1.

原式$\mathrm{F(A,B)}$可簡化為：

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

(1) 對 A 偏微：

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

且因$\mathrm{p_n}=\theta(-y_n(Az_n+B))$，故上式可簡化成：

$$\frac{\partial \mathrm{F}}{\partial \mathrm{A}}=\frac{1}{N}\sum_{n=1}^{N}-y_n z_n p_n$$

(2) 對 B 偏微：

$$\frac{\partial \mathrm{F}}{\partial \mathrm{A}}=\frac{1}{N}\sum_{n=1}^{N}\frac{1}{1+e^{-y_n(Az_n+B)}}\left(-y_n e^{-y_n(Az_n+B)}\right)=\frac{1}{N}\sum_{n=1}^{N}-y_n p_n$$

由上可知：

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

2.

首先，$\mathrm{p_n}$可化成以下：$\frac{1}{\mathrm{p_n}}=1+e^{y_n(Az_n+B)}$ 。故其對 A 和 B 的偏微為：

$$\frac{\partial \mathrm{p_n}}{\partial A}=(-p_n^2)\mathrm{y_n}z_n e^{y_n(Az_n+B)}=(-p_n^2)\mathrm{y_n}z_n\left(\frac{1-p_n}{p_n}\right)=-p_n(1-p_n)y_n z_n$$

$$\frac{\partial \mathrm{p_n}}{\partial B}=(-p_n^2)\mathrm{y_n}e^{y_n(Az_n+B)}=-p_n(1-p_n)y_n$$

而 Hessian Matrix 為：

$$\mathrm{H(F)}=\begin{bmatrix}\frac{\partial^2\mathrm{F}}{\partial\mathrm{A}^2} & \frac{\partial^2\mathrm{F}}{\partial\mathrm{A}\,\partial\mathrm{B}}\\ \frac{\partial^2\mathrm{F}}{\partial\mathrm{A}\,\partial\mathrm{B}} & \frac{\partial^2\mathrm{F}}{\partial\mathrm{B}^2}\end{bmatrix}=\begin{bmatrix}\frac{1}{N}\sum_{n=1}^{N}y_n^2 z_n^2 p_n(1-p_n) & \frac{1}{N}\sum_{n=1}^{N}y_n^2 z_n p_n(1-p_n)\\ \frac{1}{N}\sum_{n=1}^{N}y_n^2 z_n p_n(1-p_n) & \frac{1}{N}\sum_{n=1}^{N}y_n^2 p_n(1-p_n)\end{bmatrix}$$

而$\mathrm{y_n^2}=1$，故化簡可得：

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

3.

先列出原本的 dual problem 的 QP problem：

$$\min_{\alpha}\frac{1}{2}\alpha^{\mathrm{T}}\mathrm{Q}\alpha+\mathrm{p}_{\mathrm{QP}}^{\mathrm{T}}\alpha$$

$$\text{subject to } \mathrm{A}\alpha \ge \mathrm{C}_{\mathrm{QP}}, \text{即為：} \sum y_n\alpha_n = 0\,, C \ge \alpha_n \ge 0 \; for\; n = 1, 2, \dots, \mathrm{N}$$

其中，$(\mathrm{Q}, \mathrm{p}, \mathrm{A}, \mathrm{C}_{\mathrm{QP}})$如下：

$$\mathrm{Q}=\begin{bmatrix} \mathrm{y_1y_1K(x_1,x_1)} & \mathrm{y_1y_2K(x_1,x_2)} & \dots & \mathrm{y_1y_NK(x_1,x_N)} \\ \mathrm{y_2y_1K(x_2,x_1)} & \mathrm{y_2y_2K(x_2,x_2)} & & \mathrm{y_2y_NK(x_2,x_N)} \\ \dots & & & \dots \\ \mathrm{y_Ny_1K(x_N,x_1)} & \mathrm{y_Ny_2K(x_N,x_2)} & \dots & \mathrm{y_Ny_NK(x_N,x_N)} \end{bmatrix}$$

$$\mathrm{p}_{\mathrm{QP}}=-\mathrm{I_n}, \mathrm{A}=\begin{bmatrix} y^T \\ -y^T \\ I_N \\ -I_N \end{bmatrix}, \mathrm{C}_{\mathrm{QP}}=\begin{bmatrix} 0 \\ 0 \\ 0_N \\ -CI_N \end{bmatrix}$$

而因為$\mathrm{K}=\exp(-\gamma\|x-x'\|^2)$，在$\gamma\to\infty$時：

(1)若 $\mathrm{x}=\mathrm{x}'$，則 $\mathrm{K}=1$ (2) 若 $\mathrm{x}\neq\mathrm{x}'$，則 $\mathrm{K}\to 0$。故可得$\mathrm{Q}=\mathrm{I_n}$。

原本要解之 min 即為$\min_{\alpha}\frac{1}{2}\alpha^T\alpha-\mathrm{I_n}\alpha=\min_{\alpha}\frac{1}{2}(\alpha^T\mathrm{I_n}-\mathrm{I_n})^{\mathrm{T}}(\alpha^T\mathrm{I_n}-\mathrm{I_n})-\mathrm{I_n}$

當 min 發生時，$\alpha^T\mathrm{I_n}-\mathrm{I_n}=0$，即$\alpha$為全是 1 的 vector。

4.

設兩點為$(\mathrm{x_1}, \mathrm{x_1}-\mathrm{x_1^2}), (\mathrm{x_2}, \mathrm{x_2}-\mathrm{x_2^2})$。

在此兩點之下，所得之 h(x)為：$\mathrm{h}(x)=(1-\mathrm{x_1}-\mathrm{x_2})\mathrm{x}+(\mathrm{x_1x_2})=\mathrm{w_1}x+w_0$

由講義上之定義：$\bar{\mathrm{g}}=\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}g_t$，在 T 夠大時，組合出來的$\bar{\mathrm{g}}$可以表示成：

$$\bar{\mathrm{g}}(\mathrm{x})=\mathrm{E(w_1)x}+\mathrm{E(w_2)}=(1-\mathrm{E(x_1)}-\mathrm{E(x_2)})\mathrm{x}+(\mathrm{E}(x_1)E(x_2))=0.25$$

5.

$$\min_{\mathrm{w}} \mathrm{E}_{\mathrm{in}}^{\mathrm{u}}(w) = \frac{1}{N}\sum_{n=1}^{N} u_n (y_n - w^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} \sqrt{u_n}^2 (y_n - w^T x_n)^2$$

$$= \frac{1}{N}\sum_{n=1}^{N} \left(\sqrt{u_n} y_n - \sqrt{u_n} w^T x_n\right)^2$$

令$\tilde{\mathrm{y}}_n = \sqrt{u_n} y_n,\ \tilde{x}_n = \sqrt{u_n} x_n$，可得：

$$\frac{1}{N}\sum_{n=1}^{N} (\tilde{\mathrm{y}}_n - w^T \tilde{x}_n)^2$$

此即為 usual linear regression。

故題目所求$\{(\tilde{x}_n, \tilde{\mathrm{y}}_n)\}^N{}_{\mathrm{n=1}} = \left\{\left(\sqrt{u_n} x_n, \sqrt{u_n} y_n\right)\right\}^N{}_{\mathrm{n=1}}$

6.

由上課所提可知，optimal re-weighting 為：

$$multi\ correct \propto \epsilon_t; multi\ incorrect \propto (1 - \epsilon_t)$$

又由題目知，$\epsilon_1 = 0.22$，且$\mathrm{u}_+^{(1)} = \mathrm{u}_-^{(1)}$。故可得：

$$\frac{\mathrm{u}_+^{(2)}}{\mathrm{u}_-^{(2)}} = \frac{\mathrm{u}_+^{(1)}}{\mathrm{u}_-^{(1)}} * \frac{0.22}{0.78} = \frac{11}{39}$$

7.

s 有兩種可能(±1)，d 亦有兩個維度，θ 有$\left(\mathrm{M} - (-\mathrm{M})\right)$種可能。

(註：因題目有定義$\mathrm{sign}(0) = +1$，故在$\theta = -\mathrm{M}$時，為下列所提之情形)

除此之外，仍需考慮全為$+1$ or -1之情形。

故總共有：$\mathrm{d} * \mathrm{s} * (2\mathrm{M}) + 2 = 2 * 2 * 10 + 2 = 42$種不同的 decision stump。

8.

因s^2必為+1，故$\mathrm{K}_{\mathrm{ds}}(x, x')$可簡化為：

$$\mathrm{K}_{\mathrm{ds}}(\mathrm{x}, \mathrm{x}') = \sum_{i=1}^{|G|} \mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}} - \theta_\mathrm{i}\right)\mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}}' - \theta_\mathrm{i}\right)$$

設$\min\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}},\ \mathrm{x}_{\mathrm{i}_\mathrm{d}}'\right) = \mathrm{a}_\mathrm{d}, \max\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}},\ \mathrm{x}_{\mathrm{i}_\mathrm{d}}'\right) = \mathrm{b}_\mathrm{d}$，易知在$\theta = [-\mathrm{M}, \mathrm{a}_\mathrm{d}] \cup (\mathrm{b}_\mathrm{d}, \mathrm{M}]$中，$\mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}} - \theta_\mathrm{i}\right)\mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}}' - \theta_\mathrm{i}\right) = +1$；而在$(\mathrm{a}_\mathrm{d}, \mathrm{b}_\mathrm{d}]$中，$\mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}} - \theta_\mathrm{i}\right)\mathrm{sign}\left(\mathrm{x}_{\mathrm{i}_\mathrm{d}}' - \theta_\mathrm{i}\right) = -1$。

可知，原式中有$2 * \sum_{d \epsilon D}(\mathrm{b}_\mathrm{d} - \mathrm{a}_\mathrm{d})$個值為$-1$，其他均為$+1$，故原式可簡化為：

$$-2\sum_{d\epsilon D}(\mathrm{b_d}-\mathrm{a_d})+\left((2*\mathrm{d}*(2\mathrm{M})+2)-2*\sum_{d\epsilon D}(\mathrm{b_d}-\mathrm{a_d})\right)$$
$$=(4\mathrm{dM}+2)-4\sum_{d\epsilon D}(\mathrm{b_d}-\mathrm{a_d})$$
$$=(4\mathrm{dM}+2)-4\sum_{d\epsilon D}|\mathrm{x}_{\mathrm{i_d}}-\mathrm{x}_{\mathrm{i_d}}{}'|$$
其中，$\mathrm{D}=\{1,2,\dots,\mathrm{d}\}$(此 d 為題目給定之 d)

9.

此題和下題的結果如下：

lambda	Train Error	Test Error
0.05	0.3175	0.36
0.5	0.3175	0.36
5	0.32	0.36
50	0.315	0.4
500	0.33	0.37

$\mathrm{E_{in}}(g)$最小為：0.315，此時之 λ 為 50。

10.

$\mathrm{E_{out}}(g)$最小為：0.36，此時之 λ 為 0.05 or 0.5 or 5。

11.

lambda	Train Error	Test Error
0.05	0.3175	0.37
0.5	0.3225	0.37
5	0.32	0.36
50	0.315	0.39
500	0.32	0.38

$\mathrm{E_{in}}(g)$最小為：0.315，此時之 λ 為 50。

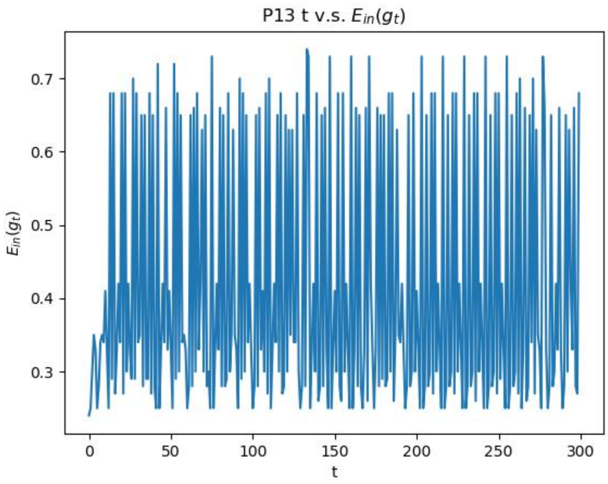
其實在 λ 為這五個值時，使用 bagging 與否對$\mathrm{E_{in}}$之影響並不大。我認為是因為本來線性的時候就分類的滿不錯的了，使用 bagging 對於降低$\mathrm{E_{in}}$效果並不大。

12.

$\mathrm{E_{out}}(g)$最小為：0.36，此時之 λ 為 5

和上題相似，使用 bagging 與否對$\mathrm{E_{out}}$之影響並不大。我認為原因和上題相同。

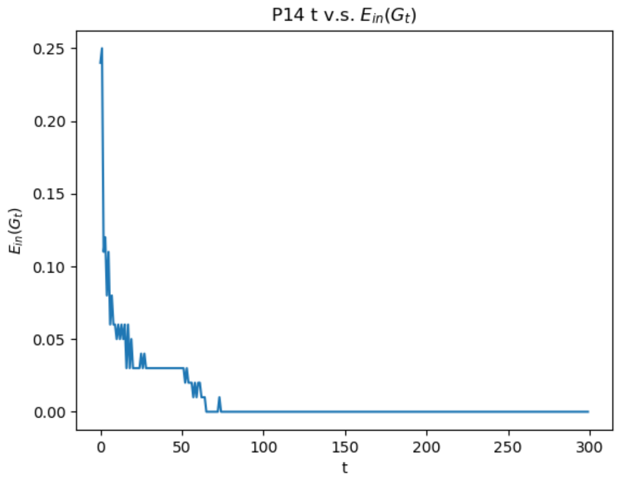
13.



$$\mathrm{E_{in}}(g_{300}) = 0.68$$

可以看出$\mathrm{E_{in}}(g_t)$呈現不規則的趨勢。因題目要我們求的是沒有乘上 u 的 error，故本來隨著不同比例的 u 來找最佳切割線，其$\mathrm{E_{in}}(g_t)$的值會浮動不定。

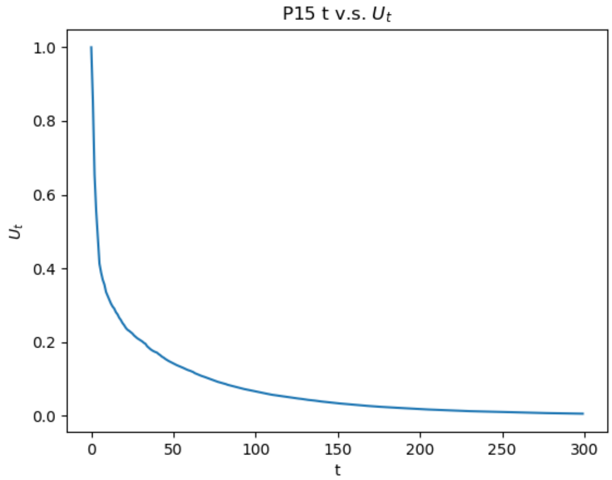
14.



$$\mathrm{E_{in}}(G_{300}) = 0$$

可以看出$\mathrm{E_{in}}(G_t)$大致呈現遞減的趨勢，而在 t 約莫為 80 之後，$\mathrm{E_{in}}(G_t)$就為 0 了。這其實和 Bonus 所說的情形相同，更新夠多次後$\mathrm{E_{in}}(G_t)$本來就會是 0。

15.



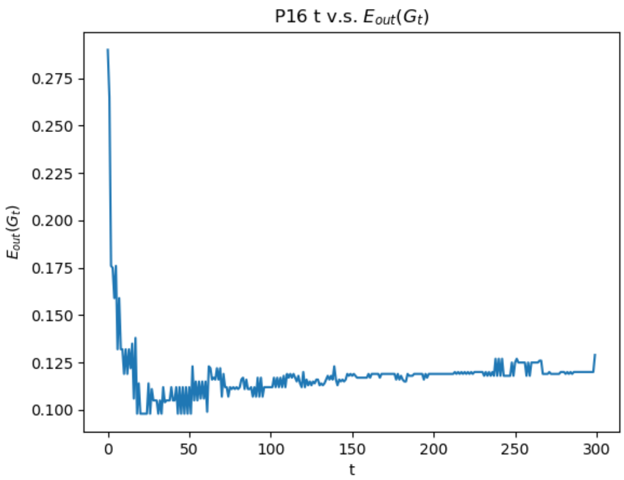
$$U_{300} = 5.47 * 10^{-3}$$

由第 17 題可知：

$$U_{t+1} = 2U_t\sqrt{(1-\epsilon_t)\epsilon_t} > U_t(\because \epsilon_t < \frac{1}{2})$$

故U_t為遞減是件十分合理的事情。

16.



$$E_{out}(G_{300}) = 0.129$$

可以看出其走勢為快速下降後，再緩慢遞增。我認為是因為在 t 較小時，正在從 underfitted 走向 fitted。但在$E_{in}(G_t)$逼近 0 甚至是等於 0 時，仍然在繼續做 Adaboost，便使其有點 overfitted 的現象，但也不會 Overfit 得太誇張。

17.

因為一開始即定$\mathrm{u}^{(1)} = \left[\frac{1}{N}, \frac{1}{N}, \dots, \frac{1}{N}\right]$，故 $\mathrm{U}_1 = \mathrm{N} * \frac{1}{N} = 1$。

設 $\mathrm{u}^{(\mathrm{t})}$ 中，判斷正確的所有 u 總和為$\mathrm{U}_\mathrm{t}^\mathrm{c}$，錯誤的為$\mathrm{U}_\mathrm{t}^\mathrm{w}$。由 u 的 update 知，原本正確/錯誤的值 update 後總和分別會是：

$$\mathrm{U}_\mathrm{t}^{\mathrm{c}\prime} = \frac{\mathrm{U}_\mathrm{t}^\mathrm{c}}{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}},\ \ \mathrm{U}_\mathrm{t}^{\mathrm{w}\prime} = \mathrm{U}_\mathrm{t}^\mathrm{w}\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

故可知：

$$\mathrm{U}_{\mathrm{t}+1} = \mathrm{U}_\mathrm{t}^{\mathrm{c}\prime} + \mathrm{U}_\mathrm{t}^{\mathrm{w}\prime} = \frac{\mathrm{U}_\mathrm{t}^\mathrm{c}}{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} + \mathrm{U}_\mathrm{t}^\mathrm{w}\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}$$

又由因$\mathrm{U}_\mathrm{t}^\mathrm{c}, \mathrm{U}_\mathrm{t}^\mathrm{w}$ 的定義可知：$\mathrm{U}_\mathrm{t}^\mathrm{c} = \mathrm{U}_\mathrm{t} * (1-\epsilon_t)$、 $\mathrm{U}_\mathrm{t}^\mathrm{w} = \mathrm{U}_\mathrm{t} * \epsilon_t$。(因$\epsilon$本來就式答錯的比例) 故上式可化成：

$$\mathrm{U}_{\mathrm{t}+1} = \frac{\mathrm{U}_\mathrm{t} * (1-\epsilon_t)}{\sqrt{\frac{1-\epsilon_t}{\epsilon_t}}} + \mathrm{U}_\mathrm{t} * \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} = 2\mathrm{U}_\mathrm{t}\sqrt{(1-\epsilon_t)\epsilon_t}$$

又因$\epsilon_\mathrm{t} \le \epsilon < \frac{1}{2}$，再加上$(1-\epsilon_t)\epsilon_t$為二次函數。易知$(1-\epsilon_t)\epsilon_t \le\ (1-\epsilon)\epsilon$。

故綜合以上可得：

$$\mathrm{U}_1 = 1, \mathrm{U}_{\mathrm{t}+1} = 2\mathrm{U}_\mathrm{t}\sqrt{(1-\epsilon_t)\epsilon_t} \le\ 2\mathrm{U}_\mathrm{t}\sqrt{(1-\epsilon)\epsilon}$$

18.

由上題得出之式子和本題給的不等式可得：

$$\mathrm{U}_{\mathrm{t}+1} \le\ 2\mathrm{U}_\mathrm{t}\sqrt{(1-\epsilon)\epsilon} \le \mathrm{U}_\mathrm{t}e^{-2\left(\frac{1}{2}-\epsilon\right)^2}$$

(1) 經由遞迴和初始條件$\mathrm{U}_1 = 1$：

$$\mathrm{U}_{\mathrm{t}+1} \le \mathrm{U}_\mathrm{t}e^{-2\left(\frac{1}{2}-\epsilon\right)^2} \le\ \mathrm{U}_{\mathrm{t}-1}e^{2*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)} \le \cdots \le e^{t*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)}$$

(2) $\mathrm{E}_{\mathrm{in}}(G_T) = \frac{1}{\mathrm{N}}\sum_{i=1}^{N}[\![G_T(x_i) \neq y_i\ ]\!] = 0$可用以下式子表達：

$$\mathrm{E}_{\mathrm{in}}(G_T) < \frac{1}{N}$$

(3) 由題目所給之式子以及$\mathrm{E}_{\mathrm{in}}(G_T)$之定義：

$$\mathrm{U}_{\mathrm{t}+1} = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\exp\left(-y_n\sum_{\tau=1}^{t}\alpha_\tau g_\tau(x_n)\right)$$

$$\mathrm{E}_{\mathrm{in}}(G_T) = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}[\![G_T(x_n)y_n \neq 1\ ]\!] = \frac{1}{\mathrm{N}}\sum_{n=1}^{N}\left[\!\!\left[sign\left(y_n\sum_{\tau=1}^{t}\alpha_\tau g_\tau(x_n)\right) \neq 1\ \right]\!\!\right]$$

易知$\mathrm{E_{in}}(G_T) \leq \mathrm{U_{t+1}}$。

故結合以上 3 點可得：

$$\mathrm{E_{in}}(G_T) \leq \mathrm{U_{t+1}} \leq e^{t*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)}$$

且若能使$e^{t*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)} < \frac{1}{\mathrm{N}}$，則$\mathrm{E_{in}}(G_T) < \frac{1}{N}$必成立，即表$\mathrm{E_{in}}(G_T) = 0$。

(註：這個上界並沒有很緊，要使$\mathrm{E_{in}}(G_T) < \frac{1}{N}$，$e^{t*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)}$不一定要比$\frac{1}{\mathrm{N}}$小)

在$e^{t*\left(-2\left(\frac{1}{2}-\epsilon\right)^2\right)} < \frac{1}{\mathrm{N}}$時：

$$\ln N < 2\mathrm{t}\left(\frac{1}{2}-\epsilon\right)^2$$

$$\mathrm{t} > \frac{\ln N}{2\left(\frac{1}{2}-\epsilon\right)^2}$$

可以得當$\mathrm{T} = \mathrm{t} = \mathrm{O}(\ln N)$時，$\mathrm{E_{in}}(G_T) = 0$。