1.

$$F(A.B) = \frac{1}{N}\sum_{n=1}^{N} \ln(1+\exp(-y_n(Az_n+B)))$$ (目的為湊出 p_n)

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

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

$$= \frac{1}{N}\sum_{n=1}^{N} \ln(1/(1-p_n))$$

$$p_n = \theta(-y_n(Az_n+B)) = \theta\left(-\begin{bmatrix} y_n z_n & y_n \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix}\right)$$

$$\nabla F(A.B) = \nabla\left(\frac{1}{N}\sum_{n=1}^{N} \ln(1/(1-p_n))\right)$$

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

$$= -\frac{1}{N}\sum_{n=1}^{N} \nabla \ln(1-p_n)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} \frac{-1}{1-p_n}\nabla p_n$$

$$\nabla p_n = \nabla\left(\frac{\exp(-y_n(Az_n+B))}{1+\exp(-y_n(Az_n+B))}\right)$$ let $-y_n(Az_n+B) = X$

$$= \frac{(1+\exp(X))\cdot\exp(X) - \exp(X)\exp(X)}{(1+\exp(X))^2}\nabla X$$

$$= \frac{\exp(X)}{1+\exp(X)}\cdot\frac{1+\cancel{\exp(X)}-\cancel{\exp(X)}}{1+\exp(X)}\nabla X = p\cdot(1-p)\nabla\left(-\begin{bmatrix} y_n z_n \\ y_n \end{bmatrix}^T\begin{bmatrix} A \\ B \end{bmatrix}\right)$$

$$= -p(1-p)\begin{bmatrix} y_n z_n \\ y_n \end{bmatrix}$$

$$\Rightarrow \nabla F(A.B) = -\frac{1}{N}\sum_{n=1}^{N}\frac{-1}{1-p_n}\nabla p_n = -\frac{1}{N}\sum_{n=1}^{N}\frac{-1}{\cancel{1-p_n}}\cdot -p_n\cancel{(1-p_n)}\begin{bmatrix} y_n z_n \\ y_n \end{bmatrix}$$

$$= \frac{-1}{N}\sum_{n=1}^{N} p_n\begin{bmatrix} y_n z_n \\ y_n \end{bmatrix}$$ #

2.

$$H(F) = \begin{bmatrix} \frac{\partial^2 F(A,B)}{\partial A^2} & \frac{\partial^2 F(A,B)}{\partial A \partial B} \\ \frac{\partial^2 F(A,B)}{\partial A \partial B} & \frac{\partial^2 F(A,B)}{\partial B^2} \end{bmatrix}$$

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

$$\frac{\partial F(A,B)}{\partial B} = -\frac{1}{N}\sum_{n=1}^{N} y_n p_n$$

$$\Rightarrow \frac{\partial^2 F(A,B)}{\partial A^2} = \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n z_n p_n\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n z_n \frac{\partial p_n}{\partial A} = +\frac{1}{N}\sum_{n=1}^{N} y_n z_n p_n (1-p_n)(+y_n z_n)$$

$$= \frac{1}{N}\sum_{n=1}^{N} \underbrace{y_n^2}_{=1} z_n^2 p_n (1-p_n) = \frac{1}{N}\sum_{n=1}^{N} z_n^2 p_n (1-p_n)$$

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

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n z_n \frac{\partial p_n}{\partial B} = +\frac{1}{N}\sum_{n=1}^{N} y_n z_n \cdot p_n (1-p_n)(+y_n)$$

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

$$\frac{\partial^2 F(A,B)}{\partial B^2} = \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} y_n p_n\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n \cdot \frac{\partial p_n}{\partial B} = +\frac{1}{N}\sum_{n=1}^{N} y_n p_n (1-p_n)(+y_n)$$

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

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

3.

for soft-Margin SVM

$b = y_s - \sum_{n=1}^{N} \alpha_n y_n K(x_n, x_s)$

$= y_s - \sum_{n=1}^{N} \alpha_n y_n \exp(-\gamma \|x_n - x_s\|^2)$

$\because \gamma \to \infty$，且所有 x_n 都不同 $\Rightarrow \|x_n - x_s\|^2 \neq 0$ for $n \neq s$

$\therefore \alpha_n y_n \exp(-\gamma \|x_n - x_s\|^2) = \alpha_n y_n \exp(-\infty) = 0$ for $n \neq s$

$b = y_s - 0 - \alpha_s y_s \exp(-\gamma \underbrace{\|x_n - x_n\|^2}_{0}) = y_s - \alpha_s y_s$

設 $y_s = 1$ 的 $\alpha_s = \alpha^+$; $y_s = -1$ 的 $\alpha_s = \alpha^-$

$1 - \alpha^+ = -1 - \alpha^-(-1) \Rightarrow \alpha^+ + \alpha^- = 2$

對於所有 x_n， $y_n(w^T z_n + b) = y_n \cdot (\sum_{i=1}^{N} \alpha_i y_i K(x_i, x_n) + b)$

$= y_n \cdot (\alpha_n y_n \cdot 1 + 0 + b)$

$= y_n(\alpha_n y_n + y_n - \alpha_n y_n) = y_n^2 = 1$

$\Rightarrow$ 所有 x_n 都是 support vector, $\alpha_n \neq 0$ for $n \in N$

$\because \sum_{n=1}^{N} y_n \alpha_n = 0$ ， $y_n = 1$ 跟 $y_n = -1$ 數目一樣

$\Rightarrow \frac{N}{2} \cdot 1 \cdot \alpha^+ + \frac{N}{2} \cdot (-1) \alpha^- = 0$

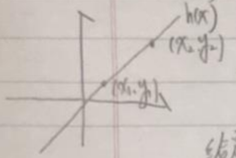
$\Rightarrow \frac{N}{2} \alpha^+ = \frac{N}{2} \alpha^- \Rightarrow \alpha^+ = \alpha^-$

$\begin{cases} \alpha^+ + \alpha^- = 2 \\ \alpha^+ = \alpha^- \end{cases} \Rightarrow \alpha^+ = 1, \alpha^- = 1$

$\alpha = [\alpha_0, \alpha_1, \cdots \alpha_N] = [1 \cdots 1]$ all-1 vector #

4.

給定平面上2點，求 minimized MSE line：



$\Rightarrow$ 即為2點連線.

給定2點 x_1, x_2，$y_1 = x_1 - x_1^2$　$y_2 = x_2 - x_2^2$

設 $h(x) = w_0 x + w_1$

$$w_0 = \frac{y_1 - y_2}{x_1 - x_2} = \frac{(x_1 - x_1^2) - (x_2 - x_2^2)}{x_1 - x_2} = \frac{x_1 - x_2}{x_1 - x_2} - \frac{x_1^2 - x_2^2}{x_1 - x_2} = 1 - (x_1 + x_2)$$

$$y_1 = w_0 x_1 + w_1 \Rightarrow \cancel{x_1 - x_1^2} = (1 - (x_1 + x_2))x_1 + w_1$$

$$= \cancel{x_1 - x_1^2} - x_1 x_2 + w_1 \Rightarrow w_1 = x_1 x_2$$

$h(x) = (1 - x_1 - x_2)x + x x_2$

$\because x_1, x_2$ 的 distribution is uniform

$\Rightarrow E(w_0) = 1 - E(x_1) - E(x_2) = 1 - \frac{1}{2} - \frac{1}{2} = 0$

$E(w_1) = E(x_1 x_2) = E(x_1)E(x_2) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$

$\bar{g}(x) = \frac{1}{4}$ #

5.

$$\min_w E_{in}^u(w) = \frac{1}{N}\sum_{n=1}^{N} u_n (y_n - w^T x_n)^2 = \frac{1}{N}\sum_{n=1}^{N} (\underbrace{\sqrt{u_n} y_n}_{\tilde{y}_n} - w^T \underbrace{\sqrt{u_n} x_n}_{\tilde{x}_n})^2$$

取 $(\tilde{x}_n, \tilde{y}_n) = (\sqrt{u_n} x_n, \sqrt{u_n} y_n)$，對這些資料作 linear regression，其結果和上面的 optimization problem 相同.

6.

$\varepsilon_1 = 1 - 78\% = 22\%$　(-1 的比率即為 incorrect 比率)

$u_+^{(2)} = 1 / \sqrt{\frac{1-\varepsilon_1}{\varepsilon_1}}$

$u_-^{(2)} = 1 \cdot \sqrt{\frac{1-\varepsilon_1}{\varepsilon_1}}$

$$\frac{u_+^{(2)}}{u_-^{(2)}} = \frac{1}{\frac{1-\varepsilon_1}{\varepsilon_1}} = \frac{\varepsilon_1}{1-\varepsilon_1} = \frac{22\%}{78\%} = \frac{22}{78}$$ #

7.

当 $\theta \le -M$，$sign(x_i-\theta)$ 必为 1
当 $\theta > M$，$sign(x_i-\theta)$ 必为 -1
$\Rightarrow$ 不论 $s=\pm1$，只有 $g(x)=1, g(x)=-1$ 两种可能.

对於 $\theta \in (t, t+1]$ for $-M \le t < M-1$ $t \in \mathbb{N}$
因为 $x_i \in \mathbb{N}$，$sign(x_i-\theta)$ 值相同，因此共有 $M-1+(M-1)=2M$ 种 sign 函数.
而每个 i 维度所形成的 g 也会不同，因此要乘上 d.
$s=1,-1$，所以共有 $2\cdot d\cdot 2M+2 = 4dM+2$ 种 g.
$d=2, M=5 \Rightarrow 4\cdot 2\cdot 5+2=42$ 种 #

8.

$$K_{ds}(x,x') = (\phi_{ds}(x))^T(\phi_{ds}(x')) = \sum_{t=1}^{|G|} g_t(x)g_t(x')$$
$$= \sum_{t=1}^{|G|} (s_t\cdot sign(x_{i_t}-\theta_t))(s_t\, sign(x'_{i_t}-\theta_t))$$
$$= \sum_{t=1}^{|G|} sign(x_{i_t}-\theta_t)\cdot sign(x'_{i_t}-\theta_t)$$

if $\theta_t \in (\min(x_{i_t}, x'_{i_t}), \max(x_{i_t}, x'_{i_t})]$，则 $sign(x_{i_t}-\theta_t)\cdot sign(x'_{i_t}-\theta_t) = -1$
if $\theta_t \notin (\min(x_{i_t}, x'_{i_t}), \max(x_{i_t}, x'_{i_t})]$，则 $sign(x_{i_t}-\theta_t)\cdot sign(x'_{i_t}-\theta_t) = 1$
$K_{ds}(x,x')$ 中 $sign(x_{i_t}-\theta_t)\, sign(x'_{i_t}-\theta_t) = -1$ 的数量为:
$2\cdot\sum_{i=1}^{d}|x_i-x'_i|$ （$(\min(x_{i_t}, x'_{i_t}), \max(x_{i_t}, x'_{i_t})]$ 中整数数量 $\times$ $s=\pm1$ 两种情形）
已知 $|G| = 4dM+2$，其中 $2\sum_{i=1}^{d}|x_i-x'_i|$ 种状况为 -1，其馀为 +1

$$K_{ds}(x,x') = 1\times|G| - 2\times\left(2\sum_{i=1}^{d}|x_i-x'_i|\right)$$
$$= 4dM - 4\sum_{i=1}^{d}|x_i-x'_i| + 2 \quad \#$$

17.

① Prove $U_1 = 1$: $U_1 = \sum_{n=1}^{N} \frac{1}{N} = 1$ #

② 已知 $U_{t+1} = \sum_{n=1}^{N} u_n^{t+1} = \frac{1}{N} \sum_{n=1}^{N} \exp\left(-y_n \sum_{\tau=1}^{t} \alpha_\tau g_\tau(x_n)\right)$

$$u_n^{t+1} = \begin{cases} u_n^t \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} & \text{for } y_n g_t(x_n) = -1 \\ u_n^t / \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}, & \text{for } y_n g_t(x_n) = 1 \end{cases}, \quad \varepsilon_t = \frac{\sum_{y_n \neq g_t(x_n)} u_n^t}{\sum_{n=1}^{N} u_n^t}$$

$$U_{t+1} = \sum_{n=1}^{N} u_n^{t+1} = \sum_{y_n \neq g_t(x_n)} u_n^t \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} + \sum_{y_n = g_t(x_n)} u_n^t \Big/ \sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}}$$

$$= \sum_{n=1}^{N} u_n^t \cdot \left(\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \frac{\sum_{y_n \neq g_t(x_n)} u_n^t}{\sum_{n=1}^{N} u_n^t} + \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \frac{\sum_{y_n = g_t(x_n)} u_n^t}{\sum_{n=1}^{N} u_n^t} \right)$$

$$= \sum_{n=1}^{N} u_n^t \left(\sqrt{\frac{1-\varepsilon_t}{\varepsilon_t}} \cdot \varepsilon_t + \sqrt{\frac{\varepsilon_t}{1-\varepsilon_t}} \cdot (1-\varepsilon_t) \right)$$

$$= U_t \left(\sqrt{(1-\varepsilon_t)\varepsilon_t} + \sqrt{\varepsilon_t (1-\varepsilon_t)} \right)$$

$$= 2 U_t \sqrt{\varepsilon_t (1-\varepsilon_t)}$$

we know that $\varepsilon_t \leq \varepsilon < \frac{1}{2} \Rightarrow U_{t+1} = U_t \cdot 2\sqrt{\varepsilon_t(1-\varepsilon_t)} \leq U_t \cdot 2\sqrt{\varepsilon(1-\varepsilon)}$ #

18.

$$U_{t+1} \leq U_t \cdot 2\sqrt{\varepsilon(1-\varepsilon)} \leq U_t \cdot \cancel{\tfrac{1}{2}}\, 2 \cdot \exp\left(-2\left(\tfrac{1}{2}-\varepsilon\right)^2\right)$$

$$\leq U_{t-1} \cdot \exp\left(-2\left(\tfrac{1}{2}-\varepsilon\right)^2\right) \cdot \exp\left(-2\left(\tfrac{1}{2}-\varepsilon\right)^2\right)$$

$$\leq U_1 \cdot \exp\left(-2t \cdot \left(\tfrac{1}{2}-\varepsilon\right)^2\right) = \exp\left(-2t \cdot \left(\tfrac{1}{2}-\varepsilon\right)^2\right) \quad (\because U_1 = 1)$$

$E_{in}(G_T) = \frac{1}{N} \cdot \sum_{n=1}^{N} [y_n \neq G_T(x_n)] = \frac{i}{N}$ for $i = [0, N]$, $i \in \mathbb{N}$

$\Rightarrow$ if $E_{in}(G_T) < \frac{1}{N} \Rightarrow E_{in}(G_T) = 0$

we know that $E_{in}(G_T) \leq U_{T+1} \leq \exp\left(-2T\left(\frac{1}{2}-\varepsilon\right)^2\right)$

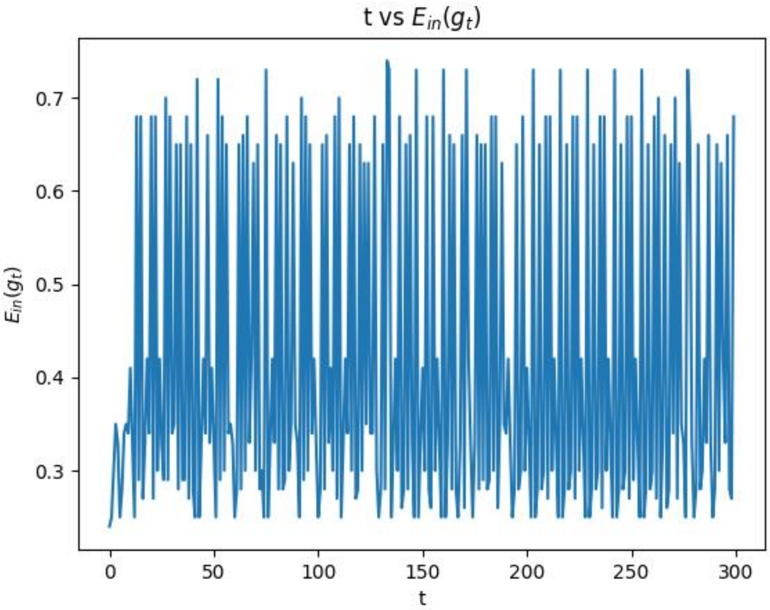
$$\exp\left(-2T\left(\tfrac{1}{2}-\varepsilon\right)^2\right) < \tfrac{1}{N} \Rightarrow N < \exp\left(2T\left(\tfrac{1}{2}-\varepsilon\right)^2\right)$$

$$\Rightarrow \ln(N) < 2T\left(\tfrac{1}{2}-\varepsilon\right)^2$$

$$\Rightarrow \ln(N) \Big/ 2\cdot\left(\tfrac{1}{2}-\varepsilon\right)^2 < T \Rightarrow T = O(\log N)$$

$\Rightarrow$ after $T = O(\log N)$ iterations, $\frac{1}{N} > E_{in}(G_T) = 0$ #

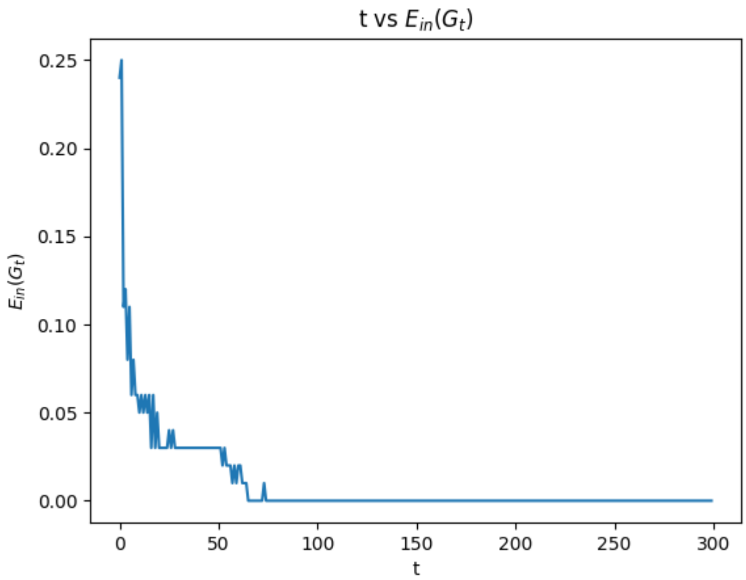
13.



$E_{in}(g_t)$的變動幅度很大，因為我們希望每次產出的 g_t 都和 g_{t-1} 差很多，所以理論上 $E_{in}(g_t)$若很小則 $E_{in}(g_{t+1})$會比較大，因此才會不斷來回變動。

$E_{in}(g_T) = 0.68$ 其實不太好，但重點是最後的 $E_{in}(G_t)$的表現。

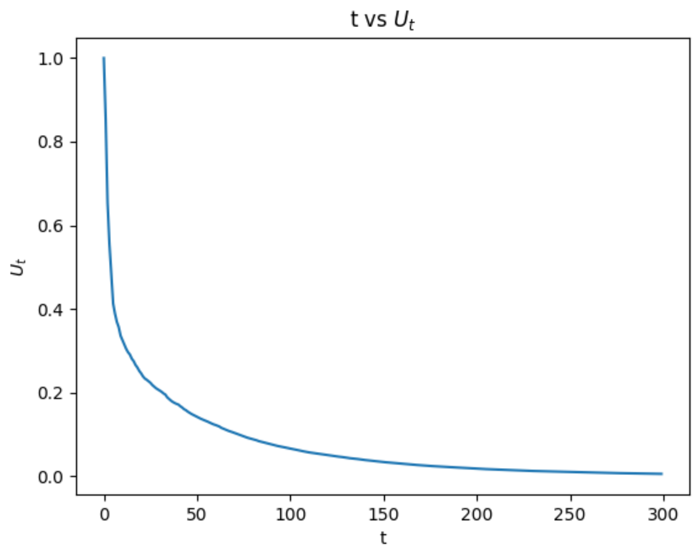
14.



$E_{in}(G_t)$一直在減少，根據 17, 18 題的證明，$E_{in}(G_t)$有個一直在減少的上限，而且在經過 O(logN) iterations 後，$E_{in}(G_t) = 0$，而在這個例字經過約 80 次迴圈後，$E_{in}(G_t) = 0$ 沒錯。

$E_{in}(G_T) = 0$

15.



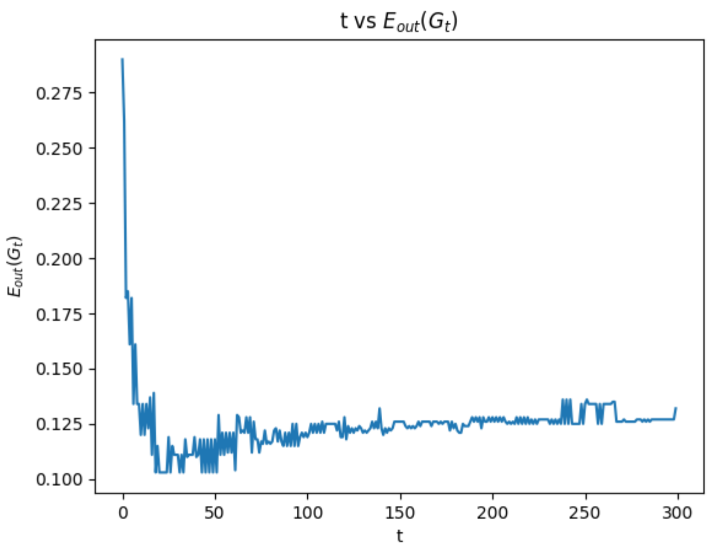
U_t是一個嚴格遞減函數，因為根據 17 題的證明，

$U_{t+1} \le U_t 2\sqrt{\epsilon(1-\epsilon)} \le U_t(\epsilon + 1 - \epsilon) = U_t$

因此他是遞減函數，而在經過 O(logN)次 iteration 後，$U_t < 1/N$ 如上圖所示。

$U_T = 0.0055 < 1/N$

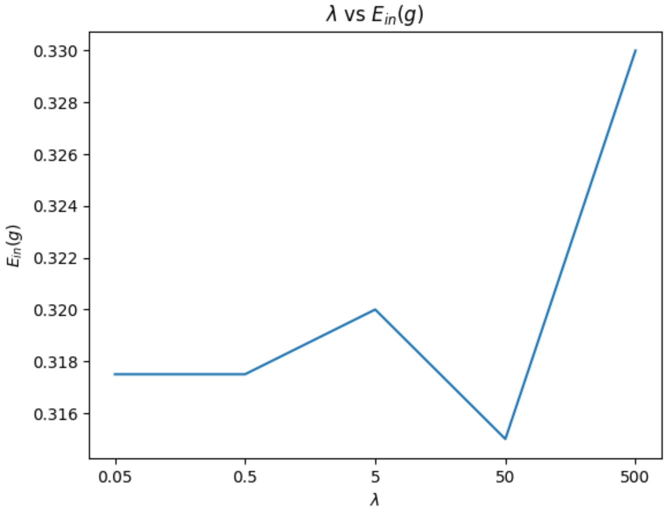
16.



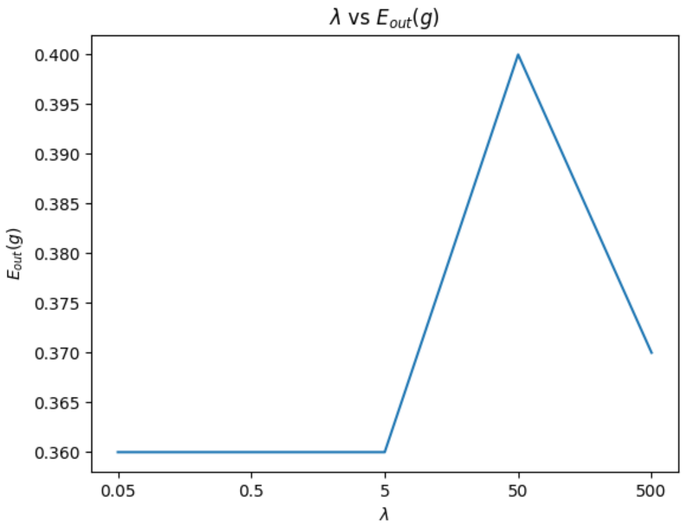
可以看到 $E_{out}(G_t)$是遞減後稍微增加的情形。我想主要是因為 dvc 在 t 增加時也會緩慢增加，因此 $E_{out}(G_t)$並不會隨著 t 不斷減少，但整體來說 $E_{out}(G_t)$還是很穩定。

$E_{out}(G_T) = 0.132$

9. $\lambda = 50$ results in the minimum $E_{in}(g)$, which $E_{in}(g) = 0.315$

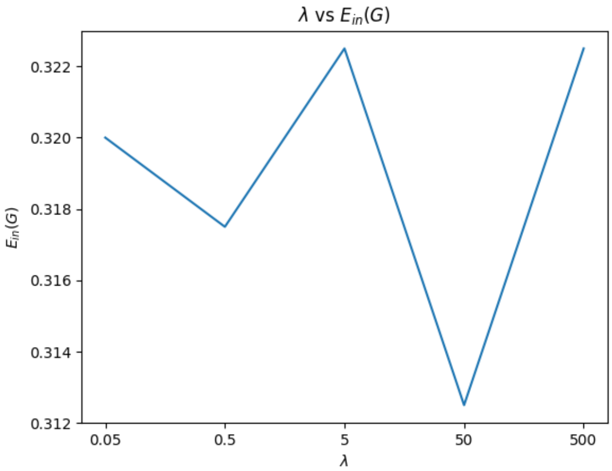


10. $\lambda = 0.05$ results in the minimum $E_{out}(g)$, which $E_{out}(g) = 0.36$



11.

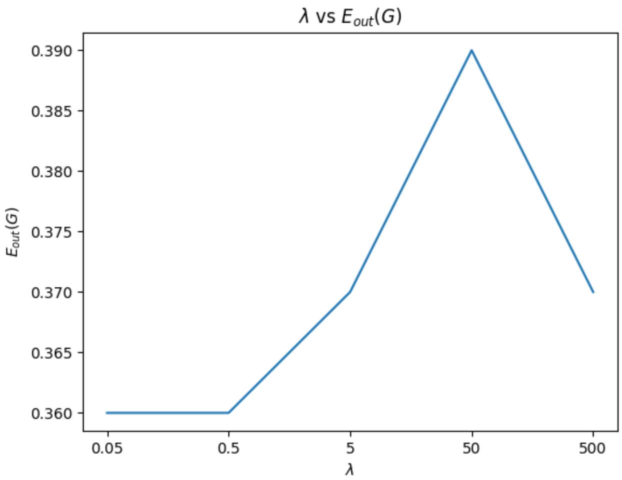
$\lambda = 50$ results in the minimum $E_{in}(g)$, which $E_{in}(g) = 0.3125$



可以看到和第 9 題相比，每個 λ 對應到的 E_{in} 其實都差不多(除了 $\lambda = 500$)，因為 bagging 雖然讓 data 變多了，但因為我們 boosting 是 400 取後放回 400 個 data，所以每個 g 都長得差不多，因此最後 E_{in} 也不會差太多。在調整 random seed 後，我發現根據取樣的不同結果也會很不穩定， $\lambda = 50$ 不一定都會對應到最小的 E_{in}，但所有的 E_{in} 都集中在 0.31~0.32 之間。

12.

$\lambda = 0.05$ results in the minimum $E_{out}(g)$, which $E_{out}(g) = 0.36$



這題的結論和 11 題很相似，最後結果也會根據取樣不同而有不同，但大部分狀況都是 $\lambda = 0.05$ 最好。可以看出這題 bagging 結果並沒有比單純 ridge regression 好，若增加 data 的 variance 應該會有所不同。