Question 1 : Perception Questions

1. Let the sequence of points be $x_i = ((-0.5)^i)$ where i ranges from 1 to ∞.
Classification of points $y_i = \text{sign}(x_i)$ or $(-1)^i$ where i ranges from 1 to ∞.

<u>Property 1</u>: Sequence is strictly linearly seperable with $b=0$

Points with +1 classification : $((0.5)^2, 0.5), (0.5^4, 0.5)$
$\ldots ((0.5)^{2i}, 0.5)$

Points with -1 classification : $(-0.5, 0.5), (-0.5^3, 0.5)$
$\ldots ((-0.5)^{2n+1}, +0.5)$

$\therefore$ All points with +1 classification is on the right side of the y axis or $x=0$ line and tends to 0 but will never become 0.
All points with -1 Classification is on the left side of the y-axis or $x=0$ line and sequence is strictly linearly seperable.

<u>Property 2</u> : $\max \|x_i\|_2 \leq 1$

$$\max \|x_i\|_2 = \sqrt{(-0.5)^2 + (0.5)^2}$$
$$= 0.71 \leq 1$$

$\therefore$ Proved

<u>Property 3</u> : Modified perceptron algorithm makes infinite mistakes

Let us consider 1 point : $((-0.5, 0.5), -1)$

$$y(wx+b) < 0$$
$$0 < 0$$

$\therefore$ mistake

update $w = (0.5, -0.5)$ $b = -1$

Point 2 : ((0.25, 0.5), 1)

$y(wx+b) = (0.25, 0.5)(0.5, -0.5) - 1$

$= 0.125 - 0.25 - 1 < 0$

$\therefore$ ~~Correct~~ mistake

New algorithm $\Rightarrow$ Update $w = (0.75, 0)$

$b = 0$

Point 3: ((−0.125, 0.5), −1)

$y(wx+b) = -1((0.75, 0)(-0.125, 0.5))$

$= 0.75 \times 0.125 > 0$

$\therefore$ ~~mistake~~ Correct → would be correct from now

$w = (0.75, 0) + (0.125, -0.5)$ if no update

$= (0.875, -0.5)$

$b = -1$

Point 4: ((0.0625, 0.5), 1)

$y(wx+b) = (0.875, -0.5)(0.0625, 0.5) - 1$

$= 0.055 - 0.25 - 1 < 0$

$\therefore$ ~~True~~ mistake ~~→ No mistakes if no update~~

$w = (0.875, -0.5) + (0.0625, +0.5)$

$= (0.94, 0)$ $b = 0$

For point 5, $b = 0$ & $y = -1$ so $wx < 0$

$\therefore ywx > 0$ and ~~mistake~~ correct

$\therefore$ This trend will go on infinitely resulting in infinite & mistake since b will shift between 0 and −1 resulting in alternating ~~trues~~ corrects and mistakes.

2. Give examples where perceptron converges to arbitrarily small margin halfspace and a seperate example where it converges to maximum.

For an arbitrarily small margin halfspace, we can use ~~the~~ part of the dataset from (1) with some changes
Let dataset be $-1 \to ((-0.5, 0), ((-0.125), 0), ((-0.03125), 0)$
$+1 \to ((0.25, 0), (0.0625, 0), (0.016, 0)$
Since perceptron treats this similar to the case in Q1.
Since this is the original algorithm, it will not make infinite mistakes. Instead since points of opposing classes move closer to the $x = 0$ line, the only possible margin halfspace will be arbitrarily small since decision boundary will be close to $x = 0$ allowing only little changes resulting in a margin of $0 < \epsilon < \frac{1}{2}$

For a maximized margin halfspace,
Let us use dataset, $+1 \to (1, 0), (2, 0), (3, 0)$
$-1 \to (-1, 0), (-2, 0), (-3, 0)$

Here alternating points of the 2 classes ~~can~~ maximizes the margin halfspace.
Let us consider the perceptron algorithm in this case
$w = \langle 0, 0 \rangle \quad b = 0$

Point 1: $((1, 0), 1)$
$$y(wx + b) \Rightarrow 1(0 + 0) = 0 \leq 0$$
$\therefore$ update $w = \langle 1, 0 \rangle$
$b = 1$

Point 2: $((-1, 0), -1)$
$$y(wx + b) \Rightarrow -1((1, 0)(-1, 0) + 1)$$
$$-1(1) \leq 0$$

update $w = (1,0)+(-1)(-1,0) = (2,0)$
$b = 1-1 = 0$

Point 3: $((2,0),1)$
$y(wx+b) = (1((2,0)(2,0)+0) = 4) > 0$

Point 4: $((-2,0),-1)$
$-1((2,0)(-2,0)) = 4 > 0$

Point 5: $((3,0),1)$
$1((2,0)(3,0)) = 6 > 0$

Point 6: $((-3,0),-1)$
$-1((2,0)(-3,0)) = 6 > 0$

$\therefore$ $w=(2,0)$ $b=0$ maximizes the perceptron margin halfspace for given data while $b=0$ passes through origin, first data points margin is 1.

3. Show how perceptron can be viewed as an instantiation of SGD

Perceptron is a specific instance of SGD where w and b updates seem like changes in gradient with iterations

Let us consider an arbitrary data point: (x_1, x_2)
If we were to consider the perceptron algorithm, it would look like: if $y(w^T x + b) \leq \delta$ then

$$w \leftarrow w + yx$$
$$b \leftarrow b + y$$

This is what stochastic gradient needs to look like

The best loss function in this case for binary classification would be ~~hinge~~ this loss function which will be defined later.

$\cancel{\ell(f(x)) = \max(\delta, f(x))}$
$\cancel{= \max(\delta, y(w^T x + b))}$

Now, let us consider the SGD update

$$w_{new} = w_{old} + \eta \nabla L_w$$
$$b_{new} = b_{old} + \eta \nabla L_b$$

$$\cancel{L = \max(0, -y(wx+b))}$$

$$\cancel{\text{If } y(wx+b)\ \ w_{new} = w_{old} + \frac{\partial}{\partial w}(0)}$$

$$\cancel{L = \max(\delta, y(wx+b))}$$

$$\cancel{\text{If } y(wx+b) > \delta, \quad w_{new} = w_{old} + \frac{\partial}{\partial w}(\delta}$$

Loss function $L = \min(y(wx+b), \delta)$

step rate $\eta = 1$

If $y(wx+b) > \delta$

$$w_{new} = w_{old} + 1 \frac{\partial \delta}{\partial w} = w_{old}$$
$$b_{new} = b_{old} + 1 \frac{\partial \delta}{\partial b} = b_{old}$$

If $y(w_x+b) \leq \delta$

$$w_{new} = w_{old} + 1 \cdot \frac{\partial}{\partial w} y(wx+b) = w_{old} + yx$$
$$b_{new} = b_{old} + 1 \frac{\partial}{\partial b} y(wx+b) = b_{old} + y$$

This is the same as perceptron algorithm

$\therefore$ Proved.