Machine Learning Techniques Homework #3

tags: `mlt` , `ntu` , `course` , `homework`

Cheng-Shih Wong, R04945028, mob5566@gmail.com (mailto:mob5566@gmail.com)

1.

For the weighted-E_{in} of linear regression

$$\begin{aligned}
\min_{\mathbf{w}} E_{\text{in}}^{\mathbf{u}} &= \min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} u_n \left(y_n - \mathbf{w}^T \mathbf{x}_n\right)^2 \\
&= \min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \left(\pm\sqrt{u_n} y_n - \mathbf{w}^T (\pm\sqrt{u_n}\mathbf{x}_n)\right)^2 \\
&= \min_{\mathbf{w}} \frac{1}{N} \sum_{n=1}^{N} \left(\tilde{y}_n - \mathbf{w}^T \tilde{\mathbf{x}}_n\right)^2 \qquad (1.1)
\end{aligned}$$

Then (1.1) is non-weightd E_{in} of linear regression, i.e. usual linear regression on “pseudo data” $(\tilde{\mathbf{x}}_n, \tilde{y}_n)$

$$\begin{aligned}
\tilde{y}_n &= \pm\sqrt{u_n} y_n \\
\tilde{\mathbf{x}}_n &= \pm\sqrt{u_n} \mathbf{x}_n \\
\text{for } n &= 1, \ldots, N
\end{aligned}$$

2.

Let total number of examples is N

From the page 12 of Lecture 8 slide, we have

$$\text{total } u_n^{t+1} \text{ of incorrect } = \text{total } u_n^{t+1} \text{ of correct}$$

After first iteration, all the positive examples are correct and all the negative examples are incorrect.

$$
\begin{aligned}
\text{total } u_n^{t+1} \text{ of incorrect } &= 0.01N \times u_-^{(2)} \\
\text{total } u_n^{t+1} \text{ of correct } &= 0.99N \times u_+^{(2)}
\end{aligned}
$$

Then

$$
u_-^{(2)} \times 0.01N = u_+^{(2)} \times 0.99N
$$
$$
\frac{u_+^{(2)}}{u_-^{(2)}} = \frac{1}{99}
$$

3

For the decision stumps on i-th dimension, there are $(R-L+1)*2$ different decision stumps, where the threshold of each decision stump is on each integer value $L,\ldots,R$ and has two directions, positive and negative.

For the decision stumps on different dimension, they share the constant hypothesis, **all positive** and **all negative** dichotomies, which are the $\theta = L$ decision stumps on each dimension of input space.

That is, for d-dimensional decision stumps, there are

$$
\begin{aligned}
&2d*(R-L+1)-2(d-1) \\
&= 2d*(R-L)+2
\end{aligned}
$$

different decision stumps.

Under $d=2, L=1, R=6$, there are 22 different decision stumps!

4

Consider integer input vectors

Denote threshold θ_t, dimension index i_t, direction s_t are for t-th decision stump.

$$
\begin{aligned}
K_{ds}(\mathbf{x}, \mathbf{x}') &= (\phi_{ds}(\mathbf{x}))^T \left(\phi_{ds}(\mathbf{x}')\right) \\
&= \sum_{t=1}^{|\mathcal{G}|} g_t(\mathbf{x}) g_t(\mathbf{x}') \\
&= \sum_{t=1}^{|\mathcal{G}|} \left(s_t \cdot \mathsf{sign}(\mathbf{x}_{i_t} - \theta_t)\right) \left(s_t \cdot \mathsf{sign}(\mathbf{x}'_{i_t} - \theta_t)\right) \\
&= \sum_{t=1}^{|\mathcal{G}|} \mathsf{sign}(\mathbf{x}_{i_t} - \theta_t) \mathsf{sign}(\mathbf{x}'_{i_t} - \theta_t)
\end{aligned}
$$

Then

$$
\mathsf{sign}(\mathbf{x}_{i_t} - \theta_t)\mathsf{sign}(\mathbf{x}'_{i_t} - \theta_t) = \begin{cases} +1, & \text{if } \mathbf{x}_{i_t}, \mathbf{x}'_{i_t} \text{ are on the same side relative to } \theta_t \\ -1, & \text{else} \end{cases}
$$

For the decision stumps $\{g_t\}_{t=1}^{|\mathcal{G}_i|}$ on i-th dimension

If $\theta_t > \max(\mathbf{x}_i, \mathbf{x}'_i)$ or $\theta_t \leq \min(\mathbf{x}_i, \mathbf{x}'_i)$, we have 2 decision stumps that $g_t(\mathbf{x})g_t(\mathbf{x}') = 1$ for $s_t = -1, +1$.

If $\min(\mathbf{x}_i, \mathbf{x}'_i) < \theta_t \leq \max(\mathbf{x}_i, \mathbf{x}'_i)$, we have 2 decision stumps that $g_t(\mathbf{x})g_t(\mathbf{x}') = -1$ for $s_t = -1, +1$.

Therefore

$$
\begin{aligned}
\sum_{t=1}^{|\mathcal{G}_i|} g_t(\mathbf{x}) g_t(\mathbf{x}') &= 2 * (R - L - |\mathbf{x}_i - \mathbf{x}'_i|) - 2 * |\mathbf{x}_i - \mathbf{x}'_i| \\
&= 2(R - L) - 4|\mathbf{x}_i - \mathbf{x}'_i|
\end{aligned}
$$

For all d dimensions, we have to add the two sharing constant decision stumps (all positive and all negative), where $g_t(\mathbf{x})g_t(\mathbf{x}')$ always $+1$.

$$
\begin{aligned}
K_{ds}(\mathbf{x}, \mathbf{x}') &= \sum_{t=1}^{|\mathcal{G}|} g_t(\mathbf{x}) g_t(\mathbf{x}') + 2 \\
&= \sum_{i=1}^{d} \left(2(R - L) - 4|\mathbf{x}_i - \mathbf{x}'_i|\right) + 2 \\
&= 2d(R - L) + 2 - 4 \sum_{i=1}^{d} |\mathbf{x}_i - \mathbf{x}'_i|
\end{aligned}
$$

For **Q.3**

$$d = 2, L = 1, R = 6$$

$$K_{ds}(\mathbf{x}, \mathbf{x}') = 22 - 4\sum_{i=1}^{d} |\mathbf{x}_i - \mathbf{x}'_i|$$

5

$$\begin{aligned}
\text{Gini index } &= 1 - \mu_+^2 - \mu_-^2 \\
&= 1 - \mu_+^2 - (1 - \mu_+)^2 \\
&= -2\mu_+^2 + 2\mu_+ \\
&= -2(\mu_+ - \frac{1}{2})^2 + \frac{1}{2}
\end{aligned}$$

The Gini index is a parabola opening downwards, and the maximum value is $\frac{1}{2}$ occurred at $\mu_+ = \frac{1}{2}$.

6

The normalize Gini index is $\frac{-2\mu_+^2 + 2\mu_+}{0.5} = -4\mu_+^2 + 4\mu_+$.

Only $[\mathrm{b}]$ is equivalent to normalized Gini index.

- [b] the squared regression error $\mu_+(1 - (\mu_+ - \mu_-))^2 + \mu_-(-1 - (\mu_+ - \mu_-))^2$

$$\begin{aligned}
&\mu_+(1 - (\mu_+ - \mu_-))^2 + \mu_-(-1 - (\mu_+ - \mu_-))^2 \\
&= \mu_+(1 - (\mu_+ - (1 - \mu_+)))^2 + (1 - \mu_+)(-1 - (\mu_+ - (1 - \mu_+)))^2 \\
&= \mu_+(2 - 2\mu_+)^2 + (1 - \mu_+)(-2\mu_+)^2 \\
&= \mu_+(4 - 8\mu_+ + 4\mu_+^2) + (4\mu_+^2 - 4\mu_+^3) \\
&= -4\mu_+^2 + 4\mu_+ \\
&= -4(\mu_+ - \frac{1}{2})^2 + 1
\end{aligned}$$

  The maximum value is 1 occurred at $\mu_+ = \frac{1}{2}$, then the normalized squared regression error is $-4\mu_+^2 + 4\mu_+$ which is equivalent to the normalized Gini index.

😄

$[a], [c], [d]$ are not equivalent to normailize Gini index shown below

- [a] the classification error $\min(\mu_+, \mu_-)$

  The maximum value is $\frac{1}{2}$, then the normalized classification error is

$$
\begin{aligned}
2\min(\mu_+, \mu_-) &= 2\min(\mu_+, 1-\mu_+) \\
&= \begin{cases} 2\mu_+, & \text{if } \mu_+ \le 0.5 \\ 2-2\mu_+, & \text{if } \mu_+ > 0.5 \end{cases} \\
&= -|2\mu_+ - 1| + 1
\end{aligned}
$$

- [c] the entropy $E = -\mu_+ \ln \mu_+ - \mu_- \ln \mu_-$

  The maximum value is occurred at $\frac{dE}{d\mu_+} = 0$

$$
E = -\mu_+ \ln \mu_+ - (1-\mu_+)\ln(1-\mu_+)
$$

$$
\begin{aligned}
\frac{dE}{d\mu_+} &= -\ln \mu_+ - 1 + \ln(1-\mu_+) + (1-\mu_+)\left(\frac{1}{1-\mu_+}\right) \\
&= \ln\left(\frac{1-\mu_+}{\mu_+}\right) \\
&= 0
\end{aligned}
$$

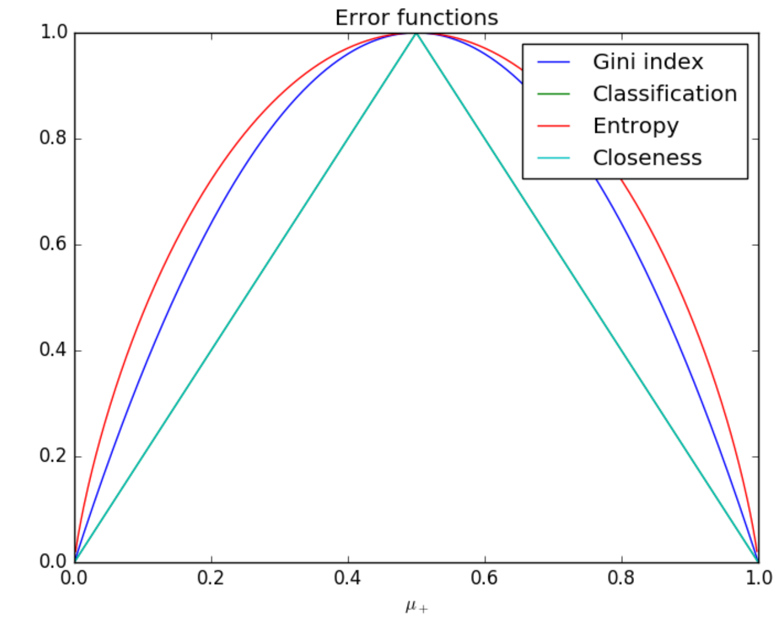
$$
\frac{1-\mu_+}{\mu_+} = 1 \Rightarrow \mu_+ = \frac{1}{2}
$$

  The maximum value is $-\ln\left(\frac{1}{2}\right)$, then the normalized entropy is $-\frac{E}{\ln 0.5}$.

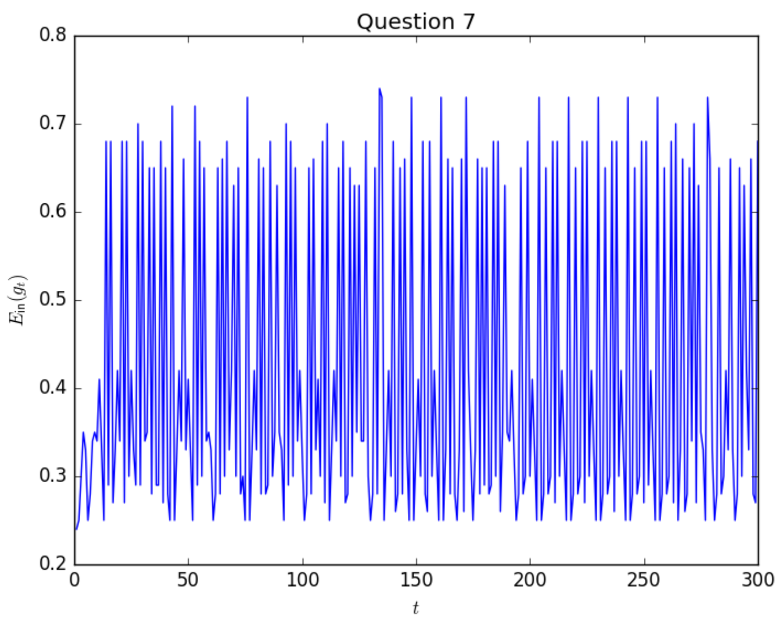
- [d] the closeness $1 - |\mu_+ - \mu_-|$

  The maximum value is 1, then the normalized form is the same to the normalized classification error.

We can plot these errors for $\mu_+ = [0, 1]$, and observe that these errors are all different from the normalized Gini index.



7



$$E_{\text{in}}(g_1) = 0.24$$
$$\alpha_1 = 0.57634$$

😄

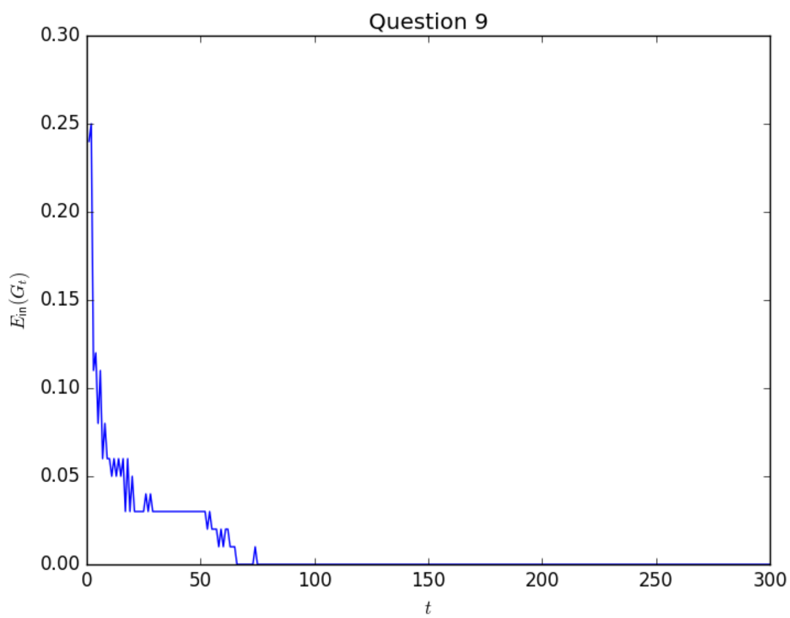
8

The $E_{\text{in}}(g_t)$ is not either decreasing or increasing, but it vibrates severely.

I thought this is because the best $E_{\text{in}}(g_t)$ is occurred at the first decision stump.

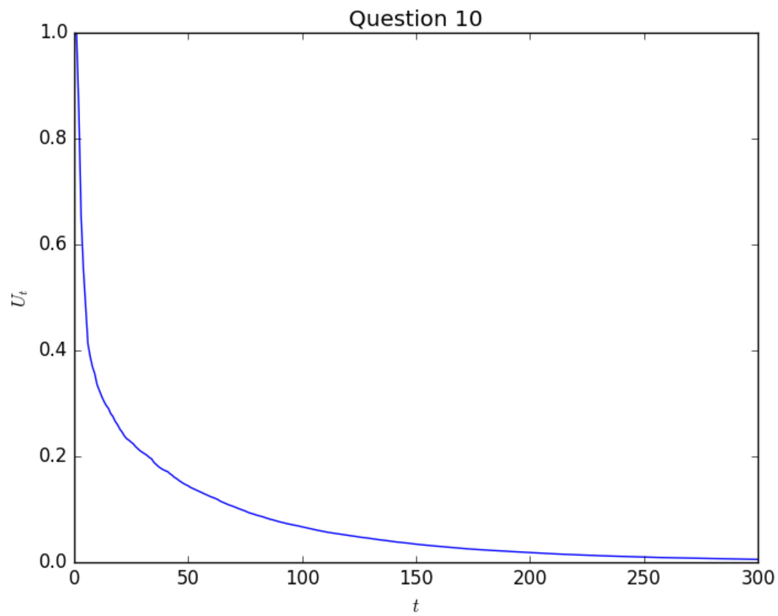
The rest of the decision stumps are trying to adjust the boundary, that they don't have to minimize E_{in}.

9



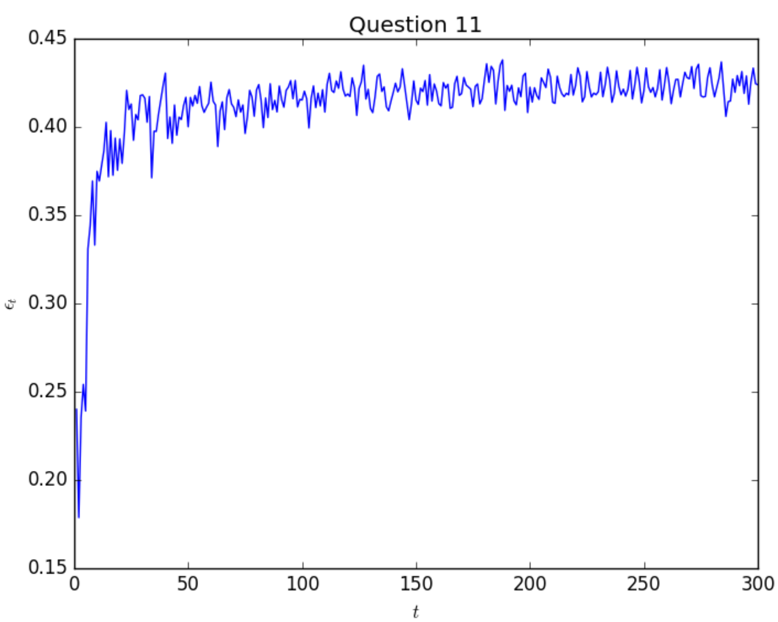
$$E_{\text{in}}(G) = 0$$

10



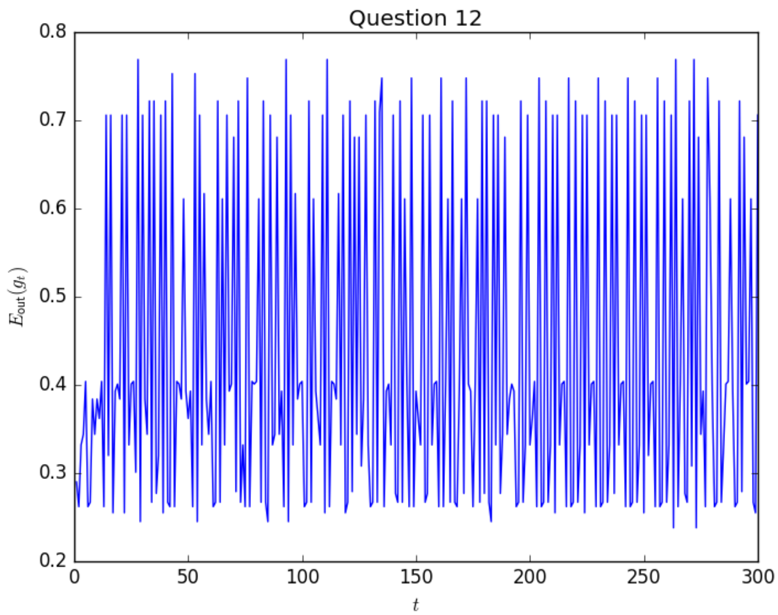
$$U_2 = 0.85417$$
$$U_T = 0.00547$$

11



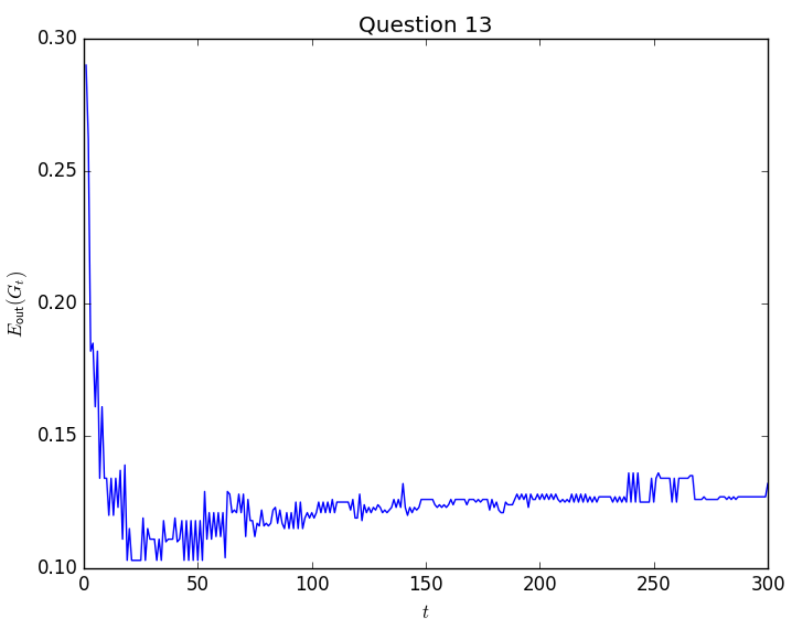
The minimal ϵ_t is 0.17873.

12



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

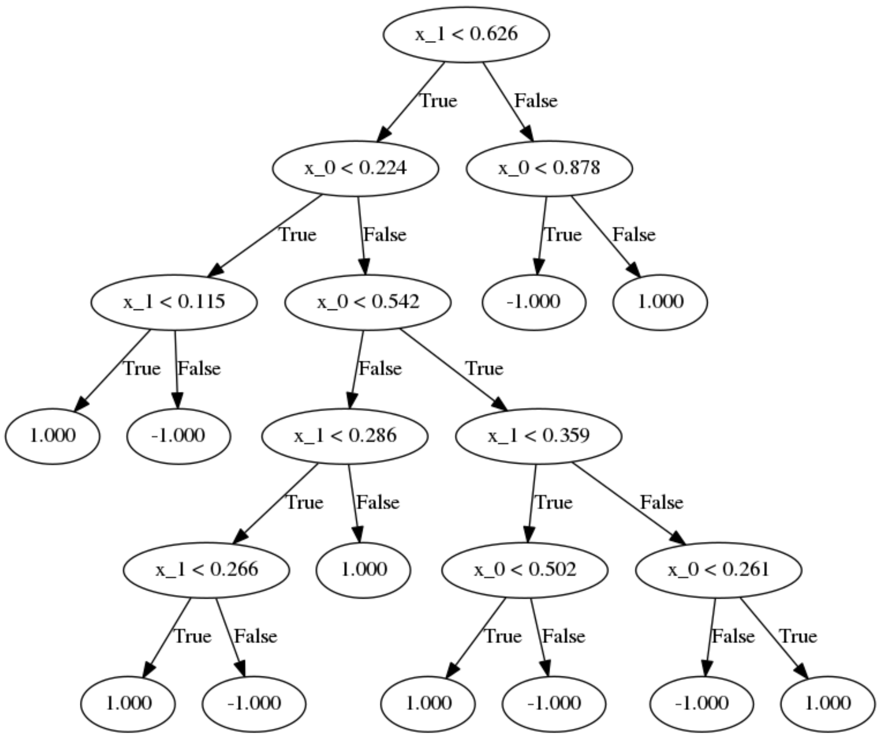
13



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

14

The graph is auto-generated by `python networkx` and `graphviz` !!!



15

$$
\begin{aligned}
E_{\text{in}} &= 0 \\
E_{\text{out}} &= 0.126
\end{aligned}
$$

16

There are 11 leaves that we can prune.

	E_{in}	E_{out}
1	0.01000	0.14400
2	0.14000	0.21500
3	0.14000	0.20300
4	0.01000	0.10900
5	0.01000	0.11700
6	0.06000	0.17300
7	0.20000	0.27900
8	0.09000	0.24200
9	0.01000	0.11600
10	0.30000	0.38300
11	0.03000	0.15300

There are 4 leaves to be pruned to have lowest $E_{\text{in}} = 0.01$, and their E_{out} are $0.144, 0.109, 0.117, 0.116$ respectively.

17

Initially

$$u_n = \frac{1}{N}, \forall n = 1, \dots, N$$

$$U_1 = \sum_{n=1}^{N} u_n = 1$$

From Lecture 11, we have

$$u_n^{(t+1)} = u_n^{(t)} \cdot \exp(-y_n \alpha_t g_t(\mathbf{x}_n))$$

$$
\begin{aligned}
U_{t+1} &= \sum_{n=1}^{N} u_n^{(t+1)} \\
&= \sum_{n=1}^{N} u_n^{(t)} \cdot \exp(-y_n \alpha_t g_t(\mathbf{x}_n)) \\
&= \sum_{n \in \text{correct}} u_n^{(t)} \cdot \exp(-\alpha_t) + \sum_{n \in \text{incorrect}} u_n^{(t)} \cdot \exp(\alpha_t) \\
&= \sum_{n \in \text{correct}} u_n^{(t)} \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \sum_{n \in \text{incorrect}} u_n^{(t)} \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \\
\text{By } \epsilon_t &= \frac{\sum_{n \in \text{incorrect}} u_n^{(t)}}{U_t} \\
&= U_t \cdot (1-\epsilon_t) \cdot \sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + U_t \cdot \epsilon_t \cdot \sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \\
&= U_t \cdot \sqrt{\epsilon_t(1-\epsilon_t)} + U_t \cdot \sqrt{\epsilon_t(1-\epsilon_t)} \\
&= U_t \cdot 2\sqrt{\epsilon_t(1-\epsilon_t)} \le U_t \cdot 2\sqrt{\epsilon(1-\epsilon)}
\end{aligned}
$$

18

And from **Q.17**, we have

$$
\begin{aligned}
E_{\text{in}}(G_T) &\le U_{T+1} \\
&\le U_T \cdot 2\sqrt{\epsilon_T(1-\epsilon_T)} \\
&\le U_T \cdot 2\sqrt{\epsilon(1-\epsilon)} \\
&\le U_1 \cdot \left(2\sqrt{\epsilon(1-\epsilon)}\right)^T \\
&\le \left(2\sqrt{\epsilon(1-\epsilon)}\right)^T \\
&\le \exp\left(-2(\frac{1}{2}-\epsilon)^2\right)^T \\
&\le \exp\left(-2T(\frac{1}{2}-\epsilon)^2\right)
\end{aligned}
$$

From Leture 11

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

We want $E_{\text{in}}(G_t) = 0$, that is

$$y_n \sum_{\tau=1}^{t} \alpha_\tau g_\tau(\mathbf{x}_n) > 0, \forall n = 1, \ldots, N$$

Then

$$U_{T+1} < \frac{1}{N}$$

Therefore

$$\begin{aligned}
E_{\text{in}}(G_T) &\leq U_{T+1} \\
&\leq \exp\left(-2T(\frac{1}{2} - \epsilon)^2\right) \\
&< \frac{1}{N} \\
&\Rightarrow -2T(\frac{1}{2} - \epsilon)^2 = -\ln N \\
&\Rightarrow T = \frac{\ln N}{2(\frac{1}{2} - \epsilon)^2}, \text{ and } \epsilon < \frac{1}{2} \\
T &= O(\log N)
\end{aligned}$$