
Machine Learning

Answer Sheet for Homework 7

Da-Min HUANG

R04942045

Graduate Institute of Communication Engineering, National Taiwan University

January 6, 2016

Problem 1

Set $\mu_- = 1 - \mu_+$, we have

$$1 - \mu_+^2 - \mu_-^2 = 1 - \mu_+^2 - (1 - \mu_+)^2 = (1 - \mu_+)(1 + \mu_+) - (1 - \mu_+)^2 \quad (1)$$

$$= 2\mu_+(1 - \mu_+) = -2\mu_+^2 + 2\mu_+ = -2\left(\mu_+ - \frac{1}{2}\right)^2 + \frac{1}{2} \quad (2)$$

$$\leq \frac{1}{2} \quad (3)$$

Hence, if $\mu_+ = 1/2 \in [0, 1]$, then the maximum value of Gini index is $1/2$.

□

Problem 2

The normalized Gini index is

$$\frac{(1 - \mu_+^2 - \mu_-^2)}{\left(\frac{1}{2}\right)} = 2(1 - \mu_+^2 - \mu_-^2) \quad (4)$$

The squared error can be rewritten as

$$\mu_+(1 - (\mu_+ - \mu_-))^2 + \mu_-(-1 - (\mu_+ - \mu_-))^2 = 4\mu_+(1 - \mu_+)^2 + 4\mu_+^2(1 - \mu_+) \quad (5)$$

$$= 4\mu_+(1 - \mu_+) \leq 4 \times \frac{1}{4} = 1 \quad (6)$$

Hence the normalized squared error is

$$4\mu_+(1 - \mu_+) = 2(2\mu_+(1 - \mu_+)) = 2((1 - \mu_+)(1 + \mu_+) - (1 - \mu_+)^2) \quad (7)$$

$$= 2(1 - \mu_+^2 - \mu_-^2) \quad (8)$$

which is equal to normalized Gini index.

□

Problem 3

The probability of one example not sampled is

$$\left(1 - \frac{1}{N}\right)^{pN} = \frac{1}{\left(\frac{N}{N-1}\right)^{pN}} = \frac{1}{\left(1 + \frac{1}{N-1}\right)^{pN}} = \left(\frac{1}{\left(1 + \frac{1}{N-1}\right)^N}\right)^p \quad (9)$$

As $N \rightarrow \infty$, we have

$$\lim_{N \rightarrow \infty} \left(\frac{1}{\left(1 + \frac{1}{N-1}\right)^N}\right)^p = \left(\lim_{N \rightarrow \infty} \frac{1}{\left(1 + \frac{1}{N-1}\right)^N}\right)^p = \left(\frac{1}{e}\right)^p = e^{-p} \quad (10)$$

So there approximately $e^{-p} \cdot N$ of the examples not sampled.

□

Problem 4

Since $G = \text{Uniform}(\{g_k\}_{k=1}^3)$, so if at least two terms of $\{g_k\}_{k=1}^3$ output wrong result, then G outputs wrong result. Let $\{E_k\}_{k=1}^3$ be the set of examples that $\{g_k\}_{k=1}^3$ got wrong results. Apparently $|E_3| > |E_2| > |E_1|$ and $|E_1| + |E_2| > |E_3|$. So

1. Maximum of $E_{\text{out}}(G)$ happens at $E_3 \subset (E_1 \cup E_2)$. Then G outputs wrong result in the region of E_3 with $E_{\text{out}}(G) = 0.35$.
2. Minimum of $E_{\text{out}}(G)$ happens at $E_i \cap E_j = \emptyset$, $i \neq j$ and $1 \leq i, j \leq 3$ with $i, j \in \mathbb{N}$.
Then G always outputs the correct result since $(E_1 \cup E_2 \cup E_3) \subset \{\text{all examples}\}$.

Hence, $0 \leq E_{\text{out}}(G) \leq 0.35$.

□

Problem 5

Since $G = \text{Uniform}(\{g_k\}_{k=1}^K)$, so if at least $(K+1)/2$ terms of $\{g_k\}_{k=1}^K$ output wrong result, then G outputs wrong result. Let $\{E_k\}_{k=1}^K$ be the set of examples that $\{g_k\}_{k=1}^K$ got wrong results.

If G outputs wrong result on some example \mathbf{x} , then we have

$$\mathbf{x} \in \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \quad (11)$$

where α_i is some index satisfies $1 \leq \alpha_i \leq K$ and $m \in (\mathbb{N} \cup \{0\})$ with $0 \leq m < (K+1)/2$. And

$$\left| \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \right| \leq \frac{2}{K+1+2m} \sum_{k=1}^K e_k \leq \frac{2}{K+1} \sum_{k=1}^K e_k \quad (12)$$

(12) holds due to

$$\bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \subseteq E_{\beta} \quad (13)$$

where β is some index such that $|E_{\beta}| = \min_{\alpha_i} |E_{\alpha_i}|$. So

$$\left(\frac{K+1}{2} + m \right) \left| \bigcap_{i=1}^{((K+1)/2)+m} E_{\alpha_i} \right| \leq \left(\frac{K+1}{2} + m \right) |E_{\beta}| \leq \sum_{k=1}^K |E_k| \quad (14)$$

(14) holds since size of E_{β} is the smallest among $((K+1)/2 + m)$ terms and $\sum_{k=1}^K |E_k|$ must contains the $((K+1)/2 + m)$ terms.

Hence, we have

$$E_{\text{out}}(G) \leq \frac{2}{K+1} \sum_{k=1}^K e_k \quad (15)$$

□

Problem 6

By the definition of U_t , we have

$$U_{t+1} = \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^t \alpha_\tau g_\tau (\mathbf{x}_n) \right) \quad (16)$$

$$= \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^{t-1} \alpha_\tau g_\tau (\mathbf{x}_n) - y_n \alpha_t g_t (\mathbf{x}_n) \right) \quad (17)$$

$$= \frac{1}{N} \sum_{n=1}^N \exp \left(-y_n \sum_{\tau=1}^{t-1} \alpha_\tau g_\tau (\mathbf{x}_n) \right) \exp (-y_n \alpha_t g_t (\mathbf{x}_n)) \quad (18)$$

$$= \sum_{n=1}^N u_n^{(t)} \exp (-y_n \alpha_t g_t (\mathbf{x}_n)) \quad (19)$$

$$= \sum_{\substack{n \\ y_n \neq g_t (\mathbf{x}_n)}} u_n^{(t)} \exp (-y_n \alpha_t g_t (\mathbf{x}_n)) + \sum_{\substack{n \\ y_n = g_t (\mathbf{x}_n)}} u_n^{(t)} \exp (-y_n \alpha_t g_t (\mathbf{x}_n)) \quad (20)$$

$$= \sum_{\substack{n \\ y_n \neq g_t (\mathbf{x}_n)}} u_n^{(t)} \exp (\alpha_t) + \sum_{\substack{n \\ y_n = g_t (\mathbf{x}_n)}} u_n^{(t)} \exp (-\alpha_t) \quad (21)$$

$$= \exp (\alpha_t) (\epsilon_t) \sum_{n=1}^N u_n^{(t)} + \exp (-\alpha_t) (1 - \epsilon_t) \sum_{n=1}^N u_n^{(t)} \quad (22)$$

$$= U_t (\exp (\alpha_t) (\epsilon_t) + \exp (-\alpha_t) (1 - \epsilon_t)) = U_t \cdot 2\sqrt{\epsilon_t (1 - \epsilon_t)} \quad (23)$$

Since

$$U_1 = \sum_{n=1}^N u_n^{(1)} = \sum_{n=1}^N \frac{1}{N} = 1 \quad (24)$$

we have

$$U_3 = U_2 \cdot 2\sqrt{\epsilon_2 (1 - \epsilon_2)} = \left(U_1 \cdot 2\sqrt{\epsilon_1 (1 - \epsilon_1)} \right) \cdot 2\sqrt{\epsilon_2 (1 - \epsilon_2)} \quad (25)$$

$$= 4\sqrt{\epsilon_1 \epsilon_2 (1 - \epsilon_1) (1 - \epsilon_2)} \quad (26)$$

which can be generalized as

$$U_{T+1} = \prod_{t=1}^T \left(2\sqrt{\epsilon_t (1 - \epsilon_t)} \right) \quad (27)$$

□

Problem 7

To compute s_n , we need to find the optimal η of

$$\min_{\eta} \frac{1}{N} \sum_{n=1}^N ((y_n - s_n) - \eta g_t(\mathbf{x}_n))^2 := A \quad (28)$$

From $\partial A / \partial \eta = 0$, we have

$$\eta = \frac{\sum_{n=1}^N g_t(\mathbf{x}_n) (y_n - s_n)}{\sum_{n=1}^N g_t^2(\mathbf{x}_n)} \quad (29)$$

Now $s_n = 0$ and $g_1(\mathbf{x}) = 2$, so

$$\eta = \frac{2 \sum_{n=1}^N y_n}{4 \sum_{n=1}^N 1} = \frac{1}{2N} \sum_{n=1}^N y_n \quad (30)$$

Since $\eta = \alpha_1$, so

$$\alpha_1 g_1(\mathbf{x}_n) = \frac{2}{2N} \sum_{n=1}^N y_n = \frac{1}{N} \sum_{n=1}^N y_n = s_n \quad (31)$$

□

Problem 8

From the equatio of optimal η , we have

$$\eta = \frac{\sum_{n=1}^N g_t(\mathbf{x}_n) (y_n - s'_n)}{\sum_{n=1}^N g_t^2(\mathbf{x}_n)} = \frac{\sum_{n=1}^N y_n g_t(\mathbf{x}_n) - \sum_{n=1}^N s'_n g_t(\mathbf{x}_n)}{\sum_{n=1}^N g_t^2(\mathbf{x}_n)} = \alpha_t \quad (32)$$

so

$$\sum_{n=1}^N y_n g_t(\mathbf{x}_n) - \sum_{n=1}^N s'_n g_t(\mathbf{x}_n) = \alpha_t \sum_{n=1}^N g_t^2(\mathbf{x}_n) = \sum_{n=1}^N \alpha_t g_t^2(\mathbf{x}_n) = \sum_{n=1}^N (s_n - s'_n) g_t(\mathbf{x}_n) \quad (33)$$

where s'_n is defined as the s_n in iteration $(t-1)$ and $s_n = s'_n + \alpha_t g_t(\mathbf{x}_n)$, so

$$\sum_{n=1}^N s_n g_t(\mathbf{x}_n) = \sum_{n=1}^N y_n g_t(\mathbf{x}_n) \quad (34)$$

□

Problem 9

$\text{OR}(x_1, x_2, \dots, x_d)$ means outputs TRUE if one input is TRUE; outputs FALSE if all inputs are FALSE.

Claim: $(w_0, w_1, \dots, w_d) = (d-1, 1, \dots, 1)$ implements OR.

Proof of Claim:

1. If all $x_i = -1$, then we have

$$\text{sign} \left(\sum_{i=0}^d w_i x_i \right) = \text{sign} \left(d-1 + \sum_{i=1}^d (-1) \right) = \text{sign}(-1) = \text{FALSE} \quad (35)$$

2. If some $x_i = +1$ and others are -1 , we have

$$\text{sign} \left(\sum_{i=0}^d w_i x_i \right) = \text{sign} (d-1 + 1 + (-1)(d-1)) = \text{sign}(+1) = \text{TRUE} \quad (36)$$

Hence, $(w_0, w_1, \dots, w_d) = (d-1, 1, \dots, 1)$ implements OR.

□

Problem 10

Claim: $D \geq 5$.

Proof of Claim:

From the conclusion of Problem 21, we have $D = 5$.

□

Problem 11

Claim: Only the gradient components with respect to $w_{01}^{(L)}$ may be non-zero, all other gradient components must be zero.

Proof of Claim:

Consider

$$\frac{\partial e_n}{\partial w_{i1}^{(L)}} = -2 \left(y_n - s_1^{(L)} \right) \left(x_i^{(L-1)} \right) = -2 \left(y_n - \sum_{j=0}^{d^{(L-2)}} w_{j1}^{(L)} x_j^{(L-1)} \right) \left(x_i^{(L-1)} \right) \quad (37)$$

Since all $w_{ij}^{(\ell)} = 0$, we have

$$\frac{\partial e_n}{\partial w_{i1}^{(L)}} = -2 y_n x_i^{(L-1)} \quad (38)$$

If $i \neq 0$, then

$$\frac{\partial e_n}{\partial w_{i1}^{(L)}} = -2y_n x_i^{(L-1)} = -2y_n \tanh\left(s_i^{(L-1)}\right) = -2y_n \tanh\left(\sum_{j=0}^{d^{(L-2)}} w_{ji}^{(L-1)} x_j^{(L-2)}\right) = 0 \quad (39)$$

since all $w_{ij}^{(\ell)} = 0$.

If $i = 0$, then

$$\frac{\partial e_n}{\partial w_{01}^{(L)}} = -2y_n x_0^{(L-1)} = -2y_n \quad (40)$$

If $y_n \neq 0$, then

$$\frac{\partial e_n}{\partial w_{01}^{(L)}} \neq 0 \quad (41)$$

Similarly, we have

$$\frac{\partial e_n}{\partial w_{0j}^{(\ell)}} = \delta_j^{(\ell)} \left(x_0^{(\ell-1)}\right) = \delta_j^{(\ell)} = \sum_k \left(\delta_k^{(\ell+1)}\right) \left(w_{jk}^{(\ell+1)}\right) \left(\frac{\partial \tanh\left(s_j^{(\ell)}\right)}{\partial s_j^{(\ell)}}\right) = 0 \quad (42)$$

since all $w_{ij}^{(\ell)} = 0$.

□

Problem 12

For $\ell = 1$ and all $w_{ij}^{(\ell)}$ initialized as 1, we have

$$\eta x_i^{(0)} \delta_j^{(1)} = \eta x_i^{(0)} \sum_k \left(\delta_k^{(2)}\right) \left(w_{jk}^{(2)}\right) \left(\frac{\partial \tanh\left(s_j^{(1)}\right)}{\partial s_j^{(1)}}\right) \quad (43)$$

$$= \eta x_i^{(0)} \sum_k \left(\delta_k^{(2)}\right) \left(\frac{\partial \tanh\left(s_j^{(1)}\right)}{\partial s_j^{(1)}}\right) \quad (44)$$

and

$$s_j^{(1)} = \sum_{i=0}^{d^{(0)}} w_{ij}^{(1)} x_i^{(0)} = \sum_{i=0}^{d^{(0)}} x_i^{(0)} \quad (45)$$

So we have

$$s_1^{(1)} = s_2^{(1)} = \dots = s_{d^{(1)}}^{(1)} \Rightarrow \delta_1^{(1)} = \delta_2^{(1)} = \dots = \delta_{d^{(1)}}^{(1)} \quad (46)$$

Hence, all update term of $w_{ij}^{(1)}$ is the same for $1 \leq j \leq d^{(1)}$, so

$$w_{ij}^{(1)} = w_{i(j+1)}^{(1)} \quad (47)$$

for $1 \leq j \leq d^{(1)} - 1$.

□

Problem 13

The roles of following tree are

1. (feature column, s , θ). The meaning of combination of numbers.
2. If the feature of \mathbf{x} is smaller than θ , then go to the left tree; if not, then go to the right tree.
3. If go to left and there is no more node, then return $s \times (+1)$; if go to the right and there is no more node, return $s \times (-1)$, where s is from the last node.

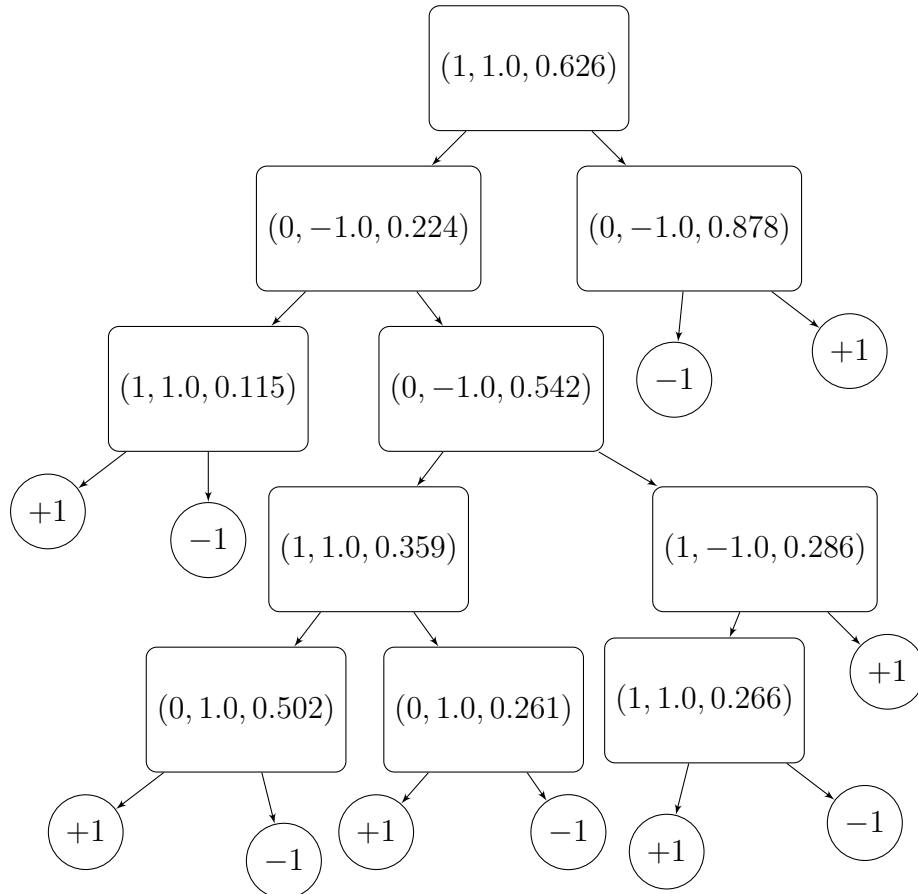


Figure 1: Tree Graph

□

Problem 14

$$E_{\text{in}} = 0.0.$$

□

Problem 15

$$E_{\text{out}} = 0.126.$$

□

Problem 16

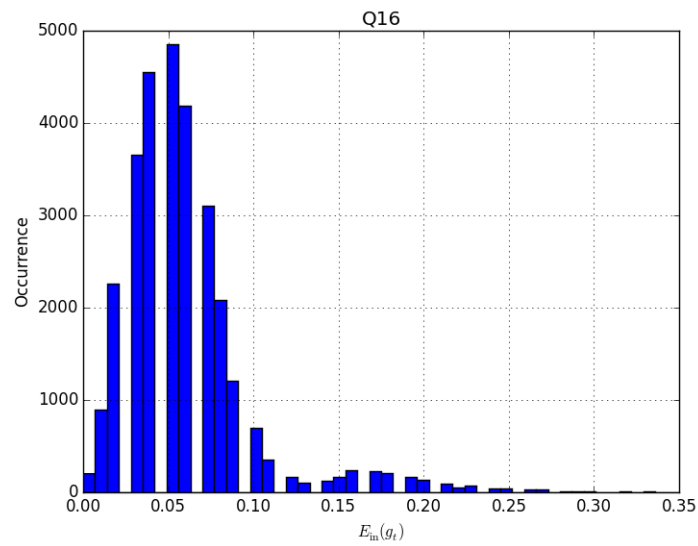


Figure 2: Q16

The average $E_{\text{in}}(g_t)$ of total 30,000 trees is 0.0593.

□

Problem 17

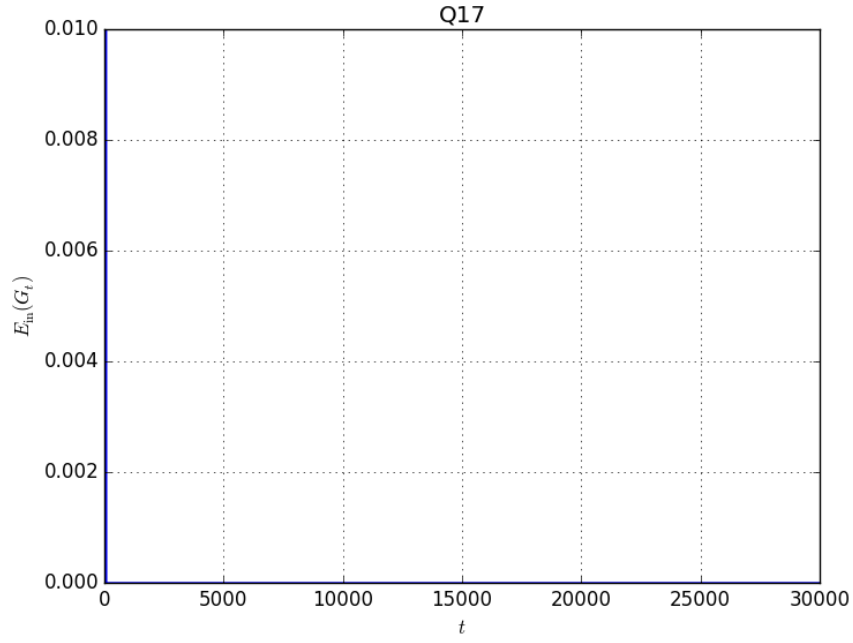


Figure 3: Q17

The average $E_{\text{in}}(G_t)$ of total 30,000 trees is 0. Since most values are 0 so the figure seems nothing left.

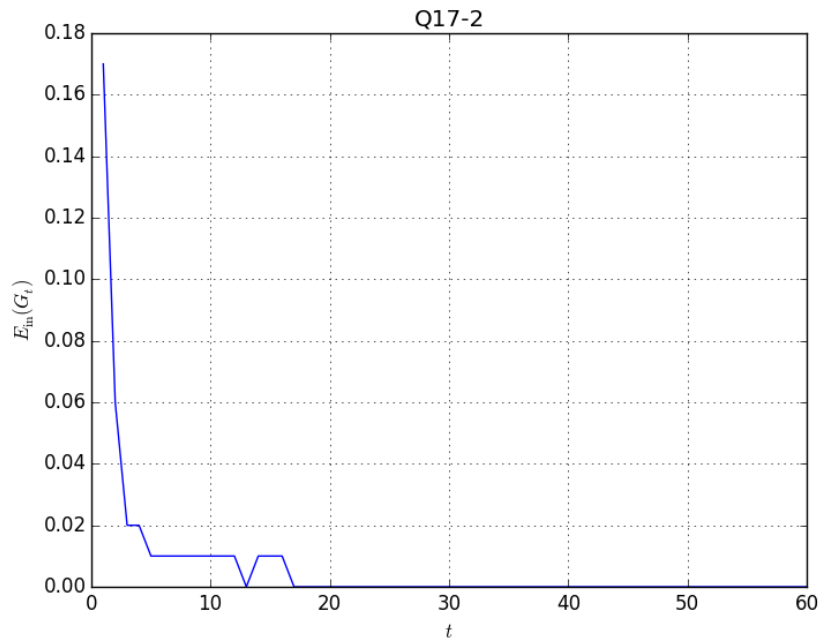


Figure 4: Q17 first 60 points

The figure of first 60 points. We can see that after the 20th point, $E_{\text{in}}(G_t)$ is almost 0.

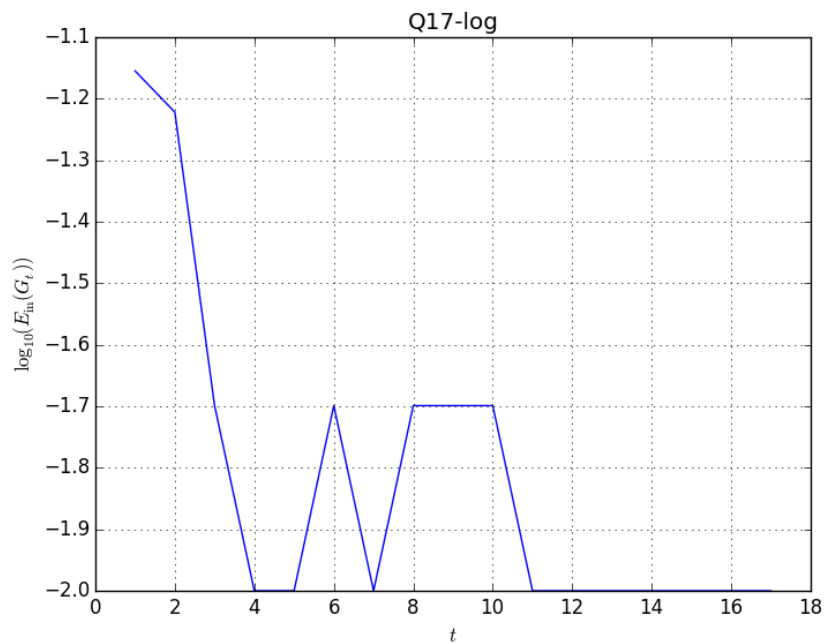


Figure 5: Q17 log₁₀(E_{in}(G_t))

The figure of log value of $E_{\text{in}}(G_t)$ (removed $E_{\text{in}}(G_t) = 0$ points).

We can see that most of the $E_{\text{in}}(G_t) = 0$ (only 17 points left).

□

Problem 18

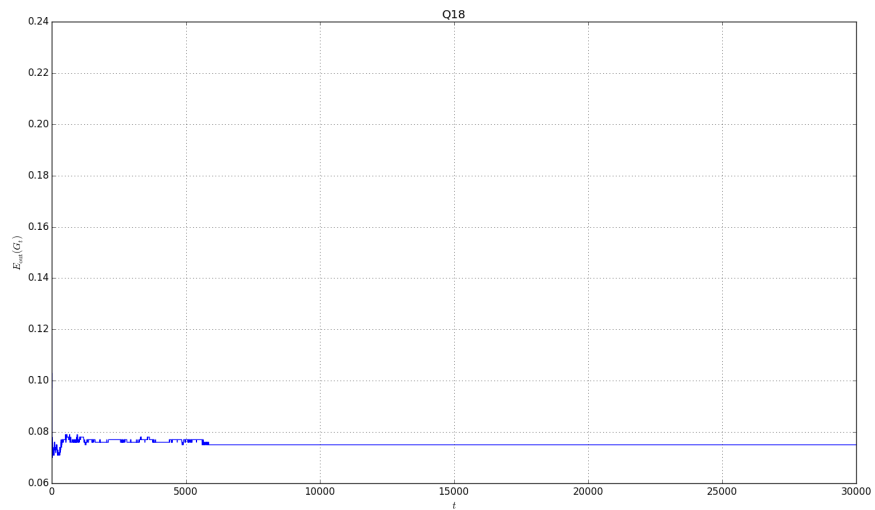


Figure 6: Q18

The average $E_{\text{out}}(G_t)$ of total 30,000 trees is 0.0753.

This curve approaches the average value as $t \rightarrow 30,000$. The curve of Problem 17 oscillates and most points are 0.

□

Problem 19

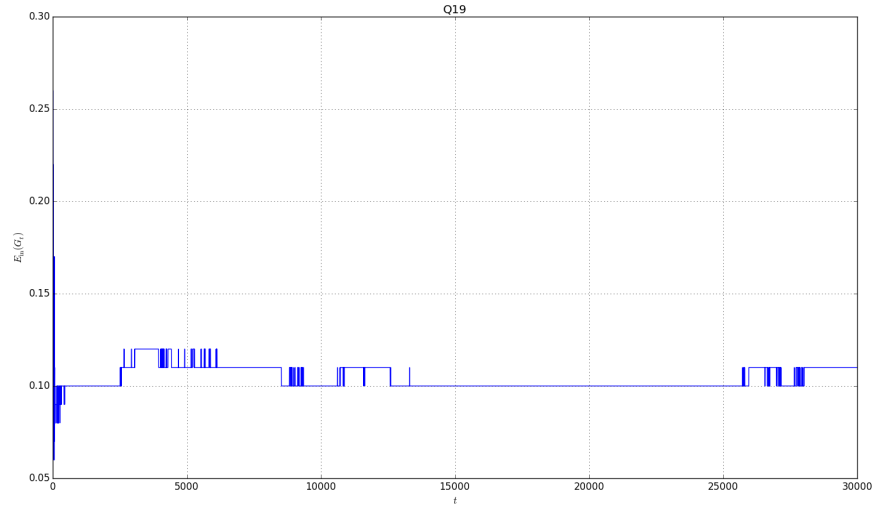


Figure 7: Q19

The average $E_{\text{in}}(G_t)$ of total 30,000 trees is 0.1042.

□

Problem 20

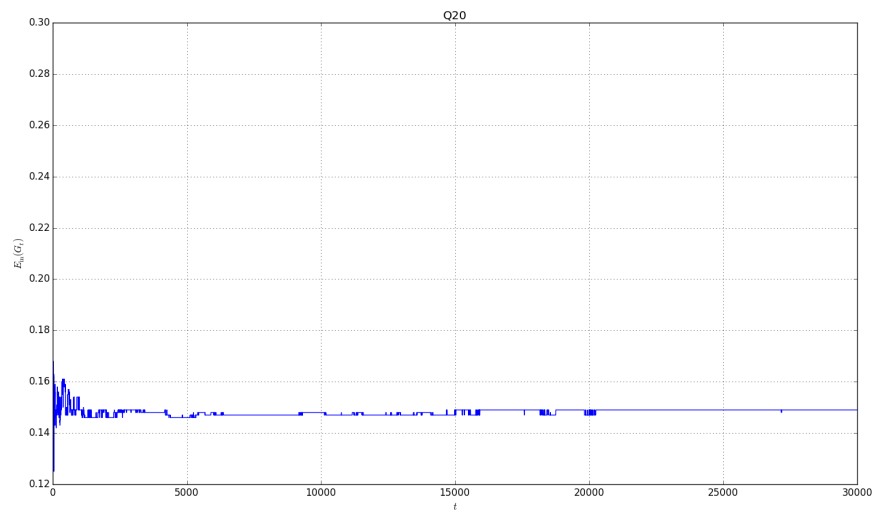


Figure 8: Q20

The average $E_{\text{out}}(G_t)$ of total 30,000 trees is 0.1483.

This curve approaches the average value as $t \rightarrow 30,000$. The curve of Problem 19 oscillates between the average value.

□

Problem 21

Consider the projection of all convex points to diagonal line of N -dimensional hypercube, the value of each function is

1. 2-D: $(-1)(1, 1)(-1)$
2. 3-D: $(-1)(1, 1, 1)(-1, -1, -1)(1)$
3. 4-D: $(-1)(1, 1, 1, 1)(-1, -1, -1, -1, -1, -1)(1, 1, 1, 1)(1)$
- \vdots

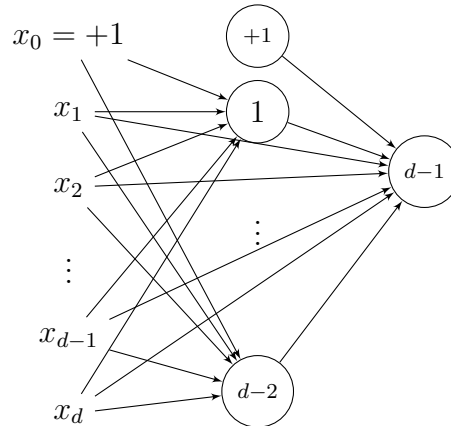
the convex points are the input of neural network. The projections are linear transformation of these convex points, which can be implemented by d linear classifier (AND or OR). By uniform blending all these linear classifier and sign ($\{\text{all linear classifier}\}$), we implement the XOR (x_1, x_2, \dots, x_d) .

The input layer to hidden layer $(d - D)$ is the linear transformation, the hidden layer to output layer $(D - 1)$ is the uniform blending of all these linear classifier.

□

Problem 22

The answer is $D = d - 1$. Like





Reference

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