6.036 Pset 1.

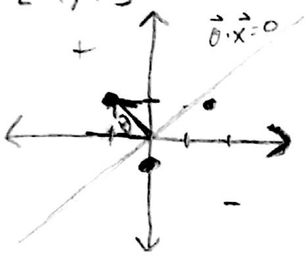
7a)(i) use Perceptron algorithm $i = \{1, 2, 3\}$

$i=1$ (first is always a mistake)

$\theta = \theta + y^{(1)} x^{(1)}$

$= [0,0] + (1)[-1,1]$

$= [-1,1]$



$\Rightarrow$ $i=2$

$h(x^{(2)}, \theta) = \text{sign}(\theta \cdot x^{(2)})$

$= \text{sign}((-1)(0) + (1)(-1))$

$= (-1) = y^{(2)}$ ✓

same as before

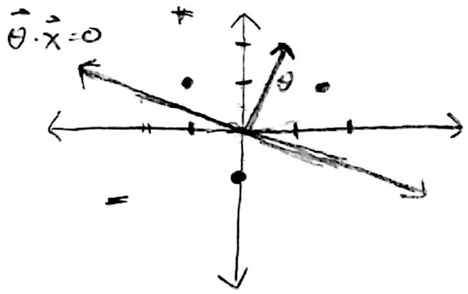
$i=3$

$h(x^{(3)}, \theta) = \text{sign}(-1(1.5) + (1)(1))$

$= \text{sign}(-.5) = -1 \neq y^{(3)} = 1$

$\Rightarrow \theta = [-1,1] + [1.5, 1](1)$

$= [0.5, 2]$



$\Rightarrow$ $i=1$

$h(x^{(1)}, \theta) = \text{sign}(\theta \cdot x^{(1)}) = \text{sign}((-1)(.5) + (1)(2))$

$= \text{sign}(-.5+2) = 1 = y^{(1)}$ ✓

same as before

$i=2$

$h(x^{(2)}, \theta) = \text{sign}((0)(.5) + (-1)(2)) = -1 = y^{(2)}$ ✓

Same as before

$\Rightarrow i \Rightarrow 3 \Rightarrow$ converges. ✓

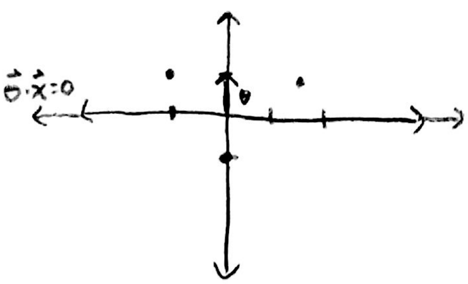
$\boxed{\text{Mistakes} = 2 \text{ before convergence.}}$

(ii) start with $x^{(2)}$

$i=2$ (first is mistake)

$\theta = [0,0] + [0,-1](-1)$

$= [0,1]$



$\Rightarrow$ $i=3$

$h(x^{(3)}, \theta) = \text{sign}(x^{(3)}\theta) = \text{sign}((1.5)(0) + (1)(1))$

$= \text{sign}(1) = 1 = y^{(3)}$ ✓

$i=1$

$h(x^{(1)}, \theta) = \text{sign}(x^{(1)}\theta) = \text{sign}((-1)(0) + (1)(1))$

$= \text{sign}(1) = 1 = y^{(1)}$ ✓

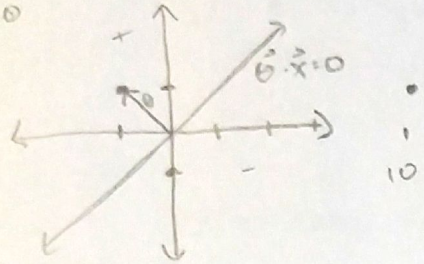
$\Rightarrow i=2 \Rightarrow$ converges

$\boxed{\text{Mistakes} = 1}$

b) $i=1$ (first is mistake)

$\theta = [0,0] + (1)[-1,1]$

$= [-1,1]$



$i=2$

$h(x^{(2)},\theta) = \text{sign}((-1)(0)+(1)(-1)) = \text{sign}(-1) = -1$

$\Rightarrow$ same figure

$\Downarrow$

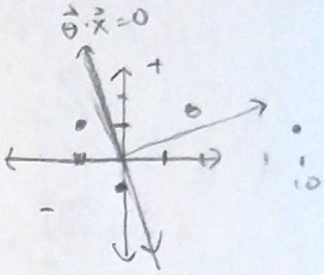
$i=3$

$h(x^{(3)},\theta) = \text{sign}((10)(-1)+(1)(1))$

$= \text{sign}(-9) \neq y^{(3)}$

$\theta = [-1,1] + (1)[10,1]$

$= [9,2]$



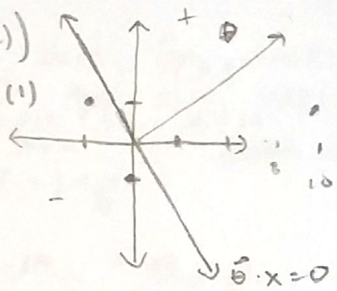
$i=1$

$h(x^{(1)},\theta) = \text{sign}((-1)(9)+(1)(2))$

$= \text{sign}(-7) = -1 \neq y^{(1)}$

$\theta = [9,2] + (1)[-1,1]$

$= [8,3]$



$i=2$

$h(x^{(2)},\theta) = \text{sign}(8(0)+3(-1)) = -1 = y^{(2)}$ ✓

$\Downarrow$

$i=3$

$h(x^{(3)},\theta) = \text{sign}((10)(8)+(1)(3)) = 1 = y^{(3)}$ ✓

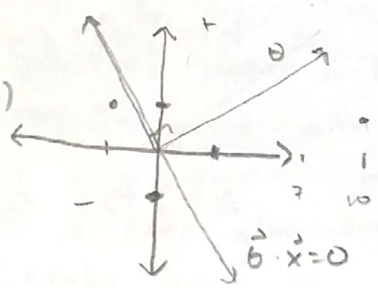
$i=1$

$h(x^{(1)},\theta) = \text{sign}((-1)(8)+(1)(3))$

$= \text{sign}(-5) = -1 \neq y^{(1)}$

$\theta = [8,3] + (1)[-1,1]$

$= [7,4]$



$\Rightarrow$ $i=2$ & $i=3$ always pass because

$[x_1, x_2] = \theta$ $\quad x_1$ & $x_2 > 0$

$\Rightarrow 0(x_1) + -x_2 = -x_2 < 0$

$\Rightarrow \text{sign}(-x_2) = -1$

& $\text{sign}(10x_1 + x_2) = 1$

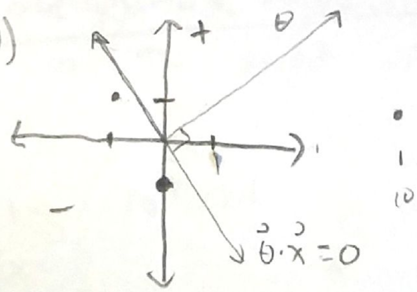
$\Rightarrow$ $i=1$

$h(x^{(1)},\theta) = \text{sign}((-1)(7)+(1)(4))$

$= \text{sign}(-3) = -1 \neq y^{(1)}$

$\theta = [7,4] + (1)[-1,1]$

$= [6,5]$



$\Rightarrow$ $i=2$ & $i=3$ ✓

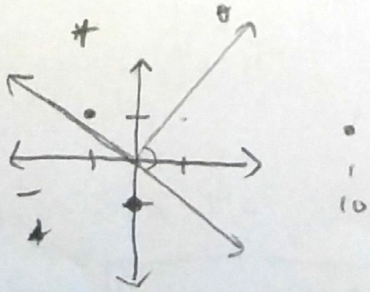
$\Rightarrow$ $i=1$

$h(x^{(1)},\theta) = \text{sign}((-1)(6)+(1)(5))$

$= \text{sign}(-1) = -1 \neq y^{(1)}$

$\theta = [6,5] + (1)[-1,1]$

$= [5,6]$



$\Rightarrow$ $i=2$ & $i=3$ ✓

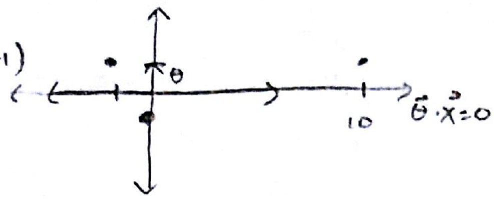
$i=1$

$\text{sign}((-1)(5)+(1)(6)) = 1$ ✓

$\boxed{\text{Mistakes} = 6}$

1(b) $i=2$

$\theta = [0,0] + [0,-1](-1)$
$= [0,1]$



$\Rightarrow$

$i=3$
$h(x^{(3)},\theta) = \text{sign}((10)(0)+(1)(1))$
$= \text{sign}(1) = 1 = y^{(3)}$ ✓

$i=1$
$h(x^{(1)},\theta) = \text{sign}((-1)(0)+(1)(1))$
$= \text{sign}(1) = 1 = y^{(1)}$ ✓

$\boxed{\text{Mistakes} = 1}$

c) The difference lies between the difference in value of the x-value in $x^{(3)}$. This cause small shifts in the graph of $\vec{\theta}$ until $y^{(0)} > x^{(0)}$ $\Rightarrow$ Greater the y-value of $x^{(3)}$ is $\Rightarrow$ more mistakes the graph will make when starting at $x^{(1)}$

The answer is the same for starting at $x^{(2)}$ because it gives a split divide of the two sides because they lie on either the positive y side or ~~point~~ negative-y.

d) Look to 1a) & b) as example., since we know there is an answer, we can sort of visualize the line. Thus, I would say start with the ~~po~~ side (positive, negative) ~~point~~ with more points. Once, we decide the side, choose the point with the smallest magnitude. Then choose the point with greatest magnitude that lies on the other side of $\vec{\theta}\cdot\vec{x}$ & keep alternating between the two opposites. This causes the most mistake because we want to put these polar opposites on the same side but it slowly increments because errors only rise on the smaller points & slowly changes θ after an error in the point with greatest magnitude.

2a) We start with the assumptions given in lecture.

A) exists an θ^* such that $y^{(i)}(\theta^* \cdot x^{(i)})/\|\theta^*\| \geq \gamma \ \forall\ i \in \{1,2, \ldots$ for some $\gamma > 0$

B) All examples are bounded $\|x^{(i)}\| \leq R$, $i > 0$

$\theta_0 = \vec{a}$

$$\cos(\theta^{(k)}, \theta^*) = \frac{\theta^{(k)} \cdot \theta^*}{\|\theta^{(k)}\| \|\theta^*\|}$$

$$\frac{\theta^{(k)} \cdot \theta^*}{\|\theta^*\|} = \frac{(\vec{a} + z(0) + z(1) \ldots z(k-1))\,\theta^*}{\|\theta^*\|}$$ where $z(n)$ = errors/changes

$$\Rightarrow \frac{\vec{a} \cdot \theta^*}{\|\theta^*\|} + \frac{\theta^{(k-1)} \cdot \theta^*}{\|\theta^*\|} + \frac{y^{(i)} x^{(i)} \cdot \theta^*}{\|\theta^*\|} \geq$$

$$\Rightarrow \frac{(a \cdot \theta^*)}{\|\theta^*\|} + \underbrace{\frac{\theta^{(k-1)} \theta^*}{\|\theta^*\|} + \gamma}_{\geq k\gamma} \geq C_1 + k\gamma$$ for some constant $C_1 = \left(\frac{a \cdot \theta^*}{\|\theta^*\|}\right)$

$$\|\theta^{(k)}\|^2 = \|\theta^{(k-1)}\|^2 + 2y^{(i)}\theta^{(k-1)} \cdot x^{(i)} + 2y^{(i)} a \cdot x^{(i)} + \|x^{(i)}\|^2$$

$$\leq \|\theta^{(k-1)}\|^2 + 2y^{(i)} a \cdot x^{(i)} + R^2$$

$$\leq \|a\|^2 + 2y^{(i)} a \cdot x^{(i)} + kR^2$$

$\to$ series of all error points $2y^{(i)} a \cdot x^{(i)}$

$$\cos(\theta^{(k)}, \theta^*) \geq \frac{C_1 + k\gamma}{\sqrt{C_2 + kR^2}}$$

$C_2 = \|a\|^2 + 2y^{(i)} a \cdot x^{(i)}$

$$\Rightarrow 1 \geq \frac{C_1 + k\gamma}{\sqrt{C_2 + kR^2}}$$

$$C_2 + kR^2 \geq C_1^2 + 2k\gamma C_1 + k^2\gamma^2$$

$$k^2\gamma^2 + k(2\gamma C_1 - R^2) + (C_1^2 - C_2) \leq 0$$

$\sqrt{b^2 - 4ac}$ as long as > 0 then good

$$\Rightarrow (2\gamma C_1 - R^2)^2 - 4\gamma^2(C_1^2 - C_2)$$

$$\cancel{4\gamma^2 C_1^2} - 4\gamma C_1 R^2 + R^4 - \cancel{4\gamma^2 C_1^2} + 4\gamma^2 C_2 > 0 \ ?$$

$$4\gamma(\gamma C_2 - C_2 R^2 + R^4) > 0 \checkmark$$

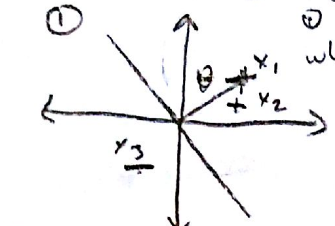
$\Rightarrow$ solution is bounded

previous

2b) No, they can give a different θ value.

Consider this example

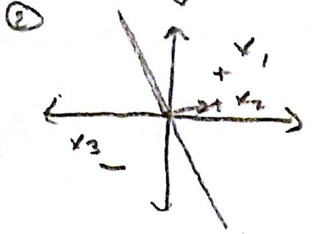
(1)



(1) where $\theta = x_1$ ~~&~~ start ~~at~~ x_2

(2) when $\theta = 0$ & starts at x_2

(2)



$\Rightarrow$ 2 different final θ-value. but same training & same order.

2c) No, because the final θ-value can be different $\Rightarrow$ a test set that works for (1) doesn't have to work for (2)

d) $\theta^{(final)} = \theta^{(0)} + \text{errors} = \quad \text{errors} = (\#\text{ misclassified})(y)(x^{(i)})$

$= [0,0] + (1)(1)([-3,2]) + (2)(-1)([-1,-1]) + (1)(-1)([2,2])$

$= [-3,2] + [2,2] + [-2,-2]$

$\boxed{\theta = [-3,2]}$

3a) And table

x_1	x_2	x_3	$f(x_1,x_2,x_3)$
0	0	0	0
0	0	1	0
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	0
1	1	1	1

1) if $\theta_0 = 0$ & we have

$\theta \cdot x > 0$ when $f(x) = 1$

$\theta \cdot x < 0$ otherwise.

$\Rightarrow$ no because if $\theta = [a,b,c]$

$\Rightarrow a+b+c > 0$ but

$a < 0$, $b < 0$, & $c < 0$

$\rightarrow$ it's impossible.

ii) going on the last ~~ex~~ example.

$(a+b+c)+[\theta_0]>0$

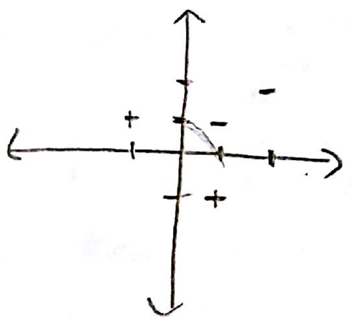
$\Rightarrow$ but $0+[\theta_0]<0$

$[\theta_0]<0$

$\Rightarrow$ $a<-\theta_0$ $b<-\theta_0$ $c<-\theta_0$

$\Rightarrow$ yes because we need to find a, b, c such that $a+b+c>-\theta_0$ but each $a, b, c, a+b, a+c, b+c < -\theta_0$

b)



i) not possible because $(-1,1)$, $(1,-1)$ & $(1,1)$ are all equidistant from $(0,0)$.

ii) possible,

$(x+3)^2+(y+3)^2 \leq 25$

iii) not possible because $(-1,1)$ & $(1,-1)$ lie on the same line through the origin.

$\Rightarrow$ adjusting θ means one point will be wrong & other correct.

iv) possible

$y=-x+1$

$\theta=[1,1]$

$\theta_0=1$

4a) $A=\begin{bmatrix} \frac{1}{6} & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{3} \\ \frac{1}{6} & \frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} \\ \frac{1}{6} & -\frac{1}{3} \end{bmatrix}$

c) $\theta=[1,1,1,1]$ $\rightarrow$ 2 dimensions.

points $[-1,1,1,1]$ - +

$[1,-1,1,1]$ - +

$[1,1,-3,-2]$ $\rightarrow$ -

$[-1,-1,0,0]$ $\rightarrow$ -

but depending on A, $\exists$ a $h(x)$ that is good. if $A=[1,1]$ $\Rightarrow$ $\forall$ no $h(x)$ that is good.

c) on previous page

4b) What we're doing essentially is that we're taking a classified set in a smaller dimension & making it to a larger dimension. If we are transforming to a larger dimension, there is a ~~log~~ larger set of posibilities of θ to get a classifier from an already existing classified set. ⇒ There should be a classifier. ~~sog~~ sign$(\theta_x$

a) If we're working in a lower dimension, we have less of a choice & thus more prone to error ~~bec~~ & mistakes. Thus, it would be slower for the smaller dimension to converge. For a higher dimension, we have much more possibilities for a linear convergence to happen.

e) Depending on the data, but in most likely considerations, we would find a much more accurate classifier in a higher dimension with more possibilities on the seen training data since we know where the classifier should typically go.

However on unseen training data, the higher number of possibilities causes us to have more doubt & less accuracy on our answer.