HW1

Name: Jerry Chen

USC ID: [redacted]

1.

$w_1 = \begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix}$, $w_2 = \begin{bmatrix} 2 & 2 \\ 2 & -3 \end{bmatrix}$, $b_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $b_2 = \begin{bmatrix} 0 \\ -4 \end{bmatrix}$

$\mathrm{ReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$ $\quad X = \begin{bmatrix} +1 \\ -1 \end{bmatrix}$

$a_1 = h(w_1 x_1 + b_1)$

$= h\left(\begin{bmatrix} 1 & -2 \\ 3 & 4 \end{bmatrix} \cdot \begin{bmatrix} +1 \\ -1 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right)$

$= h\left(\begin{bmatrix} 1+2+1 \\ 3-4+0 \end{bmatrix}\right) = h\left(\begin{bmatrix} 4 \\ -1 \end{bmatrix}\right) = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$

$a_2 = h(w_2 a_1 + b_2)$

$= h\left(\begin{bmatrix} 2 & 2 \\ 2 & -3 \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ -4 \end{bmatrix}\right)$

$= h\left(\begin{bmatrix} 8+0 \\ 8-4 \end{bmatrix}\right) = h\left(\begin{bmatrix} 8 \\ 4 \end{bmatrix}\right) = \boxed{\begin{bmatrix} 8 \\ 4 \end{bmatrix}}$

2,
$f(x,y) = 4x^2 + y^2 - xy - 13x$

(a)

$$\frac{\partial f}{\partial x} = \frac{\partial}{\partial x}(4x^2 + y^2 - xy - 13x)$$

$$= 8x + 0 - y - 13$$

$$= \boxed{8x - y - 13}$$

$$\frac{\partial f}{\partial y} = \frac{\partial}{\partial y}(4x^2 + y^2 - xy - 13x)$$

$$= 0 + 2y - x$$

$$= \boxed{2y - x}$$

(b) find critical point → minimizes f

$$\begin{cases} 8x - y - 13 = 0 & \text{—①} \\ 2y - x = 0 & \text{—②} \end{cases}$$

② $x = 2y$

① $16y - y - 13 = 0$

$15y = 13 \quad y = \frac{13}{15}$

$x = 2y \rightarrow x = \frac{26}{15}$

$$\boxed{\text{minimizes } f\left(\frac{26}{15}, \frac{13}{15}\right)}$$

3.

$$w^T x_0 - w^T x$$

(a) $\min_x \|x_0 - x\|_2$ s.t. $w^T x + b = 0$

$$L(x, \lambda) = (x_0 - x)^T (x_0 - x) - \lambda (w^T x + b)$$

$$= x_0^T x_0 - x^T x_0 - x_0^T x + x^T x - \lambda (w^T x + b)$$

$$= x_0^T x_0 - 2 x^T x_0 + x^T x - \lambda (w^T x + b)$$

$$\frac{\partial L}{\partial x} = -2 x_0 + 2x - \lambda w = 0$$

$$\Rightarrow -2(x_0 - x) - \lambda w = 0$$

Dot with w^T

$$-2 w^T (x_0 - x) - \lambda w^T w = 0 \Rightarrow \lambda = -\frac{2 w^T (x_0 - x)}{w^T w}$$

Dot with $(x_0 - x)^T$

$$-2 (x_0 - x)^T (x_0 - x) - \lambda (x_0 - x)^T w = 0$$

$$\Rightarrow (x_0 - x)^T (x_0 - x) = \frac{w^T (x_0 - x)}{w^T w} (x_0 - x)^T w$$

$$\Rightarrow (x_0 - x)^T (x_0 - x) = \frac{(w^T (x_0 - x))^2}{\|w\|^2}$$

$$\Rightarrow (x_0 - x)^T (x_0 - x) = \frac{w^T (x_0 - x)}{\|w\|}$$

$$\Rightarrow (x_0 - x)^T (x_0 - x) = \boxed{\frac{w^T (x_0 + b)}{\|w\|}}$$

3.

(b)

$w^T x + b_1 = 0$ — ①

$w^T x + b_2 = 0$ — ②

$\because$ ① and ② are parallel

$\therefore$ base on 3(a) distance $= \boxed{\dfrac{|b_2 - b_1|}{\|w\|}}$

4.

(a)

$f(x) = x^2$

$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \quad \forall x, y$ and $0 < \lambda < 1$

$f(\lambda x + (1-\lambda)y) = (\lambda x + (1-\lambda)y)^2$

$= \lambda^2 x^2 + 2\lambda(1-\lambda)xy + (1-\lambda)^2 y^2$

$\lambda f(x) + (1-\lambda) f(y) = \lambda x^2 + (1-\lambda) y^2$

$\because 0 < \lambda < 1$

$\therefore \lambda^2 x^2 + 2\lambda(1-\lambda)xy + (1-\lambda^2)y^2 \leq \lambda x^2 + (1-\lambda)y^2$

$f(x) = x^2$ is a convex function

$f(x) = x^3$

$f(\lambda x + (1-\lambda)y) = (\lambda x + (1-\lambda)y)^3$

$\lambda f(x) + (1-\lambda) f(y) = \lambda x^3 + (1-\lambda) y^3$

let $x = -2$, $y = 1$, $\lambda = 0.5$

$(0.5 \times (-2) + (0.5) \cdot 1)^3 = -0.125$ $\quad \because -0.125 > -3.5$

$0.5 \cdot -8 + 0.5 \cdot 1 = -3.5$ $\quad \therefore f(x) = x^3$ is not a convex function

4.(b)

$$f(x) = x^T A x$$

$$f(\lambda x + (1-\lambda)y) = (\lambda x + (1-\lambda)y)^T A(\lambda x + (1-\lambda)y)$$

$$= \lambda^2 x^T A x + 2\lambda(1-\lambda)x^T A y + (1-\lambda)^2 y^T A y$$

$$\lambda f(x) + (1-\lambda)f(y) = \lambda x^T A x + (1-\lambda) y^T A y$$

$$(\lambda x^T A x + (1-\lambda) y^T A y) - (\lambda^2 x^T A x + 2\lambda(1-\lambda) x^T A y + (1-\lambda)^2 y^T A y)$$

$$= \lambda(1-\lambda) x^T A x - 2\lambda(1-\lambda) x^T A y + (1-\lambda-(1-\lambda)^2) y^T A y$$

$$= \lambda(1-\lambda)(x^T A x - 2x^T A y + y^T A y)$$

$$= \lambda(1-\lambda)(x-y)^T A(x-y)$$

$\because$ A is positive semi-definite matrix

$\therefore (x-y)^T A(x-y) \geq 0$

$\therefore (\lambda x + (1-\lambda)y)^T A(\lambda x + (1-\lambda)y) \leq \lambda x^T A x + (1-\lambda) y^T A y$

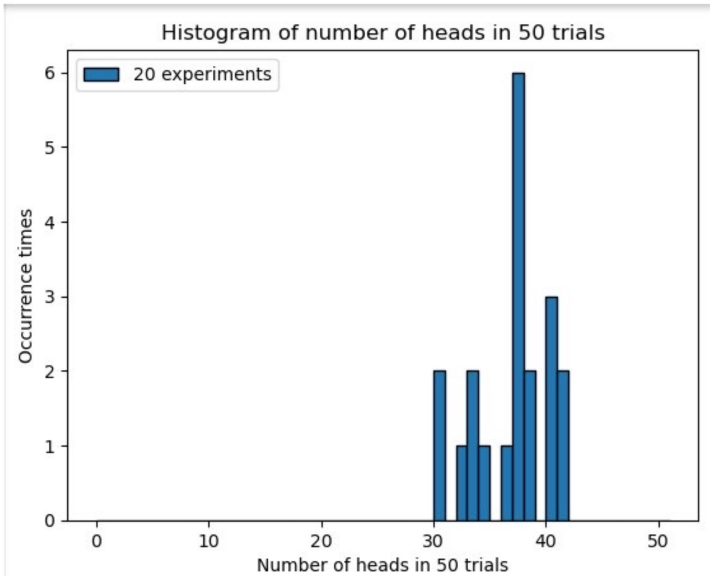
$\therefore$ A is convex function

5(a)

The total number of heads in 50 trials is 35.

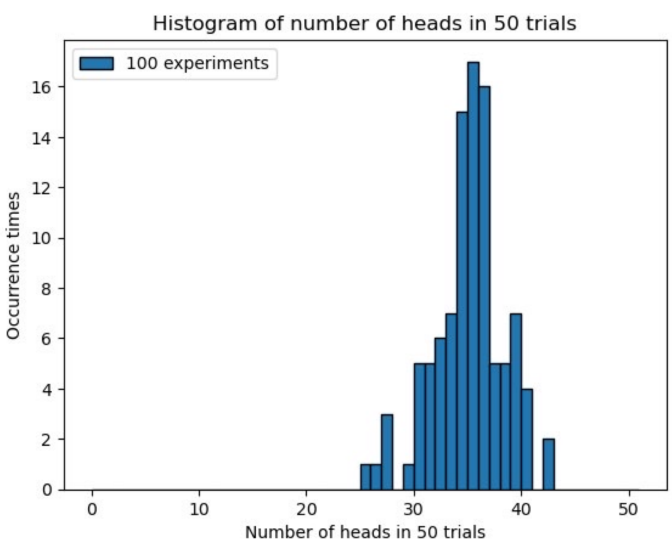
The longest run is 10 times of head.

(b) (i) 20 experiments



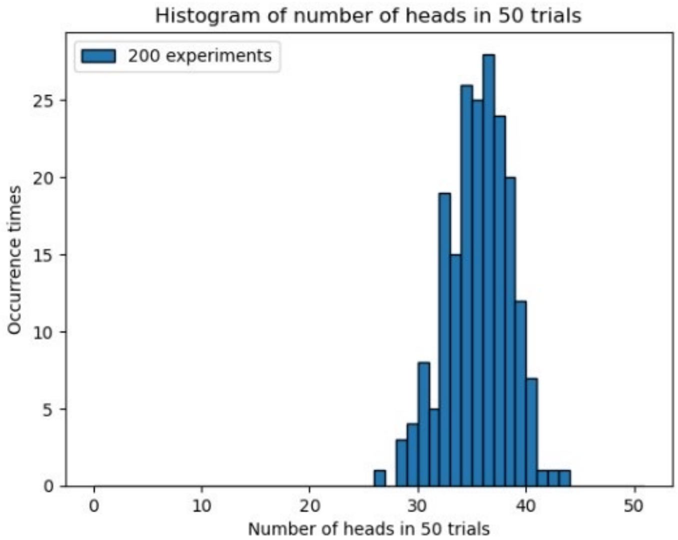
Even though the sample size is small, the histogram can show that number of heads fall between 30 and 40.

(ii) 100 experiments



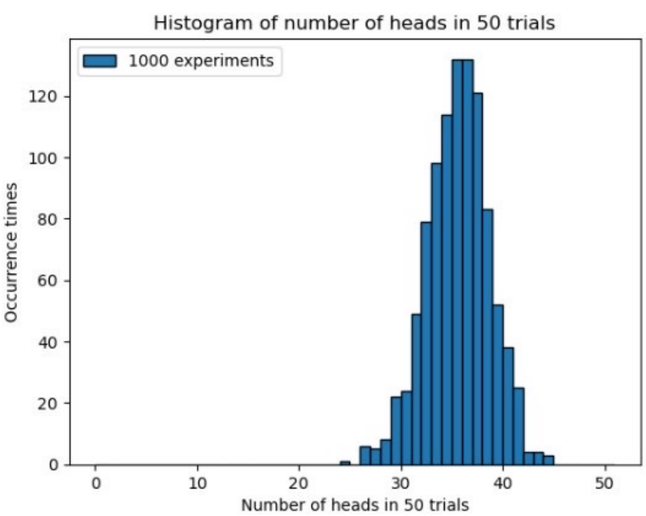
As the sample size increase, we can see that the most occurrence times is around 35.

(iii) 200 experiments



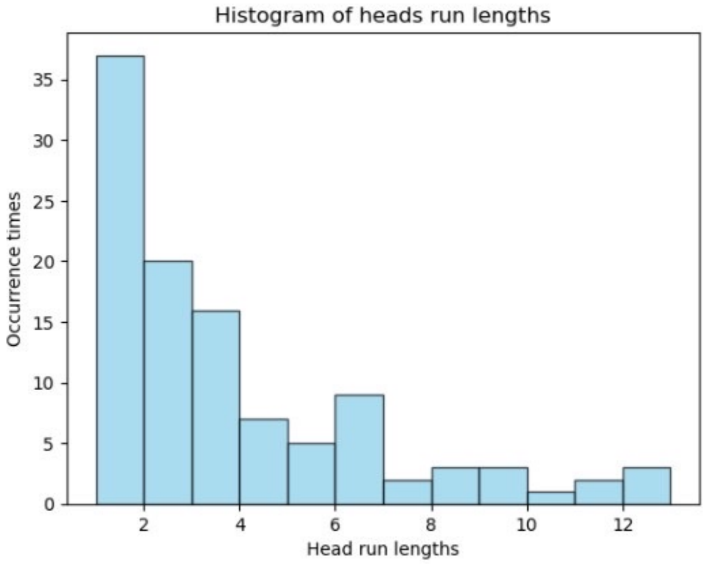
With a much larger sample size, the occurrence times become more concentrated between 30 and 40.

(iv) 1000 experiments



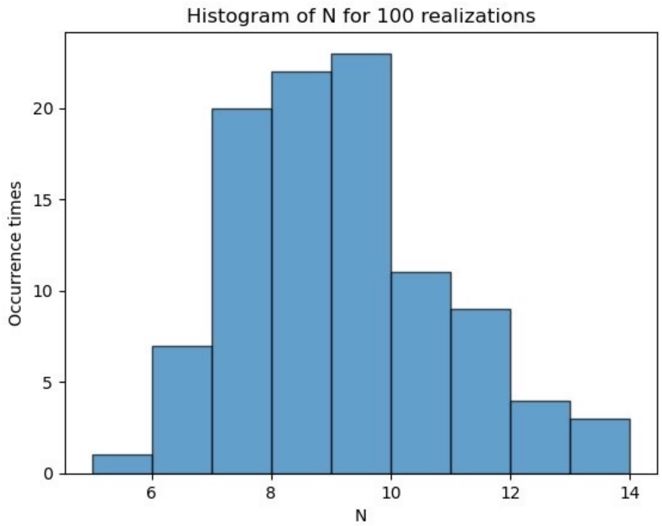
When the sample size reaches 1000, we can see that the occurrence times peak directly around 35.

5-(c)



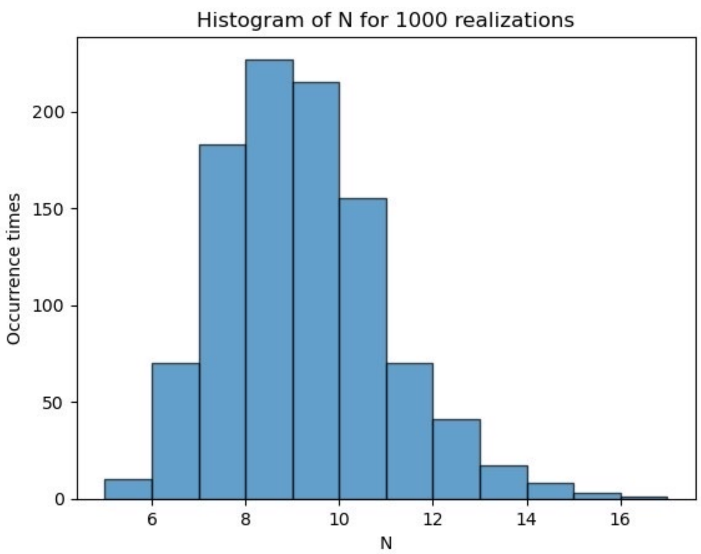
Because the probability of getting a head is 0.7, we can observe longer run lengths, with some run extending beyond 10.

6. (i) 100 realizations of N



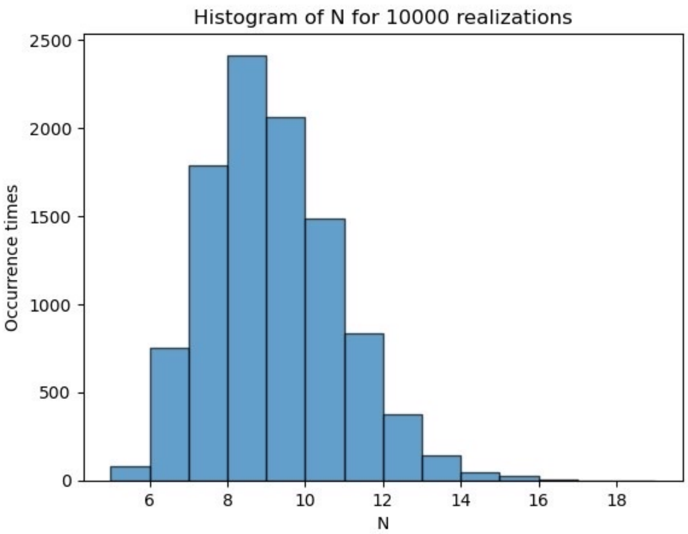
Because the expected value of each X_i is 0.5, we can see that N is typically between 6 to 10.

(ii) 1000 realizations of N



As the number of realization increases, we can see that N tends to cluster more closely to 8.

(iii) 10000 realizations of N



With 10000 realizations, the result become much more apparent, showing a clear concentration of N around 8.