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

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

So

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

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

Hence

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

□

Problem 2

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

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

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

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

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

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

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

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

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

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

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

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

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

Hence, we have

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

□

Problem 3

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

$$\lim_{\gamma\to\infty} \exp\left(-\gamma \left\|\mathbf{x}_n - \mathbf{x}_m\right\|^2\right) = 0 \tag{17}$$

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

$$\boldsymbol{\beta} = \left(\lambda I + K\right)^{-1}\mathbf{y} = \lambda^{-1}\mathbf{y} \tag{18}$$

□

Problem 4

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

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

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

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

So P_2 is equivalent to

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

with no constraints.

□

Problem 5

The first term is of course

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

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

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

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

Consider the following cases.

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

   Then we have

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

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

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

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

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

   Then we have

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

So we have

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

□

Problem 6

First, we have

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

And e_0 is

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

where we have used that $g_0 \left(\mathbf{x} \right) = 0$, $\forall \mathbf{x}$.
So e_t can be rewritten as

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

Hence,

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

□

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

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

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

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

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

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

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

   Similarly, we have

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

Hence, we have

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

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

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

□

Problem 8

Now we have

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

This is equal to

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

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

□

Problem 9

With the rule of optimal re-weighting, we have

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

□

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

$$20 + 2 = 22 \tag{45}$$

different decision stumps.
Also, this can be generalized to

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

□

Problem 11

First, we have

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

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

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

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

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

since all g_i are the same.
If $x_j \neq x'_j$, there are $\left(x_j - x'_j\right)$ 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

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

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

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

□

Problem 12

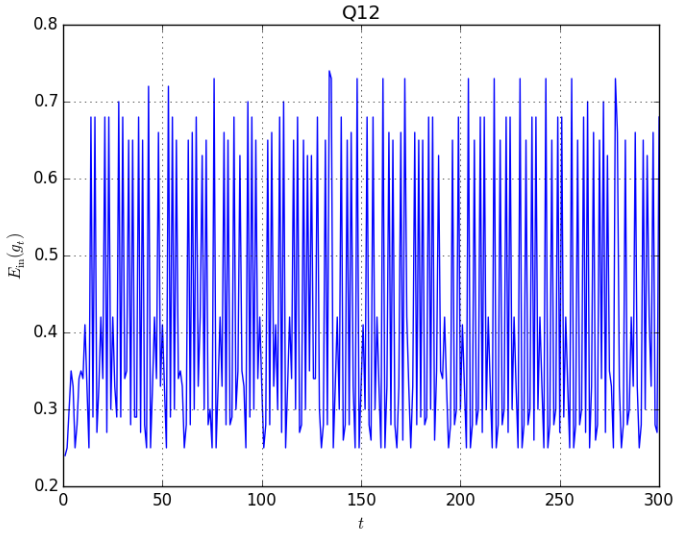

Figure 1: Q12

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

□

Problem 13

The result oscillates. Since re-weighting causes g_{t+1} and g_t to be very different in each iteration. So it oscillates.

□

Problem 14

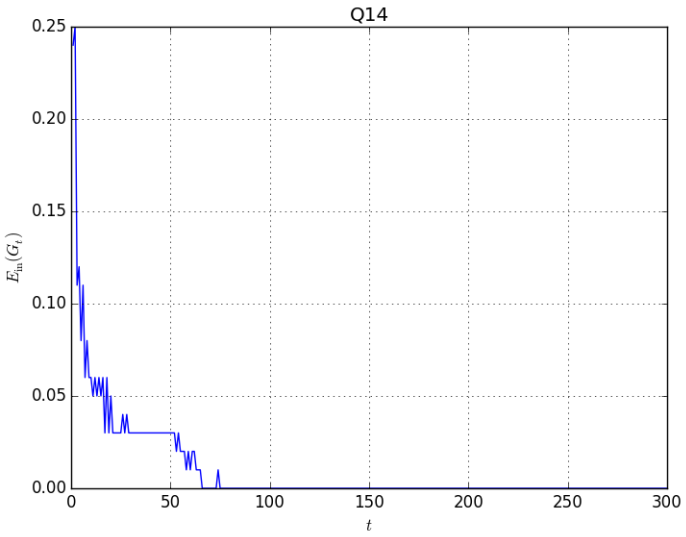

Figure 2: Q14

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

□

Problem 15

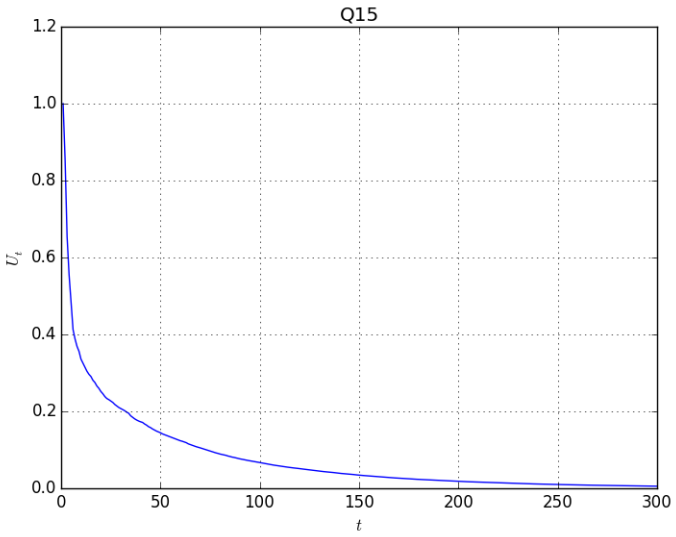

Figure 3: Q15

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

□

Problem 16

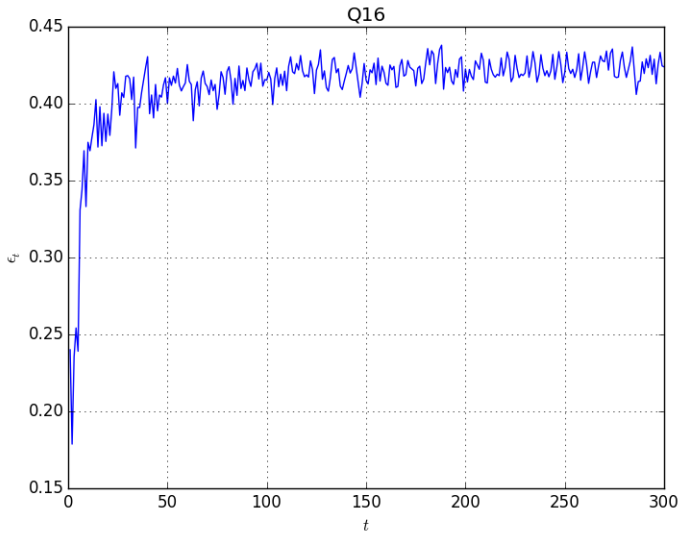

Figure 4: Q16

$\min(\epsilon_t) = 0.17873$.

□

Problem 17

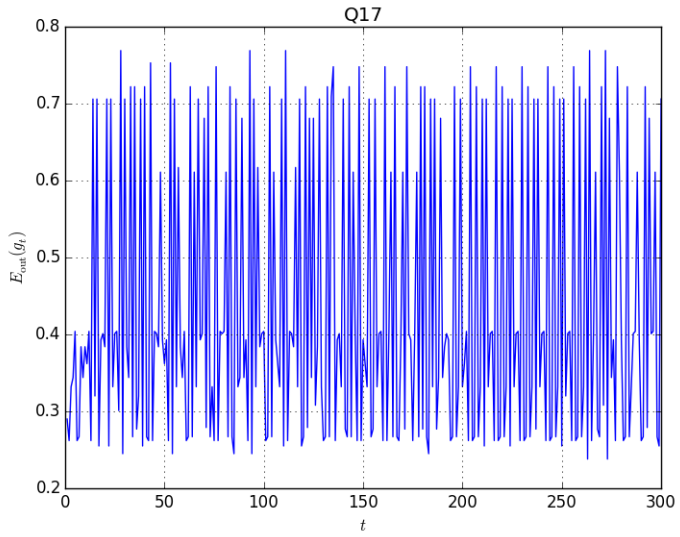

Figure 5: Q17

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

□

Problem 18

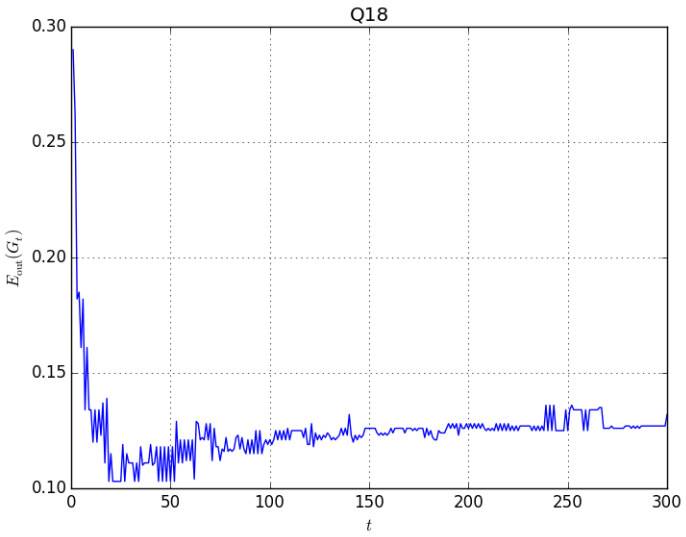

Figure 6: Q18

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

□

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

□

Problem 22

□

Reference

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