Mathematical Foundations of Machine Learning

Spring 2022

Tsinghua University

Lecturer: Yuan Zhou **Homework 1**

Posted: Feb 25, 2022 Due: Mar 14 23:59, 2022

Name: 董浚哲 ID: 2019011985

There are 6 problems in this homework. The total amount of available points is 100. Throughout this exercise, we use log *to denote the natural logarithm (i.e., the* ln *function).*

1. [10 pts.] **VC-dimension of axis aligned rectangles in $\mathbb{R}^d$.** Let $\mathcal{H}^d_{\text{rec}}$ be the class of axis aligned rectangles in $\mathbb{R}^d$. In Lecture 2, we have seen that $\text{VCdim}\left(\mathcal{H}^2_{\text{rec}}\right) = 4$. Prove that for general $d \in \mathbb{N}_+$, we have that $\text{VCdim}\left(\mathcal{H}^d_{\text{rec}}\right) = 2d$.

   Solution.

   On one hand, $\mathcal{H}^d_{rec}$ shatters the set $\{\pm e_1, \pm e_2, , \cdots, \pm e_n\}$, where $e_i = (0, \cdots, 0, 1, 0, \cdots, 0)$ with the only "1" at the i-th place. Denote the rectangle as $[a_1, b_1] \times \cdots \times [a_n, b_n]$, where $a_i = -\frac{1}{2} - h(-e_i), b_i = \frac{1}{2} + h(e_i)$

   On the other hand, consider the set of points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_{2n+1}$. Now we contruct a scheme where at least one point labeled 0 is inevitably covered by the box. $\forall i \in [1, n]$, consider $\boldsymbol{x}_{i_{floor}}, \boldsymbol{x}_{i_{ceil}}$ s.t. $x^i_{i_{floor}} = \min\{x^i_k : 1 \le k \le 2n+1\}, x^i_{i_{ceil}} = \max\{x^i_k : 1 \le k \le 2n+1\}$. Label both points with "1". If a point is not labled after this process, lable it with "0". In this way, at most $2n$ points are labled "1" while at least 1 point is labled with "0". Meanwhile, to predict the selected $2n$ points right, all points are covered by the rectangle since $a_i \le x_{i_{floor}}, b_i \ge x_{i_{ceil}}$, which results in predicting the point labeled "0" wrong. So $\mathcal{H}^d_{rec}$ cannot shatter any set of $2n+1$ points.

   □

2. [10 pts.] **VC-dimension of a vector space of real functions.** Let F be a finite-dimensional vector space of real functions on $\mathbb{R}^n, \dim(F) = r < \infty$. Let $\mathcal{H}$ be the set of hypotheses:

$$\mathcal{H} = \{\{x : f(x) \ge 0\} : f \in F\}.$$

Show that d, the VC-dimension of $\mathcal{H}$, is finite and that $d \leq r$. (**Hint:** select an arbitrary set of $m = r + 1$ points and consider $\{(f(x_1), \ldots, f(x_m)) : f \in F\}$)

Solution. Suppose the contrary is true s.t. for $m = r + 1$, $\forall C \subset \mathbb{R}^n, |C| = m$, $\mathcal{H}$ shatters $C = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_m\}$. Select a basis of F: $\mathbb{B} = \{f_1, f_2, \cdots, f_r\}$. Give points in C an arbitrary lable. Since $\mathcal{H}$ shatters C, $\exists f(x) = \sum_{i=1}^{r} \alpha_i f_i(x)$ and a partition $I \cup J = \{1, 2, \cdots, n\}, f(x_i) \geq 0 \quad \forall i \in I, f(x_j) < 0 \quad \forall j \in J$. This results in the equation:

$$\begin{bmatrix} f_1(x_1) & \cdots & f_r(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_m) & \cdots & f_r(x_m) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_r \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix}$$

where $\forall i \in I, \beta_i \geq 0; \forall j \in J, \beta_j < 0$.

This is an overdetermined equation, so unless rows of the matrix are linear related there's no solution (WLOG assume the last row can be linearly represented by others), but in this case β_m is no longer free to be positive or negative, which is a contradiction.

□

3. [20 pts.] **The weighted majority learner.** Recall the "The Halving Learner" in Lecture 1. We reached the conclusion that the halving learner makes at most $\log_2 |\mathcal{H}|$ mistakes under the assumption that there exists one hypothesis $h \in \mathcal{H}$ which consistently provides the correct label. In this exercise, we will explore the case when such an assumption does not hold, and we will design a learner trying to match up with the best hypothesis in $\mathcal{H}$ (i.e., the one that proves the least number of wrong labels).

   Now, let us assume that there exists one hypothesis $h \in \mathcal{H}$ that makes at most M mistakes. The goal of our learner is to upper bound the number of mistakes by a function of M.

   (a) **The simple method.** We run the halving learner until every hypothesis in $\mathcal{H}$ makes at least one mistake (so that all hypotheses are eliminated). We call such a process an *epoch*. In our simple method, we repeatedly run the epochs – whenever an epoch ends, we start a new one with all the (eliminated) hypotheses "alive". Please upper bound the number of epochs and the number of mistakes the simple method would make.

(b) **The weighted majority approach.** Let us assign non-negative weights w_h to each hypothesis $h \in \mathcal{H}$. Instead of elimination (which can be viewed as a hard penalty), we impose a soft penalty to a hypothesis function h whenever it makes a mistake – we reduce the weight w_h by a pre-determined factor $C \in [0, 1)$. Note that the elimination (hard penalty) is equivalent to set $C = 0$ so that the weight of the hypothesis function h is set to 0.

In Algorithm 1 we describe the details of the weighted majority approach. The name "weighted majority" comes from the prediction rule that is based on the weighted majority of the hypotheses, where in contrast, the halving learner admits either 0 or 1 weight and predicts based on the majority of the weight-1 hypotheses.

To analyze the weighted majority learner, let us consider the potential function (at time t) that is defined as $w^{(t)} = \sum_{h \in \mathcal{H}} w_h^{(t)}$. Please determine the constant C in Algorithm 1 and prove that $w^{(t+1)} \leq \frac{3}{4} w^{(t)}$ if the learner makes a mistake at time t.

Algorithm 1 The Weighted Majority Learner

Initialization: Set the initial weight for each $h \in \mathcal{H}$ as $w_h^{(1)} = 1$
for $t = 1, 2, \ldots$ **do**
    Get x_t
    Predict $\hat{y}_t = \text{WeightedMajority}\{h(x_t) : h \in \mathcal{H}\}$ i.e. $\hat{y}_t = \text{sign}(\sum_{h \in \mathcal{H}} w_h^{(t)} h(x_t))$
    Get y_t and update weights as follows –
    if $y_t = \hat{y}_t$ **then**
        Let $w_h^{(t+1)} = w_h^{(t)}$ for all $h \in \mathcal{H}$
    else
        For all $h \in \mathcal{H}$ such that $h(x_t) = \hat{y}_t$, let $w_h^{t+1} = C w_h^t$
        For all $h \in \mathcal{H}$ such that $h(x_t) \neq \hat{y}_t$, let $w_h^{(t+1)} = w_h^{(t)}$
    end if
end for

(c) Under the assumption that there exists one hypothesis in $\mathcal{H}$ that makes at most M mistakes, prove a lower bound of the potential function at any time t.

(d) Derive the mistake bound for Algorithm 1 by combining the above conclusions.

Solution.

(a) Denote the wise hypothesis who makes at most M mistakes as $\hat{h}$. In the worst case, exactly one (different) mistake is made in the first M epoch, and in the $M+1$ th epoch, $\hat{h}$ never makes an mistake. In each epoch, at most $\log_2 |H|$ mistakes are made so in total no more than $(M+1)\log_2 |\mathcal{H}| - 1$ mistakes are made.

(b) In the worst case, every prediction reaches a tie: $\hat{y}_t = 0$ and is regarded as a mistake. Here we propose $C = \frac{1}{2}$. Whenever a tie is made, hypothesises that amounts to exactly half the total weight (e.g. $\sum_{h \text{ wrong}} w_h^{(t)} = \frac{1}{2} w^{(t)}$), and their weight are reduced to half of it, so

$$w^{(t+1)} \leq \frac{1}{2} \sum_{h \text{ wrong}} w_h^{(t)} + \sum_{h \text{ correct}} w_h^{(t)} = \frac{1}{2} \cdot \frac{1}{2} w^{(t)} + \frac{1}{2} w^{(t)} = \frac{3}{4} w^{(t)}$$

(c) In the worst case, every hypothesis apart from $\hat{h}$ makes an mistake in every test, and $\hat{h}$ makes an mistake in the first M tests. So (observe that no more mistakes can be made after $2^{-M} \geq (\frac{1}{2})^{-t}(|\mathcal{H}| - 1)$ in the worst case, which means $t \leq M + \log(|\mathcal{H} - 1|)$)

$$w^{(t)} \geq \begin{cases} (\frac{1}{2})^{-t}|\mathcal{H}| & t \leq M \\ 2^{-M} + (\frac{1}{2})^{-t}(|\mathcal{H}| - 1) & M < t \leq M + \log_2(|\mathcal{H}| - 1) \end{cases}$$

(d) Above we've proved that $t \leq M + \log_2(|\mathcal{H}| - 1)$, which is an lower bound. Combine (b) and we see that the upper bound is given by $2^{-M} \geq \frac{1}{2}(\frac{3}{4})^t |H|$, which results in $t \leq (2 - \log_2 3)^{-1}(M - 1 + \log_2(|\mathcal{H}|))$.

□

4. [10 pts.] **Learning concentric circles.** Let $\mathcal{X} = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some real number r. Show that this class can be (ϵ, δ)-PAC-learned from training data of size $m \geq (1/\epsilon)\log(1/\delta)$.

(**Hint:** The proof of Lemma 6.1 in the textbook *Understanding Machine Learning* might serve as a reference.)

Solution.

Denote the (concentric) circle with radius r to be $C_r = \{(x, y) : x^2 + y^2 \leq r^2\}$. Suppose r^* is the radius s.t. $L_D(C_{r^*}) = 0$. Let D_x be the marginal dsitribution over the domain $\mathbb{X}$ and denote $r_1 < r^* < r_2$ the radii s.t.

$$P_{\boldsymbol{x}\sim D_x}(\boldsymbol{x} \in C_{r^*} \backslash C_{r_1}) = P_{\boldsymbol{x}\sim D_x}(\boldsymbol{x} \in C_{r_2} \backslash C_{r^*}) = \varepsilon$$

Given a training data set, denote $b_1 = \max\{|\boldsymbol{x}| : \boldsymbol{x} \in C_{r^*}\}, b_2 = \min\{|\boldsymbol{x}| : \boldsymbol{x} \notin C_{r^*}\}$, and denote C_{r_S} the corresponding ERM hypothesis. By definition $b_1 < r_S < b_2$, so

$$\begin{aligned}
&P_{\boldsymbol{x}\sim D^m}(L_D(C_{r_S}) > \varepsilon) \\
\leq &P_{\boldsymbol{x}\sim D^m}(r_2 < b_2 \vee r_1 > b_1) \\
\leq &P_{\boldsymbol{x}\sim D^m}(r_2 < b_2) + P_{\boldsymbol{x}\sim D^m}(r_1 > b_1) \\
= &(1-\varepsilon)^m + (1-\varepsilon)^m \leq 2e^{-\varepsilon m}
\end{aligned}$$

Take the estimate of m into the inequality above and we get the desired statement. □

5. [20 pts.] **The robust ellipsoid learner in adversarial setting.** In this problem we consider a special adversarial setting of the online learning game. Specifically, when we wish to predict the label with the parameter $\boldsymbol{w}_t$ given the instance $\boldsymbol{x}_t$, the adversary will perturb the model parameter to $\tilde{\boldsymbol{w}}_t$ (see Line 7 in Algorithm 2). Assuming $\|\tilde{\boldsymbol{w}}_t - \boldsymbol{w}_t\| \leq \epsilon_t \leq \sqrt{\lambda_{\min}(A_t)}/(10d^3)$ where $\lambda_{\min}(A_t)$ is the smallest eigenvalue of A_t, we aim to prove that, if we slightly adjust the update process of A_t (see the red constant in Line 10 in Algorithm 2), our ellipsoid learner will still work with a similar mistake bound.

   Fix any time t, let us write A_t in the form of $A_t = UD^2U^\top$ where U is orthonormal and D is diagonal with non-negative entries.

   (a) Show that if $y_t \langle UD\boldsymbol{z} + \boldsymbol{w}_t, \boldsymbol{x}_t\rangle > 0$ and $y_t \langle \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t\rangle < 0$, we have $y_t \langle UD\boldsymbol{z}, \boldsymbol{x}_t\rangle > -\epsilon\|\boldsymbol{x}_t\|$.

   (b) Show that if $y_t \langle UD\boldsymbol{z}, \boldsymbol{x}_t\rangle > -\epsilon\|\boldsymbol{x}_t\|$, we have $\left\langle \boldsymbol{z}, y_t DU^\top \boldsymbol{x}_t \right\rangle / \|DU^\top \boldsymbol{x}_t\| > -1/(10d^3)$.

   (c) Prove the Enclosing Ellipsoid Lemma: when $\hat{y}_t \neq y_t$, we have that $\mathcal{E}_t \cap \{\boldsymbol{w} : y_t \langle \boldsymbol{w}, \boldsymbol{x}_t\rangle > 0\} \subseteq \mathcal{E}_{t+1}$.

Algorithm 2 The Robust Ellipsoid Learner.

Initialization: $\boldsymbol{w}_1 = \mathbf{0}, \boldsymbol{A}_1 = I$
for $t = 1, 2, \ldots$ **do**
  Get $\boldsymbol{x}_t$
  if $\left|\mathcal{E}_t \cap G^d\right| = 1$ **then**
    Choose the only element in $\mathcal{E}_t \cap G^d$ and predict $\hat{y}_t = \text{sign}\left(\boldsymbol{w}_t^\top \boldsymbol{x}_t\right)$
  else
    Predict $\hat{y}_t = \text{sign}\left(\tilde{\boldsymbol{w}}_t^\top \boldsymbol{x}_t\right)$, where $\tilde{\boldsymbol{w}}_t$ is the perturbed parameter selected by the adversary such that $\|\tilde{\boldsymbol{w}}_t - \boldsymbol{w}_t\| \leq \epsilon_t$
  end if
  if $y_t = \hat{y}_t$ **then**
    Update:

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t + \frac{y_t}{d+1}\frac{A_t \boldsymbol{x}_t}{\sqrt{\boldsymbol{x}_t^\top A_t x_t}} = \boldsymbol{w}_t + \boldsymbol{u}_t$$

$$A_{t+1} = \frac{d^2}{d^2-1}\left(A_t - \frac{2}{d+1}\frac{A_t \boldsymbol{x}_t \boldsymbol{x}_t^\top A_t}{\boldsymbol{x}_t^\top A_t \boldsymbol{x}_t}\right) = \left(1 + \frac{1}{2d^2}\right)\frac{d^2}{d^2-1}\left(A_t - 2(d+1)\boldsymbol{u}_t \boldsymbol{u}_t^\top\right)$$

  else
    $\boldsymbol{w}_{t+1} = \boldsymbol{w}_t$ and $A_{t+1} = A_t$
  end if
end for

(d) Prove the Volume Reduction Lemma: $\text{Vol}(\mathcal{E}_{t+1}) \leq \text{Vol}(\mathcal{E}_t) e^{-1/(4d+4)}$ and also prove the mistake bound for the robust learner under the presence of the adversary.

Solution.

(a) $y_t \langle UD\boldsymbol{z} + \boldsymbol{w}_t, x_t\rangle = y_t \langle UDz, \boldsymbol{x}_t\rangle + y_t \langle \boldsymbol{w}_t - \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t\rangle + y_t \langle \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t\rangle$, so

$$y_t \langle UD\boldsymbol{z}, \boldsymbol{x}_t\rangle = y_t \langle UD\boldsymbol{z} + \boldsymbol{w}_t\rangle - y_t \langle \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t\rangle - y_t \langle \boldsymbol{w}_t - \tilde{\boldsymbol{w}}_t\rangle > 0 + 0 - \|\tilde{\boldsymbol{w}}_t - \boldsymbol{w}_t\| \|\boldsymbol{x}_t\| \geq -\epsilon \|\boldsymbol{x}_t\|$$

(b) Observe that $y_t \langle UD\boldsymbol{z}, \boldsymbol{x}_t\rangle = \left\langle \boldsymbol{z}, y_t DU^T \boldsymbol{x}_t\right\rangle$, so $LHS \geq -\frac{\sqrt{\lambda_{\min}(A_t)}}{(10d^3)}\frac{\|\boldsymbol{x}_t\|}{\|DU^T \boldsymbol{x}_t\|}$. Take it into the inequality and simplify, and we see that it's equivalent to prove:

$$\sqrt{\lambda_{\min}(A_t)}\,\|\boldsymbol{x}_t\| = \sqrt{\lambda_{\min}(A_t)}\,\left\|U^T \boldsymbol{x}_t\right\| \leq \left\|DU^T \boldsymbol{x}_t\right\|$$

which is trivialty by the definition of SVD. The equality holds when $U^T \boldsymbol{x}_t$ is a right singular vector of the smallest singular value $\sigma_n = \sqrt{\lambda_{\min}(A_t)}$

(c) When $\hat{y}_t \neq y_t$,then $y_t = -\text{sign}(\tilde{\boldsymbol{w}}_t^T \boldsymbol{x}_t) \Rightarrow y_t \langle \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t \rangle = -| \langle \tilde{\boldsymbol{w}}_t, \boldsymbol{x}_t \rangle | < 0$
The goal is:

$$\forall \boldsymbol{w} \text{ s.t. } y_t \langle \boldsymbol{w}, \boldsymbol{x}_t \rangle > 0 \wedge (\boldsymbol{w} - \boldsymbol{w}_t)^T U D^{-2} U^T (\boldsymbol{w} - \boldsymbol{w}_t) \leq 1 \Rightarrow \boldsymbol{w} \in \mathcal{E}_{t+1}$$

Denote $\boldsymbol{z} = D^{-1} U^T (\boldsymbol{w} - \boldsymbol{w}_t)$, then the goal is equivalent to:

$$\forall z \text{ s.t. } y_t \langle UD\boldsymbol{z} + \boldsymbol{w}_t, x_t \rangle > 0 \wedge \boldsymbol{z}^T \boldsymbol{z} \leq 1 \Rightarrow UD\boldsymbol{z} \in \mathcal{E}(A_{t+1}^{\frac{1}{2}}, t)$$

In this case, by (a) we have $y_t \langle UD\boldsymbol{z}, \boldsymbol{x}_t \rangle \geq -\epsilon \|\boldsymbol{x}_t\|$. So it suffices to prove that:

$$\forall \boldsymbol{z} \text{ s.t. } \left\langle \boldsymbol{z}, y_t D U^T \boldsymbol{x}_t \right\rangle \geq -\epsilon \|\boldsymbol{x}_t\| \wedge \boldsymbol{z}^T \boldsymbol{z} \leq 1 \Rightarrow UD\boldsymbol{z} \in \mathcal{E}(A_t^{\frac{1}{2}}, u_t)$$

Denote V as the Householder matrix (which is symmetric and orthogonal, so we do not bother distinguishing V and V^T) s.t. $V y_t D U^T \boldsymbol{x}_t = \left\| D U^T \boldsymbol{x}_t \right\| e_1$, and denote $\tilde{\boldsymbol{z}} = Vz$, then the new goal is equivalent to (denote the first dimension of $\tilde{\boldsymbol{z}}$ as $\tilde{\boldsymbol{z}}_1$):

$$\forall \tilde{\boldsymbol{z}} \text{ s.t. } \left\| D U^T \boldsymbol{x}_t \right\| \tilde{\boldsymbol{z}}_1 \geq -\epsilon \|\boldsymbol{x}_t\| \wedge \tilde{\boldsymbol{z}}^T \tilde{\boldsymbol{z}} \leq 1 \Rightarrow UD^T V \tilde{\boldsymbol{z}} \in \mathcal{E}(A_{t+1}^{\frac{1}{2}}, u_t)$$

by (b), it suffices to prove that:

$$\forall \tilde{\boldsymbol{z}} \text{ s.t. } \tilde{\boldsymbol{z}}_1 \geq -\frac{1}{10d^3} \wedge \tilde{\boldsymbol{z}}^T \tilde{\boldsymbol{z}} \leq 1 \Rightarrow UD^T V \tilde{\boldsymbol{z}} \in \mathcal{E}(A_{t+1}^{\frac{1}{2}}, u_t)$$

Claim: $\forall \tilde{\boldsymbol{z}}$ s.t. $\tilde{\boldsymbol{z}}_1 > -\frac{1}{10d^3}, \tilde{\boldsymbol{z}}^T \tilde{\boldsymbol{z}} \leq 1 \Rightarrow \tilde{\boldsymbol{z}} \in \mathcal{E}(A^{\frac{1}{2}}, \tilde{u})$, where

$$\tilde{u} = (\frac{1}{d+1}, 0, \cdots, 0)^T \quad A = (1 + \frac{1}{2d^2}) \frac{d^2}{d^2 - 1} (\text{diag}(\frac{d-1}{d+1}, 1, \cdots, 1))$$

proof of the claim. Denote $a = \tilde{\boldsymbol{z}}_1 \in [-\frac{1}{10d^3}], b = \sqrt{\sum_{i=2}^n \tilde{\boldsymbol{z}}_i^2} \leq \sqrt{1 - a^2}$.
It suffices to prove that $(\tilde{\boldsymbol{z}} - \tilde{u})^T A^{-1} (\tilde{\boldsymbol{z}} - \tilde{u}) \leq 1$:

$$\begin{aligned}
& (\tilde{\boldsymbol{z}} - \tilde{u})^T A^{-1} (\tilde{\boldsymbol{z}} - \tilde{u}) \\
=& (1 + \frac{1}{2d^2})^{-1} \frac{d^2 - 1}{d^2} [\frac{d+1}{d-1} (a - \frac{1}{d+1})^2 + b^2] \\
\leq& (1 + \frac{1}{2d^2})^{-1} \frac{d^2 - 1}{d^2} [\frac{d+1}{d-1} (a - \frac{1}{d+1})^2 + (1 - a^2)] \\
=& (1 + a(a-1) \frac{2(d+1)}{d^2}) (1 + \frac{1}{2d^2})^{-1}
\end{aligned}$$

So (after simplification) it suffices to prove that

$$a(a-1) \leq \frac{1}{4d^2(d+1)}$$

LHS takes its maximum at $a = -\frac{1}{10d^3}$, so it suffices to prove that

$$\frac{1}{10d^3}(1+\frac{1}{10d^3}) \leq \frac{1}{4d^2(d+1)}$$

which is true for (calculated by Wolfram Alpha) $d > 0.92809$. Thus the claim is proved. □

$\forall \widetilde{z} : \widetilde{z}^\top \widetilde{z} \leq 1$ and $\tilde{z}_1 > 0$, we have $UDV^\top \widetilde{z} \in \mathcal{E}\left(UDV^\top A^{1/2}, UDV^\top \widetilde{\boldsymbol{u}}\right)$ By defn.: $V^\top = \begin{bmatrix} \frac{y_t DU^\top \boldsymbol{x}_t}{\left\|DU^\top \boldsymbol{x}_t\right\|} & \cdots & \cdots \end{bmatrix}$. Therefore, $UDV^\top \widetilde{\boldsymbol{u}} = \frac{y_t UD^2 U^\top \boldsymbol{x}_t}{(d+1)\left\|DU^\top \boldsymbol{x}_t\right\|} = \frac{y_t}{d+1} \frac{A_t \boldsymbol{x}_t}{\sqrt{\boldsymbol{x}_t^\top A_t \boldsymbol{x}_t}} = \boldsymbol{u}_t$ $\mathcal{E}\left(UDV^\top A^{1/2}, UDV^\top \widetilde{\boldsymbol{u}}\right) = \mathcal{E}\left(\left(UDV^\top AVDU^\top\right)^{1/2}, \boldsymbol{u}_t\right) = \mathcal{E}\left(A_{t+1}^{1/2}, \boldsymbol{u}_t\right)$

$$\begin{aligned}
V^\top AV &= (1+\frac{1}{2d^2})\frac{d^2}{d^2-1} V^\top \left(I - \operatorname{diag}\left(\frac{2}{d+1}, 0, 0, \ldots, 0\right)\right) V \\
&= (1+\frac{1}{2d^2})\frac{d^2}{d^2-1}\left(I - \frac{2}{d+1}\frac{DU^\top x_t x_t^\top UD}{\left\|DU^\top x_t\right\|^2}\right) \\
UDV^\top AVDU^\top &= (1+\frac{1}{2d^2})\frac{d^2}{d^2-1}\left(UD^2U^\top - \frac{2}{d+1}\frac{UD^2U^\top x_t x_t^\top UD^2U^\top}{\left\|DU^\top x_t\right\|^2}\right) \\
&= (1+\frac{1}{2d^2})\frac{d^2}{d^2-1}\left(A_t - \frac{2}{d+1}\frac{A_t x_t x_t^\top A_t}{\left\|DU^\top x_t\right\|^2}\right) = A_{t+1}
\end{aligned}$$

(d) Repeat what's written on lecture notes word by word, and we see that

$$\begin{aligned}
&(\frac{Vol(\mathcal{E}_{t+1})}{Vol(\mathcal{E}_t)})^2 \\
&= \frac{\det(A_{t+1})}{\det(A_t)} = \left(1+\frac{1}{d^2}\right)^d \left(\frac{d^2}{d^2-1}\right)^d \left(1-\frac{2}{d+1}\right) \\
&= (1+\frac{1}{2d^2})^d \left(1+\frac{1}{d^2-1}\right)^{d-1} \left(\frac{d-1}{d+1}\cdot\frac{d^2}{d^2-1}\right) = (1+\frac{1}{2d^2})^d \left(1+\frac{1}{d^2-1}\right)^{d-1} \left(1-\frac{1}{d+1}\right)^2 \\
&\leq \exp(\frac{d}{2d^2}) \exp\left(\frac{d-1}{d^2-1}\right) \cdot \exp\left(-\frac{2}{d+1}\right) \leq \exp\left(-\frac{1}{2(d+1)}\right)
\end{aligned}$$

The lower bound of the volume remains unchanged, so the mistake bound is doubled: $d(4d+4)\log(2n)$ mistakes.

□

6. [30 pts.] **Learning in the presence of noises.** We consider a finite hypothesis set $\mathcal{H}$, assume that the target concept is in $\mathcal{H}$, and adopt the following noise model: the label of a training point received by the learner is randomly changed with probability $\eta \in \left(0, \frac{1}{2}\right)$. The exact value of the noise rate η is not known to the learner but an upper bound η' is revealed to him with $\eta \leq \eta' < 1/2$.

   (a) For any $h \in \mathcal{H}$, let $d(h)$ denote the probability that the label of a training point received by the learner disagrees with the one given by h. Let h^* be the target hypothesis, show that $d\left(h^*\right) = \eta$.

   (b) More generally, show that for any $h \in \mathcal{H}, d(h) = \eta + (1 - 2\eta) L_{\mathcal{D}}(h)$, where $L_{\mathcal{D}}(h)$ denotes the generalization error of h.

   (c) Fix $\epsilon > 0$ for this and all the following questions. Use the previous questions to show that if $L_{\mathcal{D}}(h) > \epsilon$, then $d(h) - d\left(h^*\right) \geq \epsilon'$, where $\epsilon' = \epsilon\left(1 - 2\eta'\right)$.

   (d) For any hypothesis $h \in \mathcal{H}$ and sample S of size m, let $\widehat{d}(h)$ denote the fraction of the points in S whose labels disagree with h. We will consider the algorithm L that, after receiving S, returns the hypothesis h_S with the smallest number of disagreements (thus $\widehat{d}\left(h_S\right)$ is minimized. To show the PAC-learning guarantee for L, we will show that for any h, if $L_{\mathcal{D}}(h) > \epsilon$, then with high probability $\widehat{d}(h) \geq \widehat{d}\left(h^*\right)$. First, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon'^2} \log \frac{2}{\delta}$, the following holds:

$$\widehat{d}\left(h^*\right) - d\left(h^*\right) \leq \epsilon'/2.$$

   (e) Second, show that for any $\delta > 0$, with probability at least $1 - \delta/2$, for $m \geq \frac{2}{\epsilon'^2}\left(\log |\mathcal{H}| + \log \frac{2}{\delta}\right)$, the following holds for all $h \in \mathcal{H}$:

$$d(h) - \widehat{d}(h) \leq \epsilon'/2.$$

   (f) Finally, show that for any $\delta > 0$, with probability at least $1 - \delta$, for $m \geq \frac{2}{\epsilon^2(1-2\eta')^2}\left(\log |\mathcal{H}| + \log \frac{2}{\delta}\right)$, the following holds for all $h \in \mathcal{H}$ with $L_{\mathcal{D}}(h) > \epsilon$:

$$\widehat{d}(h) - \widehat{d}\left(h^*\right) \geq 0.$$

(**Hint:** use $\widehat{d}(h) - \widehat{d}\,(h^*) = [\widehat{d}(h) - d(h)] + [d(h) - d\,(h^*)] + \left[d\,(h^*) - \widehat{d}\,(h^*)\right]$ and use previous questions to lower bound each of these three terms.)

Solution.

(a) By definition, $d(h^*)$ is the probability that the label received by the learner is different from that predicted by h^*. Since h^* always makes correct predictions, it $d(h^*)$ is the probability that the received label is different from the original one. So by the definition of η, $d(h^*) = \eta$.

(b) By definition, $L_D(h) = P_{x\sim D}\{h(x) \neq h^*(x)\}$. Denote the lable given to the learner as $\tilde{y}$. Then

- With probability $1 - \eta, h^*(x) = \tilde{y}$. In this case, $h(x)$ has to be different from $h^*(x)$ to be wrong.
- With probability $\eta, h^*(x) \neq \tilde{y}$. In this case, $h(x)$ has to be the same with $h^*(x)$ to be wrong.

So $d(h) = (1-\eta)L_D(h) + \eta(1 - L_d(h)) = \eta + (1-2\eta)L_D(h)$.

(c) Since $d(h^*) \equiv \eta$, so $d(h) - d(h^*) = (1-2\eta)L_D > (1-2\eta')L_D > (1-2\eta')\varepsilon = \varepsilon'$

(d) Now that $d(h^*) = \eta$, it's equivalent to show that $\hat{d}(h^*)$ holds for the given conditions. Since h^* always gives the correct prediction, $\hat{d}(h^*)$ is the fraction of points whose lables have been changed. Given m, the number of the changed points $X \sim B(m, \eta)$. So

$$P(\hat{d}(h^*) < \eta + \frac{\varepsilon'}{2}) = P(X < m \cdot (\eta + \frac{\varepsilon'}{2}))$$

By central limit theorem,

$$P(\hat{d}(h^*) < \frac{\varepsilon'}{2} + \eta) \geq \Phi(\frac{\varepsilon'}{2}\sqrt{\frac{m}{\eta(1-\eta)}}) > \Phi(\varepsilon'\sqrt{m}) \geq \Phi(\sqrt{2\log(\frac{2}{\delta})})$$

Meanwhile, as long as $\delta < 2e^{-\frac{1}{2}} \approx 1.21 > 1$ (which is always satisfied since $\delta < 1$), we have

$$1 - \Phi(\sqrt{2\log(\frac{2}{\delta})}) < \int_{\sqrt{2\log(\frac{\delta}{2})}}^{\infty} te^{-t^2}\,\mathrm{d}t = \frac{\delta}{2}$$

so $P(\hat{d}(h^*) - d(h^*) < \frac{\varepsilon'}{2}) > 1 - \frac{\delta}{2}$

(e) With similar approach as above, we see that for a fixed $h \in \mathcal{H}$, $P((d(h) - \hat{d}(h)) > \frac{\varepsilon'}{2}) \leq \Phi(-\varepsilon'\sqrt{m}) > \frac{1}{2}e^{-\varepsilon'^2 m} \geq \frac{\delta}{H}$. Now that there're $|H|$ hypothesis, we add these errors and get: $\forall h \in \mathcal{H}$

$$P(d(h) - \hat{d}(h) \leq \frac{\varepsilon'}{2}) \geq 1 - \frac{\delta}{2}$$

(f) Observe that

$$\hat{d}(h) - \hat{d}(h*) = [\hat{d}(h) - d(h)] + [d(h) - d(h^*)] + [d(h^*) - \hat{d}(h^*)]$$

By previous arguments:

- $\hat{d}(h) - d(h) \geq -\varepsilon'/2$
- $d(h) - d(h^*) \geq \varepsilon'$
- $d(h^*) - \hat{d}(h^*) \geq -\varepsilon'/2$

Sum these terms and we reach the desired argument.

□