1(a) $d=2=n=i$

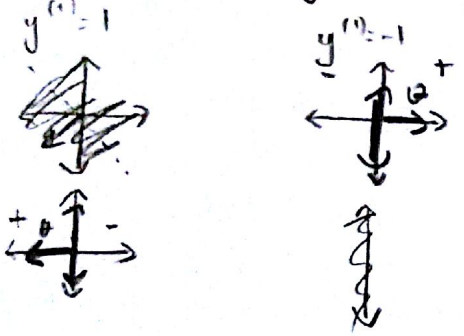
$i=2$

$x^{(1)} = [\cos(\pi), 0] = [-1, 0]$

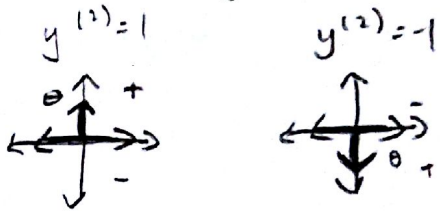
$x^{(2)} = [0, \cos(2\pi)] = [0, 1]$

First will always be mistake

$x^{(1)} \Rightarrow \theta = y^{(1)}[-1, 0]$



$x^{(2)} \Rightarrow \theta = y^{(2)}[0, 1]$



$x^{(2)}$ -> mistake.

$\Rightarrow \theta = [y^{(1)}(-1), y^{(2)}(1)]$

Same θ $\Longleftrightarrow$ $x^{(1)} \rightarrow \theta = [y^{(1)}(-1), y^{(2)}(1)]$

We know => no matter what we start with, we end up with same θ & 2 corrections.

b) for every $x^{(i)}$, $x_i^{(i)} = \cos(i\pi) = \{-1, 1\}$ $x_j^{(i)} = 0$

=> only have one non-zero term per $x^{(i)}$. => first term is always mistake & $\theta = y^{(i)} x^{(i)}$. Because, only one term θ_i switches, when we choose any $x^{(i)}$ after, it will always lie on the linear seperator => will always we zero. => Because other than $x^{(i)}$, the x_i term of all other $x^{(i)}$'s are zero

=> $\theta x^{(i)} = 0$ all the time => $\theta = y^{(1)}x^{(1)} + y^{(2)}x^{(2)} + y^{(3)}x^{(3)} + \ldots$

$y^{(d)}x^{(d)}$ => d corrections till convergence.

c) $\theta = [y^{(1)}x_1^{(1)}, y^{(2)}x_2^{(2)}, \ldots, y^{(d)}x_d^{(d)}]$

$= [-y^{(1)}, y^{(2)}, -y^{(3)}, \ldots, (-1)^d y^{(d)}]$

θ doesn't depend on ordering but does depend on labeling

d) $\|(x)^{(t)}\| \leq R$

$\|x^{(t)}\| = 1$

$1 \leq R$

$1 \leq R^2$

$$\frac{y^{(i)}(\theta^* \cdot x^{(i)})}{\|\theta^*\|} \geq \gamma$$

$\theta^* = [-y^{(1)}, y^{(2)}, -y^{(3)}, \ldots -1^d y^{(d)}]$ · $x^{(i)} = (-1^i y^i)(-1^i)$

$= y^i$

$\|\theta^*\| = \sqrt{d}$

$$\rightarrow \frac{[y^{(i)}]^2}{\sqrt{d}} \geq \gamma$$

$y^{(i)} = \{-1, +1\}$

$(y^{(i)})^2 = 1$

$$\frac{1}{\sqrt{d}} \geq \gamma$$

$$\gamma^2 \leq \frac{1}{d}$$

$d \geq \frac{1}{\gamma^2}$ $R^2 \geq 1$

$\frac{1}{\gamma^2} = d$ $R^2 = 1$

$\frac{R^2}{\gamma^2} = d$ ✓

2 a) $\frac{\lambda}{2}\|\theta - \theta^{(k)}\|^2 + Loss_h(y\theta \cdot x)$ $\theta^{(k+1)} = \theta^{(k)} + \eta yx$

$\|\theta - \theta^{(k)}\|^2$ is how close θ is to the current $\theta^{(k)}$

⇒ Minimize $\frac{\lambda}{2}\|\theta - \theta^{(k)}\|^2$, if $\frac{\lambda}{2} \uparrow$, ⇒ we want less $\|\theta - \theta^{(k)}\|^2$. This means that a higher learning rate leads our $\theta^{(k)}$ to become more closer/ accurate to θ.

if λ is large, then learning rate is high. ⇒ $\eta = \min\left\{\frac{Loss_h(y\theta^{(k)}, x)}{\|x\|^2}, \frac{1}{\lambda}\right\}$

⇒ if $\lambda \uparrow$ $\frac{1}{\lambda}$ η = small because it R_s cannot be greater than.

2b) 1) d) ⇒ larger λ means θ changes less

2) b) ⇒ 0-1 loss changes θ more to point misclassified is now classified.

3) a)

4) c)

3c) 1) hinge loss & large λ means there is almost no or no change in θ ⇒ d

2) hinge loss means very little to no change, but small λ means θ will change a bit more than large λ ⇒ b

3) 0-1 loss causes misclassified to be classified ⇒ either a or c but larger λ is less change to θ ⇒ a)

4) last on causes most change θ ⇒ c

3a) $\text{loss}_h = 1 - y\theta^{(k)} \cdot x$

$$\frac{\lambda}{2}\|\theta - \theta^{(k)}\|^2 + (1 - y\theta \cdot x) > 0$$

$$\frac{\lambda}{2}\|\theta - \theta^{(k)}\|^2 > y\theta \cdot x - 1$$

$$\|\theta - \theta^{(k)}\|^2 > \frac{2y\theta \cdot x - 2}{\lambda}$$

$$\|\theta\|^2 - 2\theta \cdot \theta^{(k)} + \|\theta^{(k)}\|^2 > \frac{2}{\lambda}(y\theta \cdot x - 1)$$

$$\|\theta^2\| - 2\|\theta\|\|\theta^{(k)}\| \cos\theta + \|\theta^{(k)}\|^2 > \frac{2}{\lambda}(y\theta \cdot x - 1)$$

$$\|\theta\|^k (\|\theta\|^k - 2\|\theta\|) > \frac{2}{\lambda}(y\theta \cdot x - 1)$$

a) $\|\theta^{k}\|^2 - 2\|\theta\|\|\theta^{k}\| - \frac{2}{\lambda}(y\theta \cdot x - 1) > 0$

$$\|\theta^{k}\| = \frac{2\|\theta\| \pm \sqrt{4\|\theta\|^2 + 4|y\|\theta\|\|x\| - 1)/\lambda}}{2}$$

$$= 4\|\theta\| \pm \sqrt{4\|\theta\|(\|\theta\| + 4y\|x\| - 1)/\lambda}$$

$$\|\theta^{k}\| = \|\theta\| \pm \sqrt{\|\theta\|^2 + \frac{(4yx \cdot \theta - 1)}{\lambda}}$$

Not sure how to continue & what not.

b) $\frac{\lambda}{2}\|\theta - \theta^{(k)}\|^2 + Loss_n(y\theta \cdot x)$

$$\theta^{(k+1)} = \theta^{(k)} + \eta yx$$

$$\eta = \min\left\{\frac{Loss}{\|x^{(i)}\|^2}, \frac{1}{\lambda}\right\} = \min\left\{\Delta loss, \frac{1}{\lambda}\right\}$$

Yes, because, even w/ updates diffent, only one $x^{(i)}$ has their specific terms.