q(a,0) = q(b,0) = q(b,1) = 0$, $b \in X$ and $b \neq a$

or $q(a,0) = 1$, $q(a,0) = q(b,0) = q(b,1) = 0$, $b \in X$ and $b \neq a$.

With this distribution q, by sampling only 1 from $\bar{P}$ we can achieve. distribution error of 0. That is, if sample is (a, y), any $h \in H$ with. $h(a) = y$ will be the best predicator and ERM algorithm on sample will output such predictor with 100 % probability.

therefore, $VCdim(H) = \infty$ and H is P-learnable for P defined above.