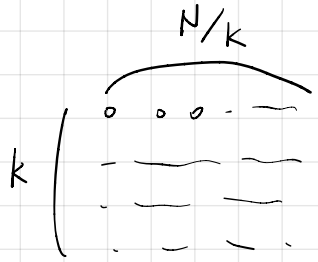
1.



total of $\binom{k}{2}$ combinations, therefore $\binom{k}{2}$ one-to-one models.

each one-to-one takes $\geq \frac{N}{K}$ values.

therefore the training process took $\binom{k}{2} \cdot a\left(\frac{2N}{K}\right)^3 = a\frac{K(K-1)}{2} \cdot \frac{8N^3}{K^3} = 4aN^3\frac{K(K-1)}{K^3}$ ✱

2.

$$\phi_Q(X_n) = \vec{Z}_n = \begin{bmatrix} 1 \\ X_n \\ X_n^2 \\ \vdots \\ X_n^{N-1} \end{bmatrix}, \text{ let } \vec{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}, \vec{y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

$$\tilde{W}^T \phi_Q(X_n) = \tilde{W}^T \vec{Z}_n = W_0 + W_1 X_n + W_2 X_n^2 + \cdots W_{N-1} X_n^{N-1}$$

$$\phi_Q(\vec{X}) = \vec{V}$$

for all $\{(X_n, Y_n)\}_{n=1}^N$, $g(\vec{X}) = \tilde{W}^T \phi_Q(\vec{X}) = \tilde{W}^T \vec{V} = \begin{pmatrix} W_0 + W_1 X_1 + W_2 X_2 + \cdots + W_{N-1} X_1^{N-1} \\ \vdots \\ W_0 + W_1 X_N + W_2 X_N + \cdots + W_{N-1} X_N^{N-1} \end{pmatrix}$

Now, we look at $\vec{V}\tilde{W} = \vec{y}$, since $\vec{V}$ is an $N \times N$ matrix,

in order for this equation to exist a solution, $\det(\vec{V}) \neq 0$, which is $\vec{V}$ is invertible.

Since we know $\det(\vec{V}) = \prod_{1 \leq n < m \leq N} (X_m - X_n)$, and for all $a_1 \sim a_N$ are different, $\det(\vec{V}) \neq 0$

then we know $\tilde{W} = \vec{V}^T \vec{y}$, there exist $\tilde{W}$ s.t. $\vec{V}\tilde{W} = \vec{y}$

$$E_{in}(g) = \sum_{n=1}^{N} (W_0 + X_n W_1 + X_n^2 W_2 + \cdots X_n^{N-1} W_{N-1} - y_n)^2$$

since $\tilde{W}\vec{V}^T = \vec{y}$

we get

$$E_{in}(g) = \sum_{n=1}^{N} (y_n - y_n)^2 = \sum_{n=1}^{N} 0 = 0$$ ✱✱

3.

let $\vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}$, $\vec{y} = \begin{pmatrix} \vec{x}_1 + \epsilon_1 \\ \vec{x}_2 + \epsilon_2 \\ \vdots \\ \vec{x}_N + \epsilon_N \end{pmatrix}$

the $\phi(\vec{x}) = \begin{pmatrix} 1 & & 0 \\ & 1 & \\ 0 & & \ddots \end{pmatrix}_{N \times N}$ = an identity matrix with size $N \times N$ = I_N

Now perform linear regression on I_N

However, since I_N is an identity matrix with size $N \times N$, it's invertible.

Hence $\tilde{w} = I_N^{-1} \vec{y} = \vec{y}$

Therefore $E_{in} = (\phi(x)\tilde{w} - \vec{y})(\phi(x)\tilde{w} - \vec{y})^T = 0$. ※

Now we check E_{out}.

First, suppose we get x from the uniform distribution (but $x \neq x_1, \dots, x_N$) where $y = x + \epsilon$

then $\phi(x) = \vec{z} = (0, 0, 0 \dots, 0)_{1 \times N}$, $g(x) = 0$, error $= (0 - x - \epsilon)^2 = x^2 + 2x\epsilon + \epsilon^2$

Next, we know the possibility for getting a specific value from a distribution is 0, which is to say $\int_x^x \frac{1}{2} dx = 0$

Therefore the possibility for getting $x_1 \sim x_N$ is 0.

Therefore $E_{out} = E(x^2 + x\epsilon + \epsilon^2) = E(x^2) + E(x\epsilon) + E(\epsilon^2) = \frac{4}{3}$ ※

1. $Var[x] = E[x^2] - (E[x])^2 \Rightarrow E[x^2] = Var[x] + (E[x])^2 = \frac{1}{12}(2)^2 + 0^2 = \frac{1}{3}$
2. $E[\epsilon^2] = Var[\epsilon] + (E[\epsilon])^2 = 1$
3. $E(x\epsilon) = E[x]E[\epsilon] = 0$
   (since x & ϵ are independent)

4. For $(X_n^T X_n)_{11} = \sum_{n=1}^{N} x_{in}^2 + \sum_{n=1}^{N} (x_{in} + \epsilon)^2 = \sum_{n=1}^{N} 2x_{in}^2 + 2x_{in}\epsilon_{in} + \epsilon^2$

$E[(X_n^T X_n)_{11}] = \sum_{n=1}^{N} 2E[x_{in}]^2 + 2E[x_{in}\epsilon_{in}] + E[\epsilon^2]$

$\epsilon \sim U[-\delta, \delta]$, $x_{in} = x_{in}$

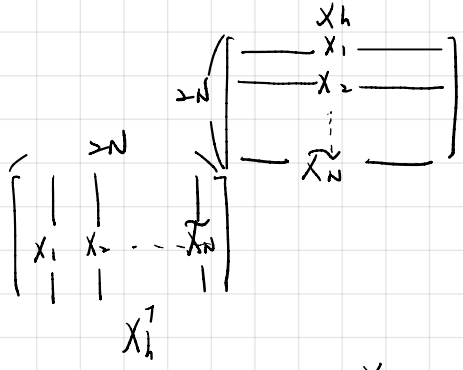
Then $E[x_{in}^2] = x_{in}^2$

$2E[x_{in}\epsilon_{in}] = 2E[x_{in}]E[\epsilon_{in}] = 0$

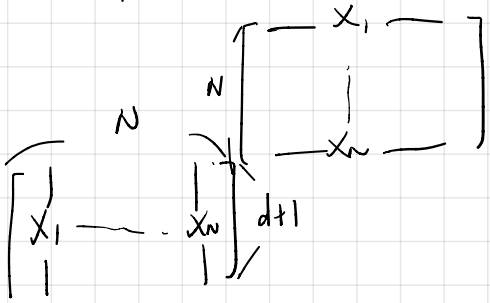
(x_{in} & ϵ_{in} are independent)

$E[\epsilon^2] = Var[\epsilon] + (E[\epsilon])^2 = \frac{1}{12}(2\delta)^2 = \frac{1}{3}\delta^2$

We get $E[(X_n^T X_n)_{11}] = \sum_{n=1}^{N} 2x_{in}^2 + \frac{1}{3}\delta^2 = \frac{N}{3}\delta^2 + \sum_{n=1}^{N} 2x_{in}^2$



X_n^T



X^T

Then $E[X_n^T X_n] = X^T X + X^T X +$

$$\begin{pmatrix} \frac{N}{3}\delta^2 & \cdots & \frac{N}{3}\delta^2 \\ \vdots & \ddots & \\ \frac{N}{3}\delta^2 & & \frac{N}{3}\delta^2 \end{pmatrix}_{(d+1)\times(d+1)}$$

5.

$E_{aug}(\vec{w}) = E_{in}(\vec{w}) + \frac{\lambda}{N}\vec{w}^T\vec{w}$

$\nabla E_{aug}(\vec{w}) = \nabla E_{in}(\vec{w}) + \frac{2\lambda}{N}\vec{w}$

$\vec{w}_{t+1} \leftarrow \vec{w}_t - \eta \nabla E_{aug}(\vec{w}_t)$

$\Rightarrow \vec{w}_t - (\eta \nabla E_{in}(\vec{w}_t) + \frac{2\lambda}{N}\eta\vec{w}_t)$

$\Rightarrow (1 - \frac{2\lambda\eta}{N})\vec{w}_t - \eta \nabla E_{in}(\vec{w}_t)$

$\Rightarrow (1 - \frac{2\lambda\eta}{N})\left(\vec{w}_t - \frac{\eta}{1 - \frac{2\lambda\eta}{N}} \nabla E_{in}(\vec{w}_t)\right)$

$\frac{\eta}{1 - \frac{2\lambda\eta}{N}} = \frac{N}{N - 2\lambda\eta}$

then $\alpha = 1 - \frac{2\lambda\eta}{N}$, $\beta = \frac{N}{N - 2\lambda\eta}$ ※

6.

let $f(w) = \frac{1}{N}\sum_{n=1}^{N}(w\cdot x_n - y_n)^2$

in order to maximize $f(w)$ under $g(w)=w^2 \le C$, we find lagrange multiplier such that

$\nabla f = \lambda \nabla g(w)$

$\Rightarrow \frac{1}{N}\sum_{n=1}^{N} 2(w\cdot x_n - y_n)(x_n) = 2w\lambda \Rightarrow \left(\sum_{n=1}^{N} w x_n^2\right) - Nw\lambda = \sum_{n=1}^{N} x_n y_n \Rightarrow w^* = \frac{\sum_{n=1}^{N} x_n y_n}{\left(\sum_{n=1}^{N} x_n^2\right) - N\lambda}$

since $(w^*)^2 = C \Rightarrow \frac{\sum_{n=1}^{N} x_n y_n}{\left(\sum_{n=1}^{N} x_n^2\right) - N\lambda} = \sqrt{C} \Rightarrow \frac{\sum_{n=1}^{N} x_n y_n}{\sqrt{C}} = \left[\left(\sum_{n=1}^{N} x_n^2\right) - N\lambda\right]$

$$\Rightarrow \lambda = \frac{-\sum_{n=1}^{N} x_n y_n}{N\sqrt{C}} + \frac{1}{N}\sum_{n=1}^{N} x_n^2$$

therefore $\alpha = \frac{-\sum_{n=1}^{N} x_n y_n}{N}$

$\beta = \frac{1}{N}\sum_{n=1}^{N} x_n^2$ ※

7.

$E(\vec{w}) = \frac{1}{N}\sum_{n=1}^{N}(\tilde{\vec{w}}^T V \vec{x}_n - y_n)^2 + \frac{\lambda}{N}\|\tilde{\vec{w}}\|$

let $\tilde{\vec{w}}^T V = \vec{w}^T = [V_{11}w_1, V_{22}w_2, \ldots V_{dd}w_d]$

$\Rightarrow \tilde{\vec{w}}^T = \vec{w}^T V^{-1}$ (V is invertible, since V is diagnal matrix storing positive values)

since V is a diagnal matrix, we know $V^{-1} = \begin{bmatrix} \frac{1}{v_{11}} & & 0 \\ & \frac{1}{v_{22}} & \\ 0 & & \ddots \\ & & & \frac{1}{v_{dd}} \end{bmatrix}$

we also have $\vec{w} = \begin{bmatrix} V_{11}w_1 \\ V_{22}w_2 \\ \vdots \\ V_{dd}w_d \end{bmatrix} = V\tilde{\vec{w}} \Rightarrow \tilde{\vec{w}} = V^{-1}\vec{w}$

then $E(\tilde{\vec{w}}) = \frac{1}{N}\sum_{n=1}^{N}(w^T x_n - y_n)^2 + \frac{\lambda}{N}\|V^{-1}\vec{w}\|$

then $\Omega(\vec{w}) = V^{-1}\vec{w}$

and $\vec{w} = V\tilde{\vec{w}}$ ※※

8.

① We remove one positive example, $A_{minority}$ will always return positive ($\because N-1 < N$)

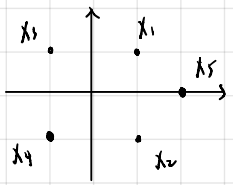
therefore $e=0$, and this will always be the same for all N positive examples.

② Then we remove one negative sample, $A_{minority}$ will always return negative ($\because N-1 < N$)

therefore $e=0$, and this will always be true for all negative examples.

Hence $E_{loocv}(A_{minority}) = 0$ ※

9.



Situation $E_{in}=0$ ⇒ $x_1 \sim x_5$ different to the rest. ⇒ 22 combinations

x_1, x_5 same

x_2, x_5 same

x_1, x_3 same

x_3, x_4 same

x_2, x_4 same

all the same

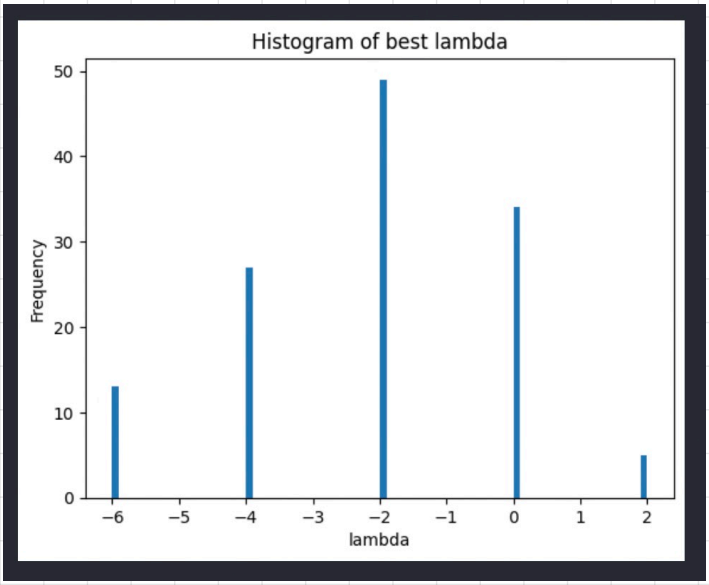
$E_{in}=1$ ⇒ (x_1, x_5), (x_4, x_5), (x_2, x_3), (x_1, x_4), (x_1, x_2) same ⇒ 10 combinations

Then the expectation $= \frac{10}{32} \times 1 + \frac{22}{32} \times 0 = \frac{5}{16}$ ※

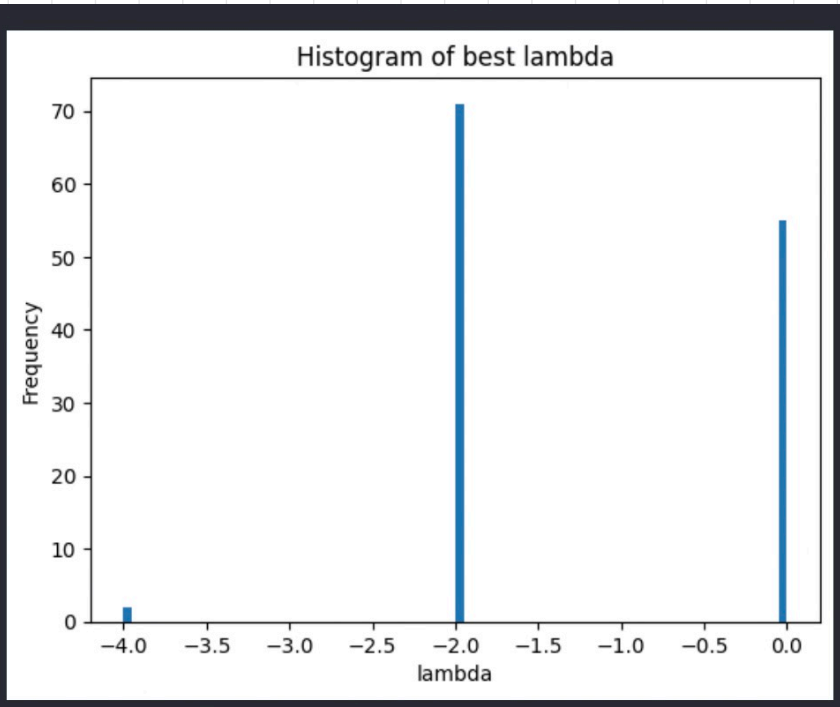
10.

```
Training with lambda: 1e-06
(200, 84) (200,)
Accuracy = 96% (192/200) (classification)

Training with lambda: 0.0001
(200, 84) (200,)
Accuracy = 92% (184/200) (classification)

Training with lambda: 0.01
(200, 84) (200,)
Accuracy = 91% (182/200) (classification)

Training with lambda: 1.0
(200, 84) (200,)
Accuracy = 87.5% (175/200) (classification)

Training with lambda: 100.0
(200, 84) (200,)
Accuracy = 80.5% (161/200) (classification)

Best lambda: 1e-06
Best lambda(log_10): -6.0
Best accuracy: 96.0
```

11.



12.



From Q10, we get the best lamda is 1e-6.

From Q11, we can see that lamda value 1e-2 appeared the most.

From Q12, we see that lambda value with 1e-2 appeared the most.

Therefore we know from cross validation that, when lamda = 1e-2, the err would be minimize.

Moreover, the distribution for Q12 is not as wide as Q11, only three values are left.

The distribution for Q11 is more like normal distribution.

We know from this experiment, cross validation can limit the amount of "good" lamdas than simply separating our data set.

Through using folds to varify which lambda is good, we reduce the amount of good lambdas into 3 rathere than 5.

13.

$\min_w E_{in}(w)$ subject to $w^T w \le C$

Is the same as finding Lagrange multiplier $\lambda > 0$ and $\vec{w}_{reg}$

such that $\nabla E_{in}(\vec{w}_{reg}) + \frac{2\lambda}{N} w = 0 \Rightarrow -\frac{2\lambda}{N}\vec{w}_{reg} = \nabla E_{in}(\vec{w}_{reg})$

Hence we know that $\vec{w}_{reg}$ and $-\nabla E_{in}(\vec{w}_{reg})$ are only different with a multiplying factor.

Moreover $-\nabla E_{in}(\vec{w}_{reg})$ point towards w_{LIN}, if and only if $X^T X$ is reversible

which is to say $X^T X = \alpha I$

therefore $\vec{w}_{reg} + \left(-\nabla E_{in}(\vec{w}_{reg})\right) t = \vec{w}_{LIN} \Rightarrow n\,\vec{w}_{reg} = \vec{w}_{LIN}$

and this this the form of $\vec{w}_C = \vec{w}_{LIN} \times \frac{\sqrt{C}}{\|w_{LIN}\|}$ when we set

$$n = \frac{\sqrt{C}}{\|w_{LIN}\|}$$

Hence we prove w_C solves the C-constrained regression

problem $\min_w E_{in}(w)$ subject to $w^T w \le C$ T.f.f. $X^T X = \alpha I$