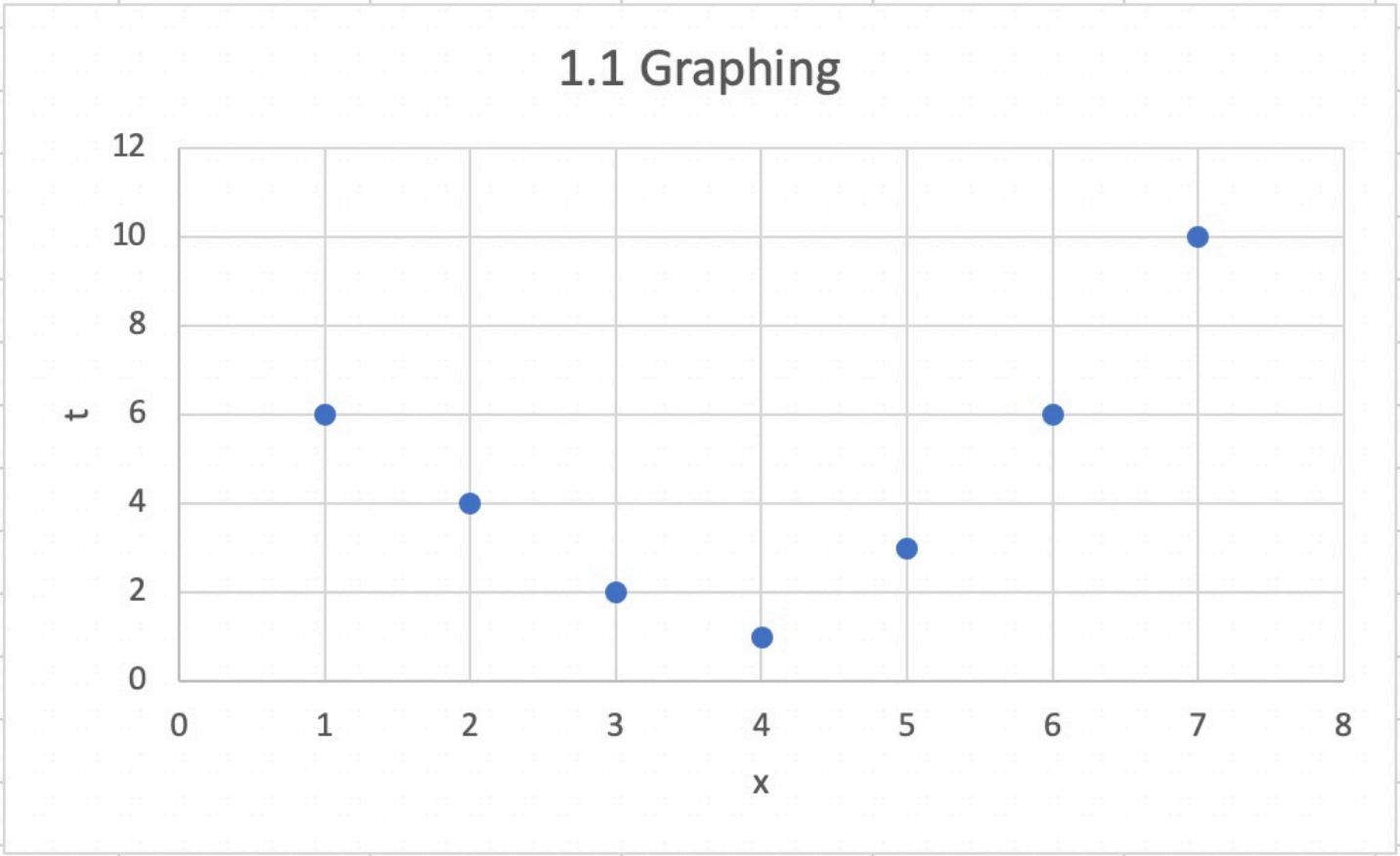
1.1

x	t
1	6
2	4
3	2
4	1
5	3
6	6
7	10



1.2. $g_{w,b}(x) = wx + b$

Least squares loss

$$\mathcal{E}(w,b) = \frac{1}{2N}\sum_{i=1}^{N}(g^{(i)} - t^{(i)})^2 \qquad g_{w,b}(wx+b)$$

first term.

$$= \frac{1}{2N}\Big[\{(wx^{(1)}+b) - t^{(1)}\}^2 + \{(wx^{(2)}+b) - t^{(2)}\}^2 +$$
$$\{(wx^{(3)}+b) - t^{(3)}\}^2 + \{(wx^{(4)}+b) - t^{(4)}\}^2 +$$
$$\{(wx^{(5)}+b) - t^{(5)}\}^2 + \{(wx^{(6)}+b) - t^{(6)}\}^2 +$$
$$\{(wx^{(7)}+b) - t^{(7)}\}^2\Big]$$

Showing computation only for the first term.

$$\{wx^{(1)} + b - t^{(1)}\}^2 = [wx^{(1)} + b - t^{(1)}][wx^{(1)} + b - t^{(1)}]$$
$$= w^2x^{(1)2} + bwx^{(1)} - t^{(1)}wx^{(1)}$$
$$+ bwx^{(1)} + b^2 - bt^{(1)}$$
$$- wt^{(1)}x^{(1)} - t^{(1)}b + t^{(1)2}$$
$$= x^{(1)2}w^2 + b^2 + 2x^{(1)}wb - 2x^{(1)}t^{(1)}w - 2t^{(1)}b + t^{(1)2}$$

It would be the same for all terms except the $x^{(1)}$ and $t^{(1)}$ will be in its own index.

$$\Rightarrow \frac{1}{2N}\Big[\sum_{i\in 1,\dots N}(x^{(i)})^2 w^2 + (1)b^2 + (2x^{(i)})wb + (-2x^{(i)}t^{(i)})w + (-2t^{(i)})b + (t^{(i)})^2\Big]$$

Form wanted $\Rightarrow \frac{1}{2N}\sum_{i\in 1,\dots N} A_i w^2 + B_i b^2 + C_i wb + D_i w + E_i b + F_i$

∴ From side by side comparison, we know

$A_i = [x^{(i)}]^2$, $B_i = 1$, $C_i = 2x^{(i)}$ ~~for $i \in 1,\dots N$~~

$D_i = [-2x^{(i)}t^{(i)}]$, $E_i = -2t^{(i)}$, $F_i = [t^{(i)}]^2$

morning glory

1.3.

From question 2,

$$E(w,b) = \frac{1}{2N}\sum_{i=1\ldots N}\left[(x^{(i)})^2 w^2 + (1)b^2 + (2x^{(i)})wb + (-2x^{(i)}t^{(i)})w + (-2t^{(i)})b + [t^{(i)}]^2\right]$$

Let $A=\sum_i A_i$, $B=\sum_i B_i$, ...

$$= \frac{1}{2N}\left\{A(Nw^2) + B(Nb^2) - C(Nwb) + D(Nw) + E(Nb) + F\right\}$$

$$= \frac{1}{2}Aw^2 + \frac{1}{2}Bb^2 + \frac{1}{2}Cwb + \frac{1}{2}Dw + \frac{1}{2}Eb + \frac{F}{2N}$$

To minimize,

$$\frac{\partial[E(w,b)]}{\partial w} = 0 \;;\; Aw + \frac{1}{2}Cb + \frac{1}{2}D = 0$$

$$2Aw + Cb + D = 0$$

$$\Rightarrow w = \frac{-D - Cb}{2A} \;;\; \text{sub 'b'}$$

$$\frac{\partial[E(w,b)]}{\partial b} = 0 \;;\; Bb + \frac{1}{2}Cw + \frac{1}{2}E = 0$$

$$b = \frac{-Cw - E}{2B}$$

To isolate without each dependency, sub in,

Sub 'b' into w,

$$w = \frac{-D - C\left[\frac{-Cw-E}{2B}\right]}{2A}$$

$$2A2Bw = -D(2B) + C^2w + CE$$

$$(2A2B - C^2)w = -2BD + CE$$

$$w = \frac{-2BD + CE}{2A2B - C^2}$$

Sub 'w' into 'b'

$$b = \frac{-C\left[\frac{-2BD+CE}{2A2B-C^2}\right] - E}{2B}$$

$$2Bb[2A2B - C^2] = 2CBD + C^2E - E[2A2B - C^2]$$

$$b = \frac{2CBD + C^2E - E[2A2B - C^2]}{2B[2A2B - C^2]}$$

morning glory

1.4.

To get numerical values, first evaluate A, B, C, D, E, F with the data.

$$A = \sum [x^{(i)}]^2 = 140$$

$$B = \sum B_i = 7$$

$$C = \sum [2x^{(i)}] = 56$$

$$D = \sum -2x^{(i)} t^{(i)} = -290$$

$$E = \sum -2t^{(i)} = -64$$

$$F = \sum [t^{(i)}]^2 = 202$$

$$w = \frac{-2BD + CE}{2AB - C^2} = \frac{476}{784} = \boxed{\frac{17}{28}}$$

$$b = \left(-C\left[\frac{-2BD + CE}{2AB - C^2}\right] - E\right) \frac{1}{2B}$$

$$= \left[-56\left(\frac{476}{784}\right) + 64\right] \frac{1}{14}$$

$$= \boxed{\frac{15}{7}}$$

1.5 verified with excel

2.1

$g_{w,b}(x) = wx + b \Rightarrow g_w(\vec{x}) = \vec{x}\,\vec{w}$; $\vec{x} = (x^{(i)}, 1)$

↳ $= [x \;\; 1]\begin{bmatrix} w \\ b \end{bmatrix} \Rightarrow = [x^{(i)} \;\; 1]\begin{bmatrix} w \\ b \end{bmatrix} = \dots$

trying to put the known expression into a vector form

$\therefore \vec{w} = \begin{bmatrix} w \\ b \end{bmatrix}$

2.2. $\bar{x} = \begin{bmatrix} x^{(1)} & 1 \\ \vdots & \\ x^{(N)} & 1 \end{bmatrix}$ $\quad \bar{x}^T = \begin{bmatrix} x^{(1)} & x^{(2)} & \cdots & x^{(N)} \\ 1 & \cdots & \cdots & 1 \end{bmatrix}$

$\bar{x}\bar{w} = \begin{bmatrix} wx^{(1)} + b \\ \vdots \\ wx^{(N)} + b \end{bmatrix}$, $\quad \bar{x}\bar{w} - t = \begin{bmatrix} wx^{(1)} + b^{(1)} - t^{(1)} \\ \vdots \\ wx^{(N)} + b^{(N)} - t^{(N)} \end{bmatrix}$

$$\|\bar{x}\bar{w} - t\|^2 = \left[wx^{(1)} + b^{(1)} - t^{(1)}\right]^2 + \cdots + \left[wx^{(N)} + b^{(N)} - t^{(N)}\right]^2$$

$$\nabla_{\bar{w}} \|\bar{x}\bar{w} - t\|^2 = \begin{bmatrix} \frac{\partial}{\partial w}\left\{\left[wx^{(1)} + b^{(1)} - t^{(1)}\right]^2 + \cdots + \left[wx^{(N)} + b^{(N)} - t^{(N)}\right]^2\right\} \\ \frac{\partial}{\partial b}\left\{\left[wx^{(1)} + b^{(1)} - t^{(1)}\right]^2 + \cdots + \left[wx^{(N)} + b^{(N)} - t^{(N)}\right]^2\right\} \end{bmatrix}$$

$$= \begin{bmatrix} 2\left[wx^{(1)} + b^{(1)} - t^{(1)}\right]x^{(1)} + \cdots + 2\left[wx^{(N)} + b^{(N)} - t^{(N)}\right]x^{(N)} \\ 2\left[wx^{(1)} + b^{(1)} - t^{(1)}\right] + \cdots + 2\left[wx^{(N)} + b^{(N)} - t^{(N)}\right] \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^{N} 2\left[wx^{(i)} + b^{(i)} - t^{(i)}\right]x^{(i)} \\ \sum_{i=1}^{N} 2\left[wx^{(i)} + b^{(i)} - t^{(i)}\right] \end{bmatrix}$$

$$= \begin{bmatrix} 2\sum_{i=1}^{N}\left[wx^{(i)} + b^{(i)}\right]x^{(i)} - 2\sum_{i=1}^{N} x^{(i)}t^{(i)} \\ 2\sum_{i=1}^{N}\left[wx^{(i)} + b^{(i)}\right] - 2\sum_{i=1}^{N}\left[t^{(i)}\right] \end{bmatrix}$$

2.2 continued

$$= 2 \begin{bmatrix} \sum_{i=1}^{N} [W x^{(i)} + b^{(i)}] \\ \sum_{i=1}^{N} [W x^{(i)} + b^{(i)}] \end{bmatrix} \begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix} - 2 \begin{bmatrix} x^{(1)} & \cdots & x^{(N)} \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} | \\ t^{(i)} \\ | \end{bmatrix}$$

$$= 2 \begin{bmatrix} - & x^{(i)} & - \\ & 1 & \end{bmatrix} X \bar{w} - 2 X^T \bar{t}$$

$$= 2 X^T X \bar{w} - 2 X^T \bar{t}$$

$$= 2 X^T (X \bar{w} - \bar{t})$$

2.3 From the question before we know

$$\nabla_{\bar{w}} \|X\bar{w} - \bar{t}\|^2 = 2X^T(X\bar{w}^* - t) \Leftrightarrow 2X^TX\bar{w}^* - 2X^T\bar{t} = 0 \quad \text{(given)}$$

→ To minimize, let the gradient equals zero.

$$2 \underbrace{\begin{bmatrix} x^{(1)} & \cdots & x^{(N)} \\ 1 & \cdots & 1 \end{bmatrix}}_{2\times N} \underbrace{\begin{bmatrix} x^{(1)}w + b - t^{(1)} \\ \vdots \\ x^{(N)}w + b - t^{(N)} \end{bmatrix}}_{N\times 1} = 0$$

$$= 2 \begin{bmatrix} \sum (x^{(i)}w + b - t^{(i)})(x^{(i)}) \\ \sum x^{(i)}w + b - t^{(i)} \end{bmatrix} = \begin{bmatrix} \sum x^{(i)2}w + x^{(i)}b - t^{(i)}x^{(i)} \\ \sum x^{(i)}w + b - t^{(i)} \end{bmatrix} \overset{!}{=} \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Let's try plugging the numbers from problem 1,

$\{x^{(i)}\}_{i=1,\dots,7} = (1, 2, \dots, 7)$, $\{t^{(i)}\}_{i=1,\dots,7} = (6.4, 2, 1.3, 6, 10)$, $w = \frac{17}{28}$ $b = \frac{15}{7}$

$$\rightarrow \begin{bmatrix} (-3.25) + (-1.29) + (5.89) + (14.29) + (10.89) + (-1.29) + (-25.25) \\ (-3.25) + (-0.64) + (1.96) + (3.57) + (2.18) + (-0.214) + (-3.61) \end{bmatrix}$$

$$= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \therefore \text{QED}$$

$$\therefore 2X^TX\bar{w}^* - 2X^T\bar{t} = 0$$

∴ QED.

While it can be numerically seen, also analytically.

from 2.2 we've driven:

$$2\begin{bmatrix} \sum_{i=1}^{N}[wx^{(i)}+b^{(i)}] \\ \sum_{i=1}^{N}[wx^{(i)}+b^{(i)}] \end{bmatrix}\begin{bmatrix} x^{(i)} \\ 1 \end{bmatrix} - 2\begin{bmatrix} x^{(1)} & \cdots & x^{(N)} \\ 1 & \cdots & 1 \end{bmatrix}\begin{bmatrix} | \\ t^{(i)} \\ | \end{bmatrix}$$

we can put this into a form

$$= 2\begin{bmatrix} — & x^{(i)} & — \\ — & 1 & — \end{bmatrix} \times \bar{w} - 2X^T\bar{t}$$

$$= 2X^TX\bar{w} - 2X^T\bar{t}$$

Then we know that for $\bar{w}^*$ that minimizes the least squares loss, the gradient $= 0$.

$$\Rightarrow \quad 2X^TX\bar{w} - 2X^T\bar{t} = 0 \quad \text{for } \bar{w}^*$$

∴ This is more of a general proof of how we land to the given. (It required bit of a back tracking from the previous question)

2.4

$$\nabla_{\bar{w}} \|X\bar{w} - t\|^2 = 2X^T X\bar{w}^* - 2X^T\bar{t} = 0 \quad \text{for } \bar{w}^*$$

$$2X^T X\bar{w}^* = 2X^T\bar{t}$$

$$\cancel{2}\left[\cancel{(X^TX)^{-1}}\overrightarrow{I}(X^TX)\bar{w}^*\right] = \cancel{2}(X^TX)^{-1}X^T\bar{t}$$

$$\bar{w}^* = (X^TX)^{-1}X^T\bar{t}$$

2.5 verified with Numpy

Date. No.

3.1.

$$A = \sum_{i \in 1 \ldots N} \bar{x}^{(i)} \bar{x}^{(i)T} \quad ; \quad x_j^{(i)} \; j \in 1 \ldots d$$

$$= \sum_{i \in 1 \ldots N} \begin{bmatrix} x_1^{(i)} \\ x_2^{(i)} \\ \vdots \\ x_d^{(i)} \end{bmatrix} \begin{bmatrix} x_1^{(i)} & \cdots & x_d^{(i)} \end{bmatrix}$$

$$= \begin{bmatrix} x_1^{(1)} x_1^{(1)} & x_1^{(1)} x_2^{(1)} & \cdots & x_1^{(1)} x_d^{(1)} \\ \vdots & & & \vdots \\ x_d^{(1)} x_1^{(1)} & \cdots & & x_d^{(1)} x_d^{(1)} \end{bmatrix} + \cdots + \begin{bmatrix} x_1^{(N)} x_1^{(N)} & \cdots & x_1^{(N)} x_d^{(N)} \\ \vdots & & \vdots \\ x_d^{(N)} x_1^{(N)} & \cdots & x_d^{(N)} x_d^{(N)} \end{bmatrix}$$

$$= \sum_{i \in 1 \ldots N} \begin{bmatrix} x_1^{(i)} x_1^{(i)} & x_1^{(i)} x_2^{(i)} & \cdots & x_1^{(i)} x_d^{(i)} \\ \vdots & & & \vdots \\ x_d^{(i)} x_1^{(i)} & \cdots & & x_d^{(i)} x_d^{(i)} \end{bmatrix}$$

$x^{(i)} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \end{bmatrix}$ $\bar{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}$ Date. No.

3.2

$b = \sum_{i=1}^{n} t^{(i)} \bar{x}^{(i)}$, show that $\nabla E(\bar{w}, D) = \frac{1}{N}(A\bar{w} - \bar{b}) + \lambda \bar{w}$

From the loss function,

$$E(\bar{w}, D) = \frac{1}{2N} \sum_{i \in 1 \ldots N} (g_{\bar{w}}(\bar{x}^{(i)}) - t^{(i)})^2 + \frac{\lambda}{2} \|w\|_2^2$$

$$= \frac{1}{2N} \sum_{i \in 1 \ldots N} \left[\underbrace{(\bar{x}^{(i)})^T \bar{w}}_{\text{scalar}} - \underbrace{t^{(i)}}_{\text{scalar}} \right]^2 + \frac{\lambda}{2} \|w\|_2^2$$

(1×d) (d×1)

Square rooting the whole sum of square of elements = l_2 norm

$$= \frac{1}{2N} \|\bar{X}^T \bar{w} - \bar{t}\|_2^2 + \frac{\lambda}{2} \|w\|_2^2 \quad ; \text{ capitals absorbed (took account) of the summation.}$$

Now, take the gradient.

$$\nabla E(\bar{w}, \bar{D})$$

known from problem 2.2 ⊛

$$= \frac{1}{N}(X)(\bar{X}^T \bar{w} - \bar{t}) + \lambda \bar{w}$$

⊛ Side proof of how $\nabla \|\bar{w}\|_2^2 = 2\bar{w}$

$$\|\bar{w}\|_2^2 = \left(\sqrt{\sum_{i=1}^{N} (w_i)^2}\right)^2$$

$$= \sum_{i=1}^{N} w_i^2$$

taking the derivative,

$$\nabla \sum_{i=1}^{N} w_i^2 = \sum_{i=1}^{N} \nabla(w_i^2) \qquad (\nabla \text{ operation } \mathbb{R}^n \to \mathbb{R})$$

$$= 2 \begin{bmatrix} \frac{\partial \Sigma w}{\partial w_1} \\ \frac{\partial \Sigma w}{\partial w_2} \\ \vdots \\ 1 \end{bmatrix}$$

$$= 2 \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ 1 \end{bmatrix}$$ partial derivative makes a column vector

$$= 2\bar{w}$$

Rewriting.

$$\frac{1}{N}(X)(X^T\bar{w} - \bar{t}) + \lambda\bar{w}$$

$$= \frac{1}{N}\left[\underbrace{XX^T}_{A}\bar{w} - X\bar{t}\right] + \lambda\bar{w}$$

$$= \frac{1}{N}\left[A\bar{w} - X\bar{t}\right] + \lambda\bar{w}$$

$X\bar{t}$ ↙

$$= \begin{bmatrix} \bar{x}^{(1)} & \bar{x}^{(2)} & \cdots & \bar{x}^{(N)} \end{bmatrix} \begin{bmatrix} t^{(1)} \\ \vdots \\ t^{(N)} \end{bmatrix}$$

$$= \bar{x}^{(1)}t^{(1)} + \cdots + \bar{x}^{(N)}t^{(N)}$$

$$= \sum \bar{x}^{(i)}t^{(i)} \quad \Rightarrow \text{this is } b \text{ by definition.}$$

On the document it is given as $\vec{b} = \sum_{i \in 1...N} t^{(i)}\bar{x}^{(i)}$

but this is equivalent to $\vec{b} = \sum_{i \in 1...N} \bar{x}^{(i)}t^{(i)}$

$\because t^{(i)}$ are scalars

$$\therefore \nabla\mathcal{E}(\vec{w}, D) = \frac{1}{N}\left[A\bar{w} - \bar{b}\right] + \lambda\bar{w} \quad \therefore \text{QED.}$$

3.3

$$\vec{w}^* = \arg\min_{\vec{w}} \varepsilon(\vec{w}, D)$$

∵ want to minimize the loss

⇒ gradient of loss function = 0.

From the previous question, found the gradient of the loss function.

$$\frac{1}{N}(A\vec{w}^* - \vec{b}) + \lambda\vec{w}^* = 0$$

$$(A\vec{w}^* - \vec{b}) + N\lambda\vec{w}^* = 0$$

$$(A + \lambda N I_b)\vec{w}^* - \vec{b} = 0$$

$$(A + \lambda N I_b)\vec{w}^* = \vec{b} \quad ; \text{ equation be to satisfied for } \vec{w}^*$$

3.4

prove all eigenvalues of A are non-negative.

⇒ prove that $\exists\, \bar{v}$ s.t. $A\bar{v} = \lambda\bar{v}$ for all $\lambda \geq 0$.

$$\bar{v}^T A\bar{v} = \bar{v}^T \lambda \bar{v}$$

LHS: let $\bar{w}$ be an arbitrary none zero vector.

$$\bar{w}^T A\bar{w} = \bar{w}^T (\Sigma x x^T)\bar{w} \quad ; \text{ from 3.1}$$

$$= \Sigma \bar{w}^T x x^T \bar{w} \quad ; \text{ took sum outside}$$

$$= \Sigma (x^T\bar{w})^T (x^T\bar{w})$$

$\because x^T x = \|x\|_2^2$

$$= \Sigma \|x^T\bar{w}\|_2^2 \geq 0$$

$\because$ squared, cannot be less than 0.

RHS:

$$\bar{w}^T \lambda \bar{w} = \lambda \bar{w}^T \bar{w} \quad ; \text{ possible since } \lambda \text{ is a scalar}$$

$$= \lambda \|\bar{w}\|_2^2$$

$\bar{w}$ is a nonzero vector and through the norm square, it will be ≥ 0

$\because$ previously shown that LHS ≥ 0

$\therefore \lambda \geq 0$.

3.5

$A + \lambda N I_d$ is invertible. $\Leftrightarrow$ none of its eigenvalues are zero.

$$(A+\lambda N I_d)\bar{v} = \gamma \bar{v}$$

$$A\bar{v} + \lambda N\bar{v} = \gamma\bar{v}$$

$$A\bar{v} = (\gamma - \lambda N)\bar{v}$$

$\because$ eigenvalue$(A) > 0$ proven in previous question

$\therefore \gamma - \lambda N > 0 \rightarrow \gamma > \lambda N$

Knowing $\lambda > 0$ and $N > 0 \Rightarrow \therefore \gamma > 0$

$\because$ proven that its eigenvalues are none zero (greater than zero)

$\therefore A + \lambda N I_d$ is invertible.

3.6

From 3.3, we know $(A + \lambda N)\vec{w}^* = \vec{b}$

$$\underbrace{(A+\lambda N I_d)^{-1}(A+\lambda N I_d)}_{\text{produces identity matrix}}\vec{w}^* = (A+\lambda N I_d)^{-1}\vec{b}$$

$$\therefore \vec{w}^* = (A+\lambda N I_d)^{-1}\vec{b}$$