
Machine Learning

Answer Sheet for Homework 6

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

December 23, 2015

Problem 1

With the definition of z_n , rewrite the equation

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

So

$$\frac{\partial F}{\partial A} = \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{\exp(-y_n(Az_n + B))}{1 + \exp(-y_n(Az_n + B))} \right)^T z_n = \frac{1}{N} \sum_{n=1}^N -y_n p_n^T z_n \quad (2)$$

$$\frac{\partial F}{\partial B} = \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{\exp(-y_n(Az_n + B))}{1 + \exp(-y_n(Az_n + B))} \right)^T = \frac{1}{N} \sum_{n=1}^N -y_n p_n^T \quad (3)$$

Hence

$$\nabla F(A, B) = \frac{1}{N} \sum_{n=1}^N [-y_n z_n p_n, -y_n p_n]^T \quad (4)$$

□

Problem 2

Use the result of Problem 1 and define $\exp(-y_n(Az_n + B)) = \exp(\xi_n)$, we have

$$\frac{\partial^2 F}{\partial A^2} = \frac{\partial}{\partial A} \left(\frac{1}{N} \sum_{n=1}^N -y_n p_n^T z_n \right) \quad (5)$$

$$= \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{-y_n \exp(\xi_n) (1 + \exp(\xi_n)) z_n - y_n (\exp(\xi_n))^2 z_n}{(1 + \exp(\xi_n))^2} \right) z_n \quad (6)$$

$$= \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{-y_n \exp(\xi_n) z_n}{(1 + \exp(\xi_n))^2} \right) z_n \quad (7)$$

$$= \frac{1}{N} \sum_{n=1}^N (y_n)^2 \left(\frac{\exp(\xi_n)}{1 + \exp(\xi_n)} \left(1 - \frac{\exp(\xi_n)}{1 + \exp(\xi_n)} \right) \right) z_n^2 \quad (8)$$

$$= \frac{1}{N} \sum_{n=1}^N z_n^2 p_n (1 - p_n) \quad (9)$$

where $y_n^2 = 1$ since $y_n \in \{-1, +1\}$.

The other term is

$$\frac{\partial^2 F}{\partial A \partial B} = \frac{\partial}{\partial A} \left(\frac{1}{N} \sum_{n=1}^N -y_n p_n^T \right) = \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{-y_n \exp(\xi_n) z_n}{(1 + \exp(\xi_n))^2} \right) \quad (10)$$

$$= \frac{1}{N} \sum_{n=1}^N z_n p_n (1 - p_n) \quad (11)$$

$$\frac{\partial^2 F}{\partial B \partial A} = \frac{\partial}{\partial B} \left(\frac{1}{N} \sum_{n=1}^N -y_n p_n^T z_n \right) = \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{-y_n \exp(\xi_n)}{(1 + \exp(\xi_n))^2} \right) z_n \quad (12)$$

$$= \frac{1}{N} \sum_{n=1}^N z_n p_n (1 - p_n) \quad (13)$$

$$\frac{\partial^2 F}{\partial B^2} = \frac{\partial}{\partial B} \left(\frac{1}{N} \sum_{n=1}^N -y_n p_n^T \right) = \frac{1}{N} \sum_{n=1}^N -y_n \left(\frac{-y_n \exp(\xi_n)}{(1 + \exp(\xi_n))^2} \right) \quad (14)$$

$$= \frac{1}{N} \sum_{n=1}^N p_n (1 - p_n) \quad (15)$$

Hence, we have

$$H(F) = \frac{1}{N} \sum_{n=1}^N \begin{bmatrix} z_n^2 p_n (1 - p_n) & z_n p_n (1 - p_n) \\ z_n p_n (1 - p_n) & p_n (1 - p_n) \end{bmatrix} \quad (16)$$

□

Problem 3

As $\gamma \rightarrow \infty$, we have

$$\lim_{\gamma \rightarrow \infty} \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2) = 0 \quad (17)$$

So K should be a zero matrix with size $N \times N$, which is $\mathbf{0}_{N \times N}$.

And β is

$$\beta = (\lambda I + K)^{-1} \mathbf{y} = \lambda^{-1} \mathbf{y} \quad (18)$$

□

Problem 4

If $|y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| \geq \epsilon$, then

$$\begin{cases} |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon = \xi_n^\wedge \text{ and } \xi_n^\vee = 0, & \text{if } y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b > 0 \\ |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon = \xi_n^\vee \text{ and } \xi_n^\wedge = 0, & \text{if } y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b < 0 \end{cases} \quad (19)$$

and if $|y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| < \epsilon$, then $\xi_n^\wedge = 0$ and $\xi_n^\vee = 0$. Hence, we have

$$(\xi_n^\wedge)^2 + (\xi_n^\vee)^2 = (\max(0, |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon))^2 \quad (20)$$

So P_2 is equivalent to

$$\min_{b, \mathbf{w}} \left(\frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{n=1}^N (\max(0, |y_n - \mathbf{w}^T \phi(\mathbf{x}_n) - b| - \epsilon))^2 \right) \quad (21)$$

with no constraints.

□

Problem 5

The first term is of course

$$\frac{\partial}{\partial \beta_m} \left(\frac{1}{2} \sum_{m=1}^N \sum_{n=1}^N \beta_n \beta_m K(\mathbf{x}_n, \mathbf{x}_m) \right) = \sum_{n=1}^N \beta_n K(\mathbf{x}_n, \mathbf{x}_m) \quad (22)$$

With \mathbf{w}_* , rewrite the result of Problem 4,

$$\text{something} = C \sum_{n=1}^N \left(\max \left(0, \left| y_n - \sum_{m=1}^N \beta_m K(\mathbf{x}_n, \mathbf{x}_m) - b \right| - \epsilon \right) \right)^2 \quad (23)$$

$$= C \sum_{n=1}^N (\max(0, |y_n - s_n| - \epsilon))^2 \quad (24)$$

Consider the following cases.

1. $|y_n - s_n| \geq \epsilon$.

Then we have

$$\frac{\partial}{\partial \beta_m} (\text{something}) = \frac{\partial}{\partial \beta_m} (C (|y_n - s_n| - \epsilon)^2) \quad (25)$$

$$= (2C (|y_n - s_n| - \epsilon)) \frac{\partial}{\partial \beta_m} |y_n - s_n| \quad (26)$$

$$= -2C (|y_n - s_n| - \epsilon) \text{sign} (y_n - s_n) \frac{\partial s_n}{\partial \beta_m} \quad (27)$$

$$= -2C (|y_n - s_n| - \epsilon) \text{sign} (y_n - s_n) K (\mathbf{x}_n, \mathbf{x}_m) \quad (28)$$

2. $|y_n - s_n| < \epsilon$.

Then we have

$$\frac{\partial}{\partial \beta_m} (\text{something}) = 0 \quad (29)$$

So we have

$$\frac{\partial F(b, \boldsymbol{\beta})}{\partial \beta_m} = \sum_{n=1}^N \beta_n K (\mathbf{x}_n, \mathbf{x}_m) - 2C \sum_{n=1}^N \mathbb{I} [|y_n - s_n| \geq \epsilon] (|y_n - s_n| - \epsilon) \text{sign} (y_n - s_n) K (\mathbf{x}_n, \mathbf{x}_m) \quad (30)$$

□

Problem 6

First, we have

$$e_t = \frac{1}{M} \sum_{m=1}^M (g_t (\tilde{\mathbf{x}}_m))^2 - 2g_t (\tilde{\mathbf{x}}_m) \tilde{y}_m + (\tilde{y}_m)^2 \quad (31)$$

And e_0 is

$$e_0 = \frac{1}{M} \sum_{m=1}^M (0)^2 - 2 \cdot 0 \cdot \tilde{y}_m + (\tilde{y}_m)^2 = \frac{1}{M} \sum_{m=1}^M (\tilde{y}_m)^2 \quad (32)$$

where we have used that $g_0 (\mathbf{x}) = 0, \forall \mathbf{x}$.

So e_t can be rewritten as

$$e_t = e_0 + \frac{1}{M} \sum_{m=1}^M (g_t (\tilde{\mathbf{x}}_m))^2 - 2g_t (\tilde{\mathbf{x}}_m) \tilde{y}_m = e_0 + s_t - \frac{2}{M} \sum_{m=1}^M g_t (\tilde{\mathbf{x}}_m) \tilde{y}_m \quad (33)$$

Hence,

$$\sum_{m=1}^M g_t(\tilde{\mathbf{x}}_m) \tilde{y}_m = \frac{M}{2} (e_0 + s_t - e_t) \quad (34)$$

□

Problem 7

Suppose the input is (a, b) with following cases.

1. $0 \leq a \leq b \leq 1$.

Then the output should be (a^2, b^2) . The line equation of these two points (a, a^2) and (b, b^2) is

$$y = \frac{b^2 - a^2}{b - a} (x - a) + a^2 \quad (35)$$

Then $\bar{g}_1(x)$ should be

$$\int_0^1 \int_0^b \left(\frac{b^2 - a^2}{b - a} (x - a) + a^2 \right) da db = \int_0^1 \left(\frac{1}{2} a^2 (x - b) + abx \right) \Big|_{a=0}^{a=b} db \quad (36)$$

$$= \int_0^1 \left(\frac{3}{2} b^2 x - \frac{1}{2} b^3 \right) db \quad (37)$$

$$= \left(\frac{1}{2} b^3 x - \frac{1}{8} b^4 \right) \Big|_{b=0}^{b=1} = \frac{1}{2} x - \frac{1}{8} \quad (38)$$

2. $0 \leq b < a \leq 1$.

Similarly, we have

$$\bar{g}_2(x) = \frac{1}{2} x - \frac{1}{8} \quad (39)$$

Hence, we have

$$\bar{g}(x) = \bar{g}_1(x) + \bar{g}_2(x) = x - \frac{1}{4} \quad (40)$$

This makes sense since $\bar{g}(x)$ and $f(x)$ should be the same at the average value of $[0, 1]$, which is

$$\bar{g}\left(\frac{1}{2}\right) = \frac{1}{2} - \frac{1}{4} = \frac{1}{4} = \left(\frac{1}{2}\right)^2 = f\left(\frac{1}{2}\right) \quad (41)$$

□

Problem 8

Now we have

$$u_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2 = u_n y_n^2 - 2u_n y_n \mathbf{w}^T \mathbf{x}_n + u_n (\mathbf{w}^T \mathbf{x}_n)^2 \quad (42)$$

This is equal to

$$u_n (y_n - \mathbf{w}^T \mathbf{x}_n)^2 = (\sqrt{u_n} y_n)^2 - 2(\sqrt{u_n} y_n) (\mathbf{w}^T (\sqrt{u_n} \mathbf{x}_n)) + (\mathbf{w}^T (\sqrt{u_n} \mathbf{x}_n))^2 \quad (43)$$

Hence, the pseudo data is $\{(\tilde{\mathbf{x}}_n, \tilde{y}_n)\}_{n=1}^N = \{(\sqrt{u_n} x_n, \sqrt{u_n} y_n)\}_{n=1}^N$.

□

Problem 9

With the rule of optimal re-weighting, we have

$$u_+^{(2)} = u^{(1)} \cdot 1\%, \quad u_-^{(2)} = u^{(1)} \cdot 99\% \Rightarrow \frac{u_+^{(2)}}{u_-^{(2)}} = \frac{1}{99} \quad (44)$$

□

Problem 10

Consider the following cases.

1. $1 < \theta \leq 6$:

Since $s \in \{+1, -1\}$, $d = 2$ and $R - L = 5$ regions to put θ_i , so there are $2 \times 2 \times 5 = 20$ different decision stumps.

2. $\theta \leq 1$ or $\theta > 6$:

Then there are only 2 decision stump: $g(\mathbf{x}) = +1$ with $(s = +1, \theta \leq 1)$ or $(s = -1, \theta > 6)$; $g(\mathbf{x}) = -1$ with $(s = +1, \theta > 6)$ or $(s = -1, \theta \leq 1)$ for all \mathbf{x} .

So there are

$$20 + 2 = 22 \quad (45)$$

different decision stumps.

Also, this can be generalized to

$$\underbrace{2}_{s \in \{+1, -1\}} \underbrace{d}_{\text{dimension}} \underbrace{(R - L)}_{\theta \text{ region}} + \underbrace{2}_{\text{left and right region}} = 2d(R - L) + 2 \quad (46)$$

□

Problem 11

First, we have

$$K_{ds}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{|\mathcal{G}|} (g_i(\mathbf{x}))^T g_i(\mathbf{x}') \quad (47)$$

$$= \sum_{i=1}^{|\mathcal{G}|} (s_i \text{sign}(x_j - \theta_i)) (s_i \text{sign}(x'_j - \theta_i)) \quad (48)$$

$$= \sum_{i=1}^{|\mathcal{G}|} \text{sign}(x_j - \theta_i) \text{sign}(x'_j - \theta_i) \quad (49)$$

where $(s_i)^2 = 1$ since $s_i \in \{+1, -1\}$.

Now if $x_j = x'_j$, applying the general result of Problem 11, we have

$$K_{ds}(\mathbf{x}, \mathbf{x}') = 2d(R - L) + 2 \quad (50)$$

since all g_i are the same.

If $x_j \neq x'_j$, there are $(x_j - x'_j)$ different output by decision stump g_i in j^{th} dimension with some fixed s_i . One different output causes the result from $+1$ to -1 , so the summation minus by 2 with each different output. Hence

$$K_{ds}(\mathbf{x}, \mathbf{x}') = 2d(R - L) + \underbrace{(-2)}_{\text{from } +1 \text{ to } -1} \times \underbrace{2}_{s \in \{+1, -1\}} \times \|\mathbf{x} - \mathbf{x}'\|_1 + 2 \quad (51)$$

$$= 2d(R - L) - 4\|\mathbf{x} - \mathbf{x}'\|_1 + 2 \quad (52)$$

where $\|\mathbf{x} - \mathbf{x}'\|_1$ denotes the one-norm of $(\mathbf{x} - \mathbf{x}')$.

□

Problem 12

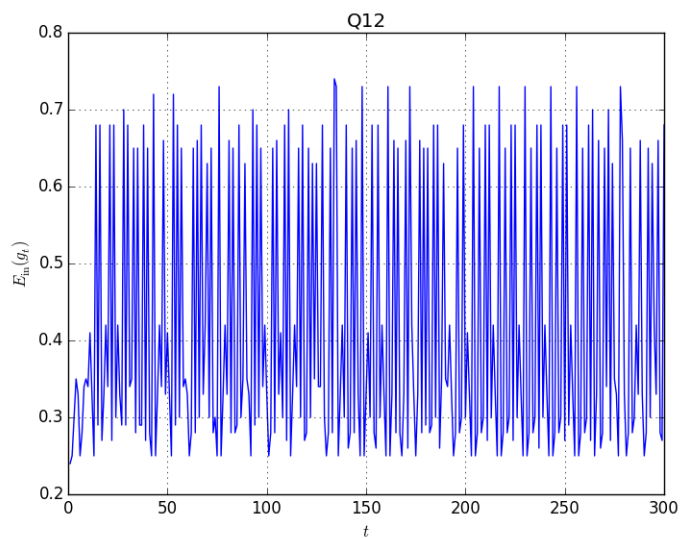


Figure 1: Q12

$E_{\text{in}} = 0.24$, $\alpha_1 = 0.576339754969$.

□

Problem 13

The result oscillates. Since re-weighting tries to correct the wrong result and ignore the correct result, this causes g_{t+1} and g_t to oscillate between 0.5. In fact, the average of $E_{\text{in}} = 0.42 \approx 0.5$.

□

Problem 14

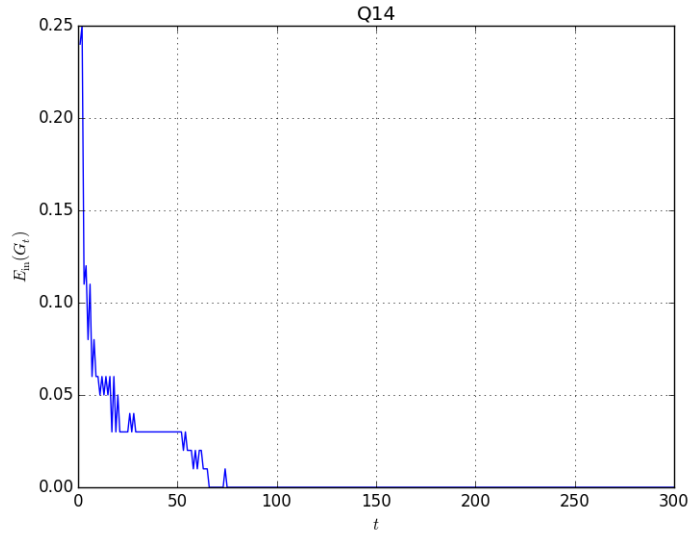


Figure 2: Q14

$$E_{\text{in}}(G) = 0.0.$$

□

Problem 15

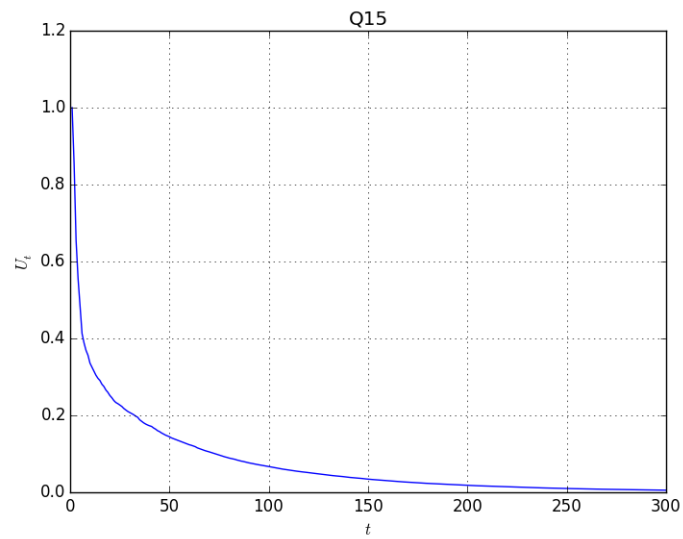


Figure 3: Q15

$$U_2 = 0.8542, U_T = 0.0055.$$

□

Problem 16

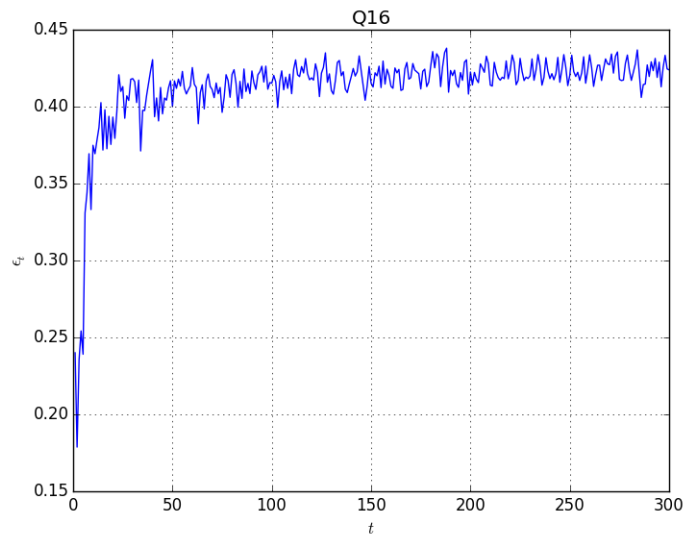


Figure 4: Q16

$$\min(\epsilon_t) = 0.17873.$$

□

Problem 17

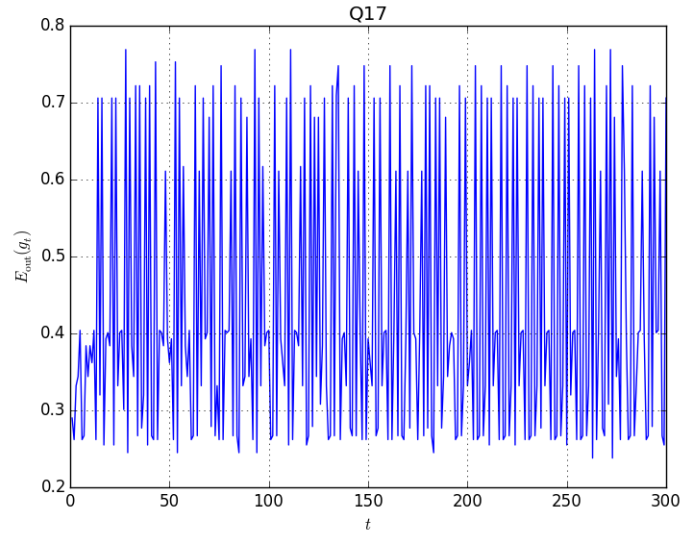


Figure 5: Q17

$$E_{\text{out}}(g_1) = 0.29.$$

□

Problem 18

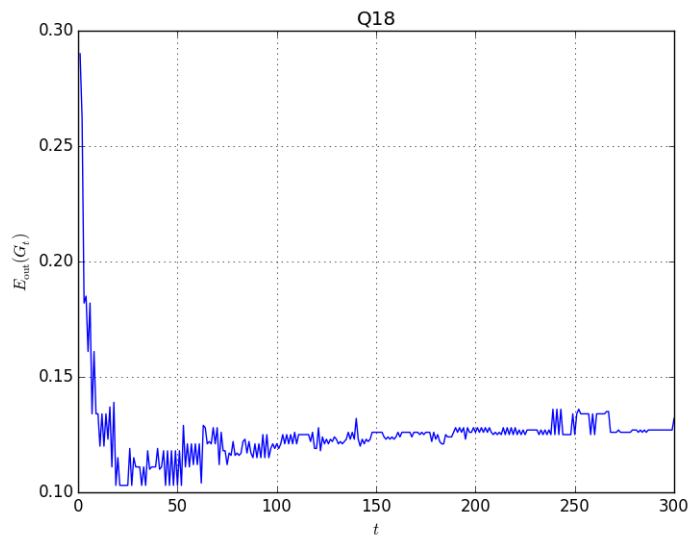


Figure 6: Q18

$$E_{\text{out}}(G) = 0.132.$$

□

Problem 19

Minimum $E_{\text{in}}(g)$ is 0.0, with $\lambda = 0.001$, $\gamma = 32$.

□

Problem 20

Minimum $E_{\text{out}}(g)$ is 0.39, with $\lambda = 1000$, $\gamma = 0.125$.

□

Problem 21

First, we have

$$U_1 = \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^0 \alpha_{\tau} g_{\tau}(\mathbf{x}_n) \right) = \frac{1}{N} \sum_{n=1}^N \exp(0) = \frac{1}{N} \times N = 1 \quad (53)$$

Then, we have

$$U_{t+1} = \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^t \alpha_{\tau} g_{\tau}(\mathbf{x}_n) \right) \quad (54)$$

$$= \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^{t-1} \alpha_{\tau} g_{\tau}(\mathbf{x}_n) - y_n \alpha_t g_t(\mathbf{x}_n) \right) \quad (55)$$

$$= \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^{t-1} \alpha_{\tau} g_{\tau}(\mathbf{x}_n) \right) \exp(-y_n \alpha_t g_t(\mathbf{x}_n)) \quad (56)$$

$$= \sum_{n=1}^N u_n^{(t)} \exp(-y_n \alpha_t g_t(\mathbf{x}_n)) \quad (57)$$

$$= \sum_{\substack{n \\ y_n \neq g_t(\mathbf{x}_n)}} u_n^{(t)} \exp(-y_n \alpha_t g_t(\mathbf{x}_n)) + \sum_{\substack{n \\ y_n = g_t(\mathbf{x}_n)}} u_n^{(t)} \exp(-y_n \alpha_t g_t(\mathbf{x}_n)) \quad (58)$$

$$= \sum_{\substack{n \\ y_n \neq g_t(\mathbf{x}_n)}} u_n^{(t)} \exp(\alpha_t) + \sum_{\substack{n \\ y_n = g_t(\mathbf{x}_n)}} u_n^{(t)} \exp(-\alpha_t) \quad (59)$$

$$= \exp(\alpha_t) (\epsilon_t) \sum_{n=1}^N u_n^{(t)} + \exp(-\alpha_t) (1 - \epsilon_t) \sum_{n=1}^N u_n^{(t)} \quad (60)$$

$$= U_t (\exp(\alpha_t) (\epsilon_t) + \exp(-\alpha_t) (1 - \epsilon_t)) = U_t \cdot 2\sqrt{\epsilon_t(1 - \epsilon_t)} \quad (61)$$

Since $\epsilon_t \leq \epsilon < 1/2$, we have

$$\epsilon_t (1 - \epsilon_t) \leq \epsilon (1 - \epsilon) \Rightarrow U_t \cdot 2\sqrt{\epsilon_t (1 - \epsilon_t)} \leq U_t \cdot 2\sqrt{\epsilon (1 - \epsilon)} \quad (62)$$

□

Problem 22

From the result of Problem 21, we have

$$U_{t+1} = U_t \cdot 2\sqrt{\epsilon_t (1 - \epsilon_t)} \leq U_t \cdot 2\sqrt{\epsilon (1 - \epsilon)} \leq U_t \exp \left(-2 \left(\frac{1}{2} - \epsilon \right)^2 \right), \quad \epsilon < \frac{1}{2} \quad (63)$$

At $t = T$, we have

$$0 \leq E_{\text{in}}(G_T) \leq U_{T+1} \leq U_T \exp \left(-2 \left(\frac{1}{2} - \epsilon \right)^2 \right) \quad (64)$$

And this can be rewritten as

$$0 \leq E_{\text{in}}(G_T) \leq U_T \exp \left(-2 \left(\frac{1}{2} - \epsilon \right)^2 \right) \leq U_{T-1} \exp \left(-4 \left(\frac{1}{2} - \epsilon \right)^2 \right) \quad (65)$$

$$\leq \cdots \leq U_1 \exp \left(-2T \left(\frac{1}{2} - \epsilon \right)^2 \right) = \exp \left(-2T \left(\frac{1}{2} - \epsilon \right)^2 \right) \quad (66)$$

Let $O(T) = O(\ln N)$, say that $T = \ell \ln N + k$, where $\ell, k \in \mathbb{R}$ and $\ell > 0$. Then we have

$$0 \leq E_{\text{in}}(G_T) \leq \exp \left(-2\ell \ln N \left(\frac{1}{2} - \epsilon \right)^2 - 2k \left(\frac{1}{2} - \epsilon \right)^2 \right) \quad (67)$$

Suppose ℓ is large enough that

$$-2\ell \ln N \left(\frac{1}{2} - \epsilon \right)^2 - 2k \left(\frac{1}{2} - \epsilon \right)^2 \approx -2\ell \ln N \quad (68)$$

So

$$0 \leq E_{\text{in}}(G_T) \leq \exp(-2\ell \ln N) = \frac{1}{N^{2\ell}} \quad (69)$$

As ℓ large enough, we have $1/N^{2\ell} \approx 0 \Rightarrow E_{\text{in}}(G_T) = 0$.

□

Reference

- [1] Lecture Notes by Hsuan-Tien LIN, Department of Computer Science and Information Engineering, National Taiwan University, Taipei 106, Taiwan.