Machine Learning HW2

B0490169 電機四 林志皓

Problem 1

①

$$F(A,B) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1+\exp\left(-y_n(Az_n+B)\right)\right)$$

$$\frac{\partial F}{\partial A} = \frac{1}{N}\sum_{n=1}^{N} \frac{\exp\left(-y_n(Az_n+B)\right)\cdot(-y_n z_n)}{1+\exp\left(-y_n(Az_n+B)\right)}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \theta\left(-y_n(Az_n+B)\right)\cdot(-y_n z_n) = \frac{-1}{N}\sum_{n=1}^{N} y_n z_n p_n$$

$$\frac{\partial F}{\partial B} = \frac{1}{N}\sum_{n=1}^{N} \frac{\exp\left(-y_n(Az_n+B)\right)\cdot(-y_n)}{1+\exp\left(-y_n(Az_n+B)\right)}$$

$$= \frac{1}{N}\sum_{n=1}^{N} \theta\left(-y_n(Az_n+B)\right)\cdot(-y_n) = \frac{-1}{N}\sum_{n=1}^{N} y_n p_n$$

$$\therefore \nabla F(A,B) = \left(\frac{\partial F}{\partial A}, \frac{\partial F}{\partial B}\right) = \left(\frac{-1}{N}\sum_{n=1}^{N} y_n z_n p_n,\ \frac{-1}{N}\sum_{n=1}^{N} y_n p_n\right)$$

Problem 2

②

$$\theta(x) = \frac{e^x}{1+e^x}$$

$$\frac{\partial \theta}{\partial x} = \frac{(1+e^x)e^x - e^x \cdot e^x}{(1+e^x)^2} = \frac{e^x}{(1+e^x)^2} = \frac{1}{1+e^x} \cdot \frac{e^x}{1+e^x} = (1-\theta(x)) \cdot \theta(x)$$

$$p_n = \theta(-y_n(Az_n + B))$$

$$\frac{\partial p_n}{\partial A} = (1-p_n)p_n \cdot (-y_n z_n) = -(1-p_n)p_n y_n z_n$$

$$\frac{\partial p_n}{\partial B} = (1-p_n)p_n \cdot (-y_n) = -(1-p_n)p_n y_n$$

$$H(F) = \begin{bmatrix} \frac{\partial^2 F}{\partial A^2}, & \frac{\partial^2 F}{\partial A \partial B} \\ \frac{\partial^2 F}{\partial B \partial A}, & \frac{\partial^2 F}{\partial B^2} \end{bmatrix};$$

by problem ①. $\frac{\partial F}{\partial A} = \frac{-1}{N}\sum_{n=1}^{N} y_n z_n p_n$. $\frac{\partial F}{\partial B} = \frac{-1}{N}\sum_{n=1}^{N} y_n p_n$

$$= \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n^2 (1-p_n)p_n, & \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n (1-p_n)p_n \\ \frac{1}{N}\sum_{n=1}^{N} y_n^2 z_n (1-p_n)p_n, & \frac{1}{N}\sum_{n=1}^{N} y_n^2 (1-p_n)p_n \end{bmatrix} \quad \#$$

Problem 3

③ Gaussian kernel: $k(x, x') = \exp(-\gamma\|x - x'\|^2)$

when $\gamma \to \infty$ $k(x, x') = \begin{cases} 1. & \text{if } x = x' \\ 0. & \text{if } x \neq x' \end{cases}$

consider the dual problem for SVM:

$$\min_\alpha \left(\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m k(x_n, x_m) - \sum_{n=1}^{N} \alpha_n \right)$$

$$= \min_\alpha \left(\frac{1}{2} \sum_{n=1}^{N} \alpha_n^2 - \sum_{n=1} \alpha_n \right) \quad (\because k(x_n, x_m) = 0 \text{ if } n \neq m)$$

$$= \min_\alpha \left(\frac{1}{2} \sum_{n=1}^{N} (\alpha_n^2 - 2\alpha_n) \right) = \min_\alpha \left(\frac{1}{2} \sum_{n=1}^{N} (\alpha_n - 1)^2 - 1 \right)$$

minimum happens when $\alpha_n = 1$, $n = 1, 2, \cdots N$.

check whether the solution satisfy the constrain:

① $\sum_{n=1}^{N} \alpha_n y_n = 0 \Rightarrow$ True, because there are same number of positive and engtive examples.

② $0 \leq \alpha_n \leq C \Rightarrow$ True. $\because C > 1$

$\therefore$ The optimal α is all-1 vector #

Problem 4

④

Let input $x \sim U(0,1)$

$E(x) = \frac{1+0}{2} = \frac{1}{2}$ $Var(x) = \frac{(1-0)^2}{12} = \frac{1}{12}$

$E(x^2) = [E(x)]^2 + Var(x) = \frac{1}{4} + \frac{1}{12} = \frac{1}{3}$

Let two examples generated for each time:

$(x_1, x_1 - x_1^2)$, $(x_2, x_2 - x_2^2)$.

The linear regression model would find the linear function fit the training set perfectly. ($E_{in} = 0$).

The function: $y - (x_1 - x_1^2) = \frac{(x_2 - x_2^2) - (x_1 - x_1^2)}{x_2 - x_1}(x - x_1)$

Let the expeted value hypothesis: $h(x) = Wx + b$.

$$W = E\left[\frac{(x_2 - x_2^2) - (x_1 - x_1^2)}{x_2 - x_1}\right] = E\left[\frac{(x_2 - x_1) - (x_2 - x_1)(x_2 + x_1)}{x_2 - x_1}\right]$$

$$= E[1 - (x_2 + x_1)] = 1 - E[x_2] - E[x_1] = 1 - \frac{1}{2} - \frac{1}{2} = 0.$$

$$b = E\left[\frac{(x_2 - x_2^2) - (x_1 - x_1^2)}{x_2 - x_1}(-x_1) + (x_1 - x_1^2)\right]$$

$$= E[-x_1 + x_1^2 + x_1 x_2 + x_1 - x_1^2] = E[x_1 x_2]$$

$$= E[x_1] \cdot E[x_2] \text{ (}\because\text{ generated independently)} = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

$\therefore$ Expected value of hypothesis: $h(x) = \frac{1}{4}$ #

Problem 5

(5)

My psuedo data : $(\tilde{x}_n, \tilde{y}_n) = (x_n\sqrt{u_n},\ y_n\sqrt{u_n})$ $n=1\ldots N$

Let w be the optimal solution for pseudo data

I want to prove that w is also the optimal solution for original data.

Assume w' is the optimal solution for original data and w' is different from w. That is,

$$E_{in}(w') = \frac{1}{N}\sum_{n=1}^{N} u_n(y_n - w'^T x_n)^2 < \frac{1}{N}\sum_{n=1}^{N} u_n(y_n - w^T x_n)^2$$

$$\Rightarrow \frac{1}{N}\sum_{n=1}^{N}(y_n\sqrt{u_n} - w'^T x_n\sqrt{u_n})^2 < \frac{1}{N}\sum_{n=1}^{N}(y_n\sqrt{u_n} - w^T x_n\sqrt{u_n})^2$$

$\Rightarrow$ w' is more optimal than w on psuedo dada $\Rightarrow$ contradiction

$\therefore$ w is also the optimal for original data

$\therefore$ Solve the optimization problem of linear regression on my psuedo data also solve the original problem.

#

Problem 6

(6)

Before the first iteration, weight for all examples are all $\frac{1}{N}$. $u_+^{(1)} = u_-^{(1)} = \frac{1}{N}$

with the constant classifier $g_1(x) = +1$.

$$\epsilon = \frac{\text{negtive example numble}}{\text{all example numbe}} = 1-0.78 = 0.22$$

$$u_+^{(2)} = u_+^{(1)} / \sqrt{\frac{1-\epsilon}{\epsilon}} \qquad u_-^{(2)} = u_-^{(1)} \cdot \sqrt{\frac{1-\epsilon}{\epsilon}}$$

$$\frac{u_+^{(2)}}{u_-^{(2)}} = \frac{u_+^{(1)}}{u_-^{(1)}} \sqrt{\frac{\epsilon}{1-\epsilon}} \sqrt{\frac{\epsilon}{1-\epsilon}} = \frac{\epsilon}{1-\epsilon} = \frac{0.22}{0.78} = \frac{11}{39}$$ #

Problem 7

⑦

consider the extrem case :

g^+ that predict +1 for $x \in X$ (e.g. $\theta = -6$, $s = +1$)

g^- that predit -1 for $x \in X$ (e.g. $\theta = -6$, $s = -1$)

and consider the first dimension (x_1) all the possible values are $[-M, M]$.

we consider g that predic +1 for some x, and -1 for others.

There are $2M = 2 \cdot 5 = 10$ intervals

There are 2 hypothesis for each θ in these intervals ($\because s \in \{-1, +1\}$)

$\Rightarrow$ There are $2 \times 10 = 20$ hypothesis for first dimension.

$d = 2$, and consider g^+ and g^- I declare before,

There are total $20 \cdot 2 + 2 = 42$ different decision stumps #

Problem 8

(8)

· Consider the hypothesis g^{+} that always predict $+1$
g^{-} that always predict -1.

$g_{+}(x)g_{+}(x') = 1$. $g_{-}(x)g_{-}(x') = 1$

consider any hypothesis $g = s \cdot \text{sign}(x_i - \theta)$.

$s \in \{-1, +1\}$, $\theta \in R$, $i \in \{1, 2, \dots, d\}$.

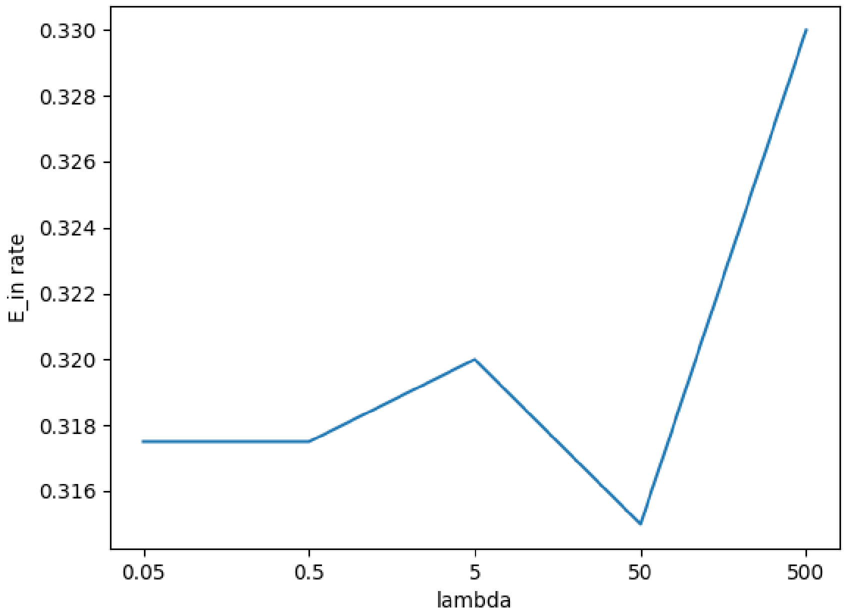
$$g(x)g(x') = s^2 \,\text{sign}(x_i - \theta) \cdot \text{sign}(x'_i - \theta)$$
$$= \text{sign}\left((x_i - \theta)(x'_i - \theta)\right)$$

for each θ, there are 2 hypothesis ($s = +1$ or -1).
but have the same value $g(x)g'(x')$

$$\therefore (\phi_{ds}(x)^T)(\phi_{ds}(x))$$
$$= 2 + \sum_{i=1}^{d} \sum_{m=-M}^{M-1} 2 \cdot \text{sign}\left((x_i - m - 0.5)(x'_i - m - 0.5)\right)$$

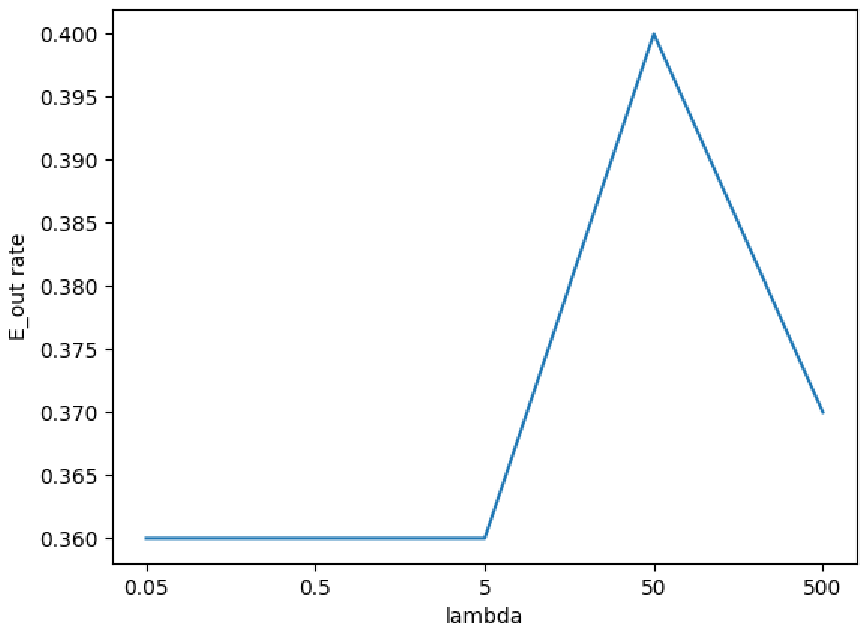
(I select $\theta = m + 0.5$ for each interval)

Problem 9



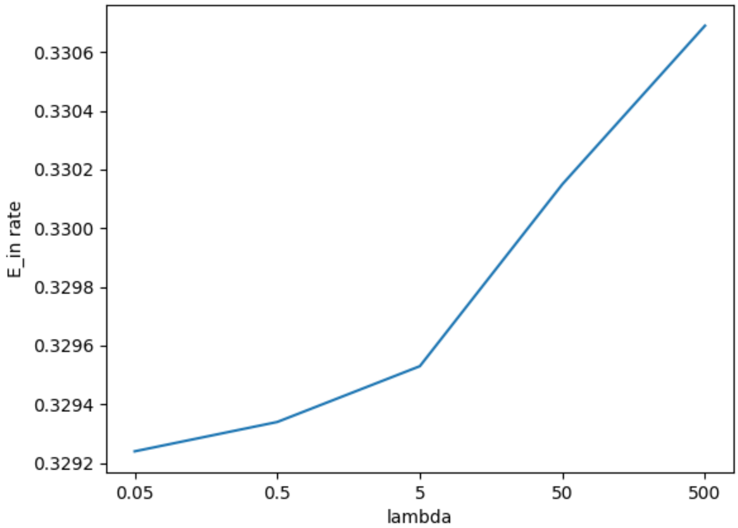
When lambda = 50, reaches the minimum E_in(g) value 0.315

Problem 10



When lambda = 0.05 or lambda = 0.5, reaches the minimum E_out(g) value 0.36

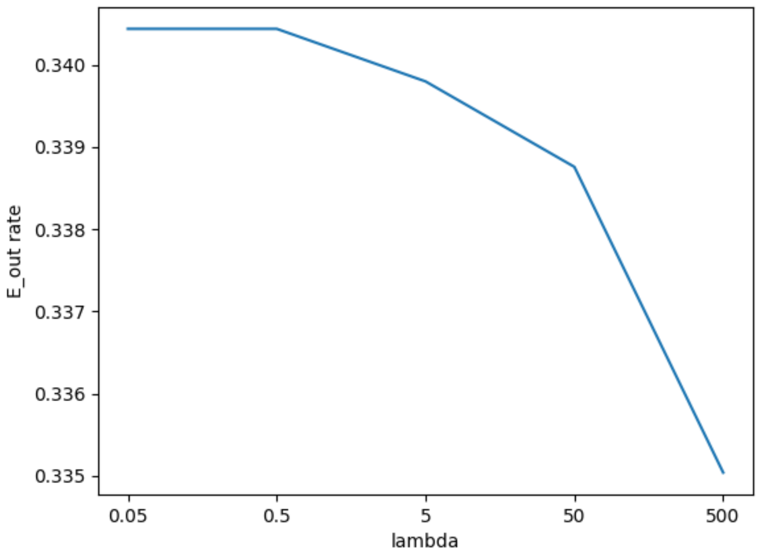
Problem 11



lambda = 0.05 reaches the minimum E_in(G) = 0.3292

Compare with the result of the problem 9, we can see that when using bootstrap for many iterations, E_in(G) rate increases with bigger lambda in average, and the result is similar among several experiments.

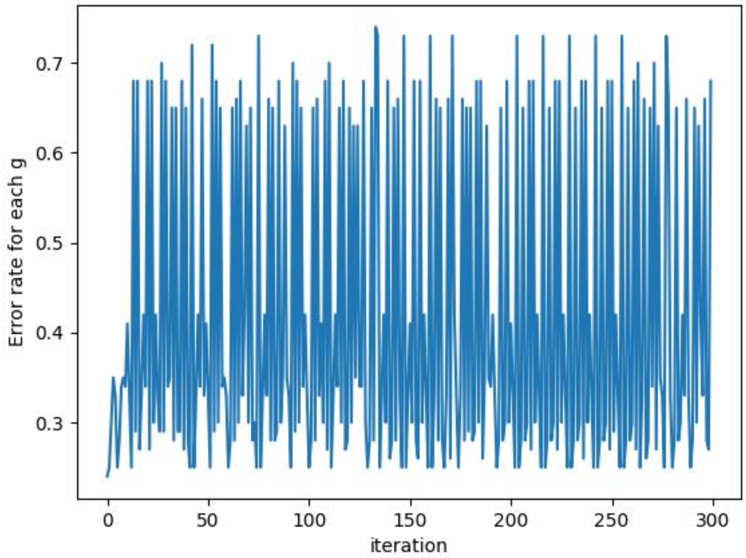
Problem 12



lambda = 500 reaches the minimum E_out(G) = 0.335.

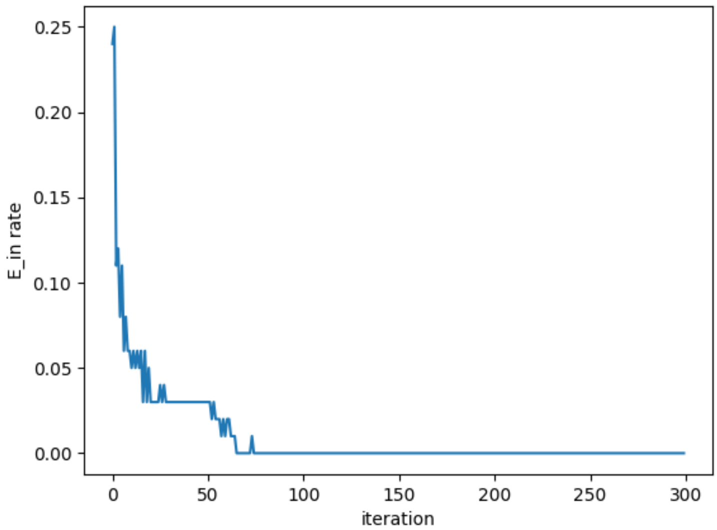
Compare with the result of the problem 10, we can see when using bootstrap for many times, E_out(G) decreases with bigger lambda value in average. Using bootstrap leads to more stable result, and is useful for parameter tuning.

Problem 13



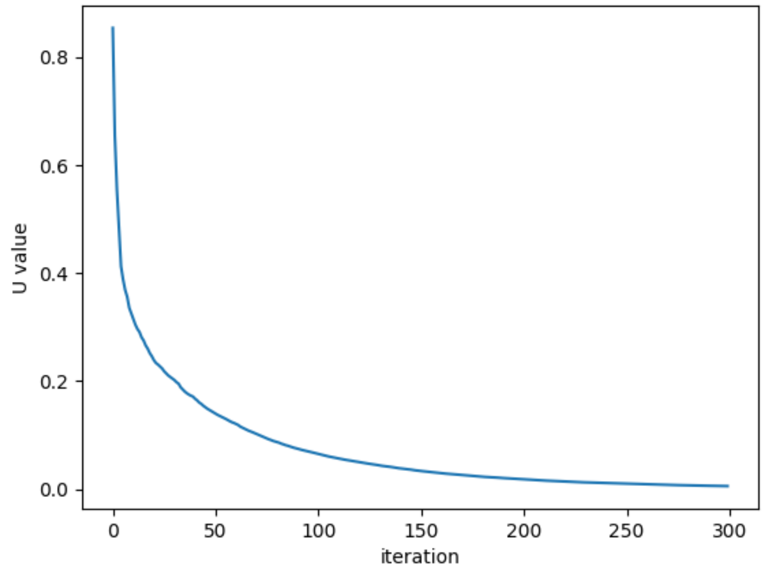
E_in(gT) = 0.68 (evaluated without weights for each sample). The E_in(g) is very unstable and random in each iteration, I think it's because after re-weighting each time, the new hypothesis tries to fit the samples with bigger weight, which might not be the majority of the data, and the E_in value is not good.

Problem 14



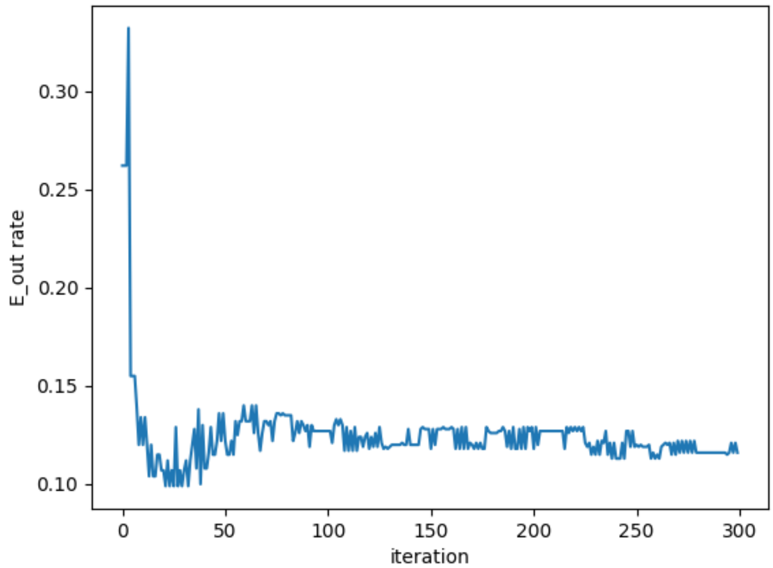
E_in(GT) = 0. The E_in(G) decreases roughly for each iteration, so we can see that for each iteration, the new hypothesis and its alpha value somehow helps the performance of the G. I think its because the new hypothesis will try to correctly classify the misclassified samples for previous hypothesis, and make them classified correctly at the end.

Problem 15



UT = 0.0054. the U value is decreasing for each iteration. See the proof in problem 17, if the hypothesis for each iteration is not too bad (error rate < 0.5), then the new U value is the old one multiplied with some value < 1, so it's decreasing.

Problem 16



E_out(G) = 0.132. The E_out(G) decreases roughly for each iteration, and it's not overfitting! Adaboost is quite a surprising machine learning algorithm that have great power and can generalize well. I think the reason that leads to not overfitting easily is the target of the algorithm can be regarded as finding a large margin in a high dimensional vector space.

Problem 17

(1)

At the beginning, the weights are uniform for all sample. $\because u_{n,1} = \frac{1}{N}$ $\therefore U_1 = \sum_{n=1}^{N} u_{n,1} = 1$.

$$u_{n,t+1} = \begin{cases} u_{n,t}\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, & \text{if } y(x_n) \neq g_t(x_n) \\ u_{n,t}/\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}, & \text{if } y(x_n) = g_t(x_n) \end{cases}$$

$$U_{t+1} = \sum_{n=1}^{N} u_{n,t+1} = \sum_{\forall n\, y \neq g_t} u_{n,t} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + \sum_{\forall n,\, y = g_t} u_{n,t} / \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

$$= \sum_{\forall n} u_{n,t} \cdot \left(\frac{\sum_{\forall n\, y \neq g_t} u_{n,t}}{\sum_{\forall n} u_{n,t}} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + \frac{\sum_{\forall n\, y = g_t} u_{n,t}}{\sum_{\forall n} u_{n,t}} \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \right)$$

$$= U_t \cdot \left(\epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} + (1-\epsilon_t) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} \right)$$

$$= U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)}$$

$x(1-x) = -(x-\frac{1}{2})^2 + \frac{1}{4}$ $\quad \because \epsilon_t \leq \epsilon < \frac{1}{2}$ $\therefore \epsilon_t - \frac{1}{2} \leq \epsilon - \frac{1}{2}$

$\therefore \epsilon_t(1-\epsilon_t) \leq \epsilon(1-\epsilon)$ $\therefore 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq 2\sqrt{\epsilon(1-\epsilon)}$

$\therefore U_{t+1} = U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)} \leq U_t \cdot 2\sqrt{\epsilon(1-\epsilon)}$ #

Problem 18

(18)

$$E_{in}(G_T) = \frac{1}{N}\sum_{n=1}^{N}\left[y_n \sum_{t=1}^{T}\alpha_t g_t(x_n) \le 0 \right]$$

$$\le \frac{1}{N}\sum_{n=1}^{N} \exp\left(-y_n \sum_{t=1}^{T}\alpha_t g_t(x_n)\right)$$

$$= \sum_{n=1}^{N} u_n^{(T+1)} = U_{T+1}$$

$U_1 = 1$. $U_{T+1} = U_T \cdot 2\sqrt{\epsilon_T(1-\epsilon_T)} = U_1 \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)}$

$$\le \prod_{t=1}^{T} \exp\left(-2\left(\tfrac{1}{2}-\epsilon_t\right)^2\right) \quad \left(\because \sqrt{\epsilon(1-\epsilon)} \le \tfrac{1}{2}\exp\left(-2\left(\tfrac{1}{2}-\epsilon\right)^2\right)\right)$$

$$= \exp\left(-2\sum_{t=1}^{T}\left(\tfrac{1}{2}-\epsilon_t\right)^2\right)$$

We want $E_{in}(G_T) \le U_{T+1} \le \frac{1}{N} \Rightarrow E_{in}(G_T) = 0$.

$\therefore \exp\left(-2\sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2\right) \le \frac{1}{N} \Rightarrow \sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2 \ge \frac{1}{2}\ln(N)$

let $\left(\frac{1}{2}-\epsilon_t\right)^2 \ge k > 0$, for all t,

$\because \sum_{t=1}^{T}\left(\frac{1}{2}-\epsilon_t\right)^2 \ge T \cdot k \ge \frac{1}{2}\ln(N)$ if $T \ge \frac{\ln(N)}{2k}$.

$\therefore$ after $T = O(\log N)$ iteration.

$E_{in}(G_T) \le U_{T+1} \le \frac{1}{N} \Rightarrow E_{in}(G_T) = 0$