Exercise 2: Perceptron Questions (5 pts)

1. (3 pts) The perceptron algorithm makes an update every time it witnesses a mistake. What if it makes an update on every point, regardless of whether or not that point is correctly classified? For simplicity, consider a setting where where b is fixed to 0. Give an example of an infinite sequence of points (x_i, y_i) with the following properties:
   I The sequence is strictly linearly separable with $b = 0$ (i.e., the margin is some constant $\gamma > 0$) ,
   II The sequence has max $\|x_i\|_2 \leq 1$,
   III The modified perceptron algorithm makes an infinite number of mistakes on this sequence .
   Prove that it has these properties. Note that the perceptron convergence theorem and the first two conditions imply that, at some point, the unmodified perceptron algorithm would only make a finite number of mistakes.
   Ans: **Answer 2.1**
2. (1 pt) Give examples of where the perceptron algorithm converges to a 0 margin halfspace, and a separate example where it converges to a maximum margin halfspace. As pointed out by some on Piazza, tech- nically, if a halfspace has 0 margin, then it would misclassify anything that still lies on the hyperplane, and the perceptron algorithm would not yet have halted. If you came up with an example that ignores these cases and halts with a point on the hyperplane, that's fine. However, the intent was more like the following, so consider solving the problem where it converges to an arbitrarily small margin halfspace. More precisely: for any $0 < \varepsilon < 1/2$, give a dataset (with margin at least 1) and a order in which to process the points such that the perceptron algorithm halts providing a halfspace with margin $\leq \varepsilon$. This problem has to do with the original perceptron, not the modified perceptron from part 1.
   Ans: **Answer 2.2**
3. (1 pt) Suppose that in each iteration of the perceptron algorithm, a point is chosen uniformly at random from the dataset. Show how the perceptron algorithm can be viewed as an instantiation of stochastic gradient descent (SGD). In particular, you must provide a loss function and a learning rate such that SGD with these parameters and perceptron are identical.
   Ans: **Answer 2.3**

Answer 2.1: Infinite Mistakes

From the Property III and Property I, we may conclude the following:

$$\because \quad \text{mistake } \forall i, \text{ and set threshold } (\delta = 0) \tag{1}$$

$$\therefore \quad y_i(\langle \mathbf{x}_i, \mathbf{w}_{i-1}\rangle + b) \leq 0, \forall i \in [1, \infty) \tag{2}$$

$$\Rightarrow \mathbf{w}_i = \mathbf{w}_{i-1} + y_i \mathbf{x}_i, \forall i \in [1, \infty) \tag{3}$$

$$\therefore \quad \mathbf{w}_k = \mathbf{w}_0 + y_1 \mathbf{x}_1 + \cdots + y_k \mathbf{x}_k, \forall k \in [1, \infty) \tag{4}$$

$$\mathbf{w}_k = \mathbf{w}_0 + \sum_{i=1}^{k} y_i \mathbf{x}_i, \forall k \in [1, \infty) \tag{5}$$

$$\because \quad \text{For simplicity, } b = 0 \tag{6}$$

$$\therefore \quad b_i = 0, \forall i \tag{7}$$

$$\because \quad \text{By convention, } \mathbf{w}_0 = \mathbf{0} \text{ will make the first always as mistake} \tag{8}$$

$$\therefore \quad w_k = \sum_{i=1}^{k} y_i \mathbf{x}_i, \forall k \in [1, \infty) \tag{9}$$

$$y_{k+1}\langle \mathbf{x}_{k+1}, \mathbf{w}_k\rangle \leq 0 \forall k \in [1, \infty) \tag{10}$$

$$\therefore \quad \langle y_{k+1}\, \mathbf{x}_{k+1}, \mathbf{w}_k\rangle \leq 0 \tag{11}$$

$$\langle y_{k+1}\, \mathbf{x}_{k+1}, \sum_{i=1}^{k} y_i \mathbf{x}_i\rangle \leq 0 \tag{12}$$

$$\because \quad \text{Let } \mathbf{a}_i = y_i\, \mathbf{x}_i, \forall i \in [1, \infty) \tag{13}$$

$$\therefore \quad \langle \mathbf{a}_{k+1}, \sum_{i=1}^{k} \mathbf{a}_i\rangle \leq 0 \Rightarrow \sum_{j=1}^{d} \left(a_{k+1,j} \sum_{i=1}^{k} a_{i,j} \right) \leq 0 \Rightarrow \sum_{j=1}^{d} \sum_{i=1}^{k} a_{k+1,j}\, a_{i,j} \leq 0 \tag{14}$$

Hence:

$$\therefore \quad \text{if} \quad \mathscr{D} = \left\{ (\mathbf{x}_i, y_i) : \sum_{j=1}^{d} \sum_{i=1}^{k} a_{k+1,j}\, a_{i,j} \leq 0, \mathbf{a}_i = y_i\, \mathbf{x}_i, i \in [1, \infty) \right\} \tag{15}$$

$$\rightarrow \textbf{Make mistakes on infinity many points} \tag{16}$$

From the Property II, we may add an additional constraint on the example:

$$\sum_{j=1}^{d} \mathbf{x}_{ij}^2 \leq 1 \Rightarrow \sum_{j=1}^{d} a_{ij}^2 \leq 1 \tag{17}$$

Hence, in order to satisfy all three properties, the dataset would be:

$$\mathscr{D} = \left\{ (\mathbf{x}_i, y_i) : \sum_{j=1}^{d} \sum_{i=1}^{k} a_{k+1,j}\, a_{i,j} \leq 0, \sum_{j=1}^{d} a_{i,j}^2 \leq 1, \mathbf{a}_i = y_i\, \mathbf{x}_i, i \in [1, \infty), y_i \in \{1, -1\} \right\} \tag{18}$$

Alert 2.1

here are infinity many examples for this question. Here, we will mathematically derive one with assumptions.

For simplicity, let's consider a 2D case (Assumption II), which allows us to visualize geometrically, since $w_i^T \mathbf{a}_i = |w_i||\mathbf{a}_i| cos(\theta_{w_i, a_i})$. The thought process here is that if we make all labels as $+1$ (Assumption III), then $a_i = x_i$ and $w_{i+1} = w_i + x_i$, which will simplifies many things. In order to ensure the linearly separable property, we can assume all the points are on the right half of the plane and may only include one side of the vertical axis (Assumption IV).

Let's summarize up these assumptions:

I Property II $\Rightarrow \|x_i\|_2 \le 1$
II 2-D data: $d = 2$
III $y_i = 1 \forall i$
IV All the points are on the right hand plane: $a_{i,1} \ge 0$

Hence, from Assumption II, we may simplify Equation (14) to:

$$\sum_{j=1}^{d=2}\sum_{i=1}^{k} a_{k+1,j}\, a_{i,j} \le 0 \tag{19}$$

$$a_{k+1,1}\sum_{i=1}^{k} a_{i,1} + a_{k+1,2}\sum_{i=1}^{k} a_{i,2} \le 0 \tag{20}$$

$$a_{k+1,1}\sum_{i=1}^{k} a_{i,1} \le -a_{k+1,2}\sum_{i=1}^{k} a_{i,2} \tag{21}$$

Expanding per $(k+1)$:

$$k+1=1 \quad 0 \le 0 \tag{22}$$

$$k+1=2 \quad a_{2,1}\left(a_{1,1}\right) \le -a_{2,2}\left(a_{1,2}\right) \tag{23}$$

$$k+1=3 \quad a_{3,1}\left(a_{1,1}+a_{2,1}\right) \le -a_{3,2}\left(a_{1,2}+a_{2,2}\right) \tag{24}$$

$$k+1=4 \quad a_{4,1}\left(a_{1,1}+a_{2,1}+a_{3,1}\right) \le -a_{4,2}\left(a_{1,2}+a_{2,2}+a_{3,2}\right) \tag{25}$$

$$\vdots \tag{26}$$

From this inequality, we can see the solution would be alternating in quadrant 2 and 4.

In addition,

$$\because \quad w_i^T \mathbf{a}_i = |w_i||\mathbf{a}_i|cos(\theta_{w_i,a_i}) \quad w_i^T \mathbf{a}_i \le 0 \tag{27}$$

$$\therefore \quad \pi \ge \theta_{w_i,a_i} \ge \frac{\pi}{2}, \forall \theta_{w_i,a_i} \in [0,\pi] \tag{28}$$

This would ensure all points locate on one side of the plane, and makes mistakes.

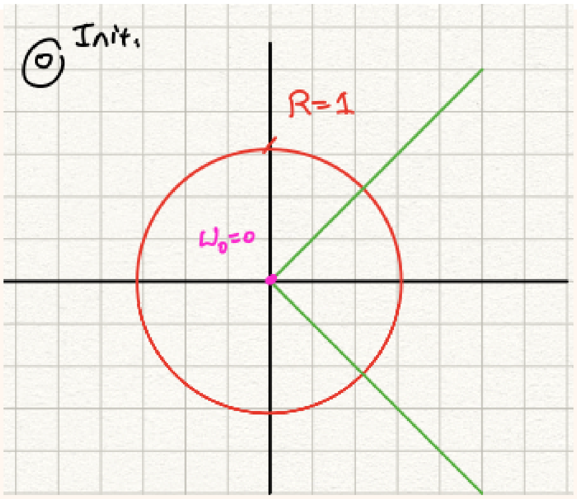
Let's draw out the concept in 2D:

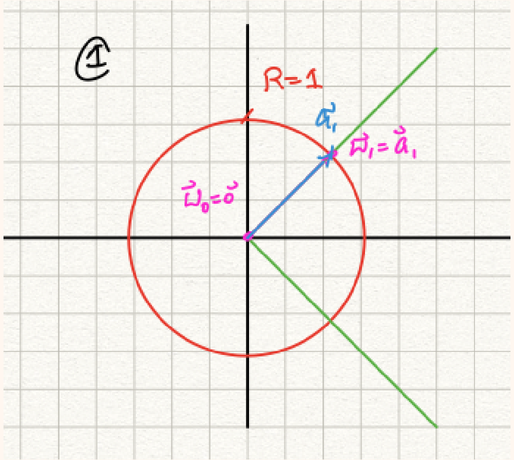
Initially, w_0 starts from the origin. Updating $w_{i+1} = w_i + \mathbf{a}_i$ essentially means w_i vector keeps accumulates the $\mathbf{a}_i$. If the w_i vector forms an angle larger than or equal to 90 degree, it guarantee it makes a mistake. Property III indicates an infinity mistakes, but permits some mistakes in between. Hence, the goal is to make this aggressive updating method make mistakes and never converges.

Say if we assume there exists repetitive series of points, that ensures the last point return back to the starting axis, but also guarantee makes an angle larger or equal than 0 all the time between the last point and the prior weight before updating.

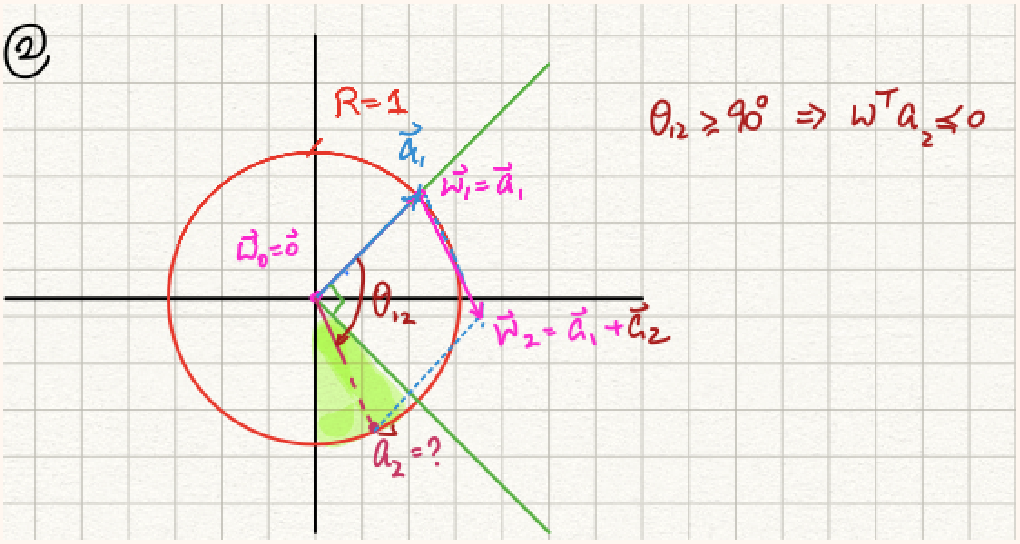
As a result, we shall try to make a set of three vectors that can bring the $w_{3n}, n \in \mathbb{Z}$ back to the x-axis ($w_{3n,2} = 0$), and we want $w_{3n-1,2}$ and $\mathbf{a}_{3n}$ form an angle larger or equal than zero ($\theta_{3n-1\to 3n} = 0$)

For simplicity, let's start with a 45 degree angle ray sourced from the origin and intersect with R (L2 norm boundary), where, we obtain first point $\mathbf{a}_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right)$.



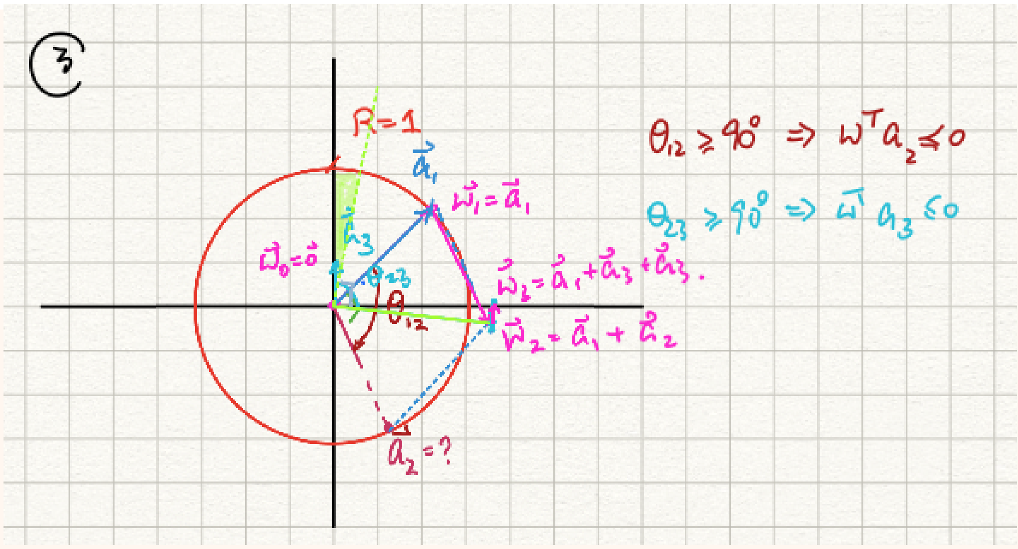


Hence, the second point has to be in the green region in quadrant 2, so that it forms an angle larger than 90 degree. In addition, we want the updated w_2 point fall in quadrant 4, so that it can be pulled back to x-axis with the third point in quadrant 2 ($a_{22}+a_{12}<0$). For simplicity, we can assume a point that forms a 60 degree from the y axis with an l2 norm of 1. (You may pick any points from this region.) As a result, we compute $\mathbf{a}_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})$.



2
R=1
$\theta_{12} \geq 90^\circ \Rightarrow w^T a_2 \leq 0$
$\vec{a}_1$
$\vec{w}_1=\vec{a}_1$
$\vec{w}_0=\vec{o}$
θ_{12}
$\vec{w}_2=\vec{a}_1+\vec{a}_2$
$a_2=?$

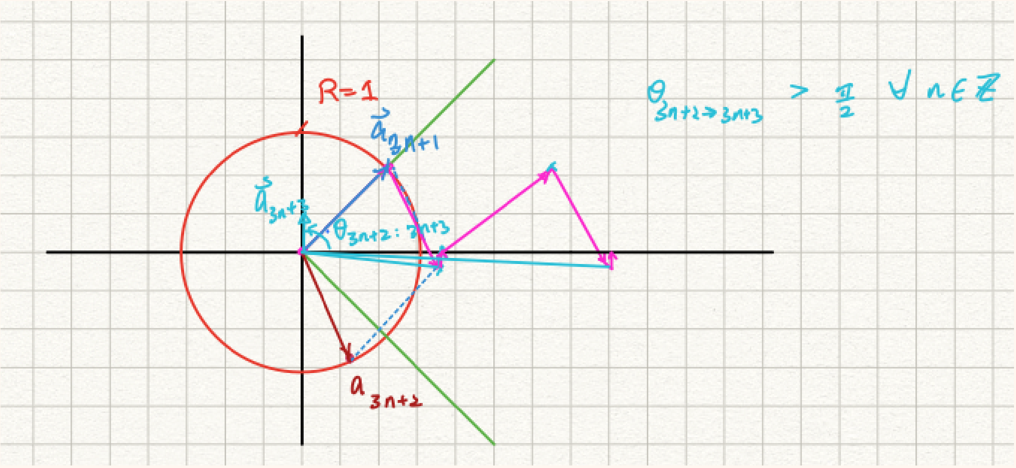
Now, we shall pull the weight vector back to x-axis. Hence, $a_{12}+a_{22}+a_{32}=0$, resulting $a_{32}=\frac{\sqrt{3}}{2}-\frac{1}{\sqrt{2}}$. Most important, we need to make sure this vector $\mathbf{a}_3$ would form an angle larger than 90 degree. Since the weight vector w_2 will always fall below the x-axis, hence, we shall make every third point locates on positive y-axis ($a_{31}=0$), so that it guarantees a minimum of 90 degree as $i\to\infty$. Hence, $\mathbf{a}_3=(0,\frac{\sqrt{3}}{2}-\frac{1}{\sqrt{2}})$.


3
R=1
$\vec{a}_1$
$\vec{w}_1 = \vec{a}_1$
$\theta_{12} \geq 90^\circ \Rightarrow w^T a_2 \leq 0$
$\theta_{23} \geq 90^\circ \Rightarrow \vec{w}^T a_3 \leq 0$
$\vec{a}_3$
$\vec{w}_0 = \vec{0}$
θ_{23}
θ_{12}
$\vec{w}_3 = \vec{a}_1 + \vec{a}_3 + \vec{a}_3$.
$\vec{w}_2 = \vec{a}_1 + \vec{a}_2$
$a_2 = ?$

Recall $\mathbf{x}_i = \mathbf{a}_i, y_i = 1$, therefore, we have obtained the infinity sequence:

Remark 2.1: Final Derived Example of an infinite sequence of points

$$(\mathbf{x}_i, y_i) = \begin{cases} ((\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}), 1) & i = 2n+1 \\ ((\frac{1}{2}, -\frac{\sqrt{3}}{2}), 1) & i = 2n+2\,, \quad n \in \mathbb{Z} \\ ((0, \frac{\sqrt{3}}{2} - \frac{1}{\sqrt{2}}), 1) & i = 2n+3 \end{cases} \tag{29}$$

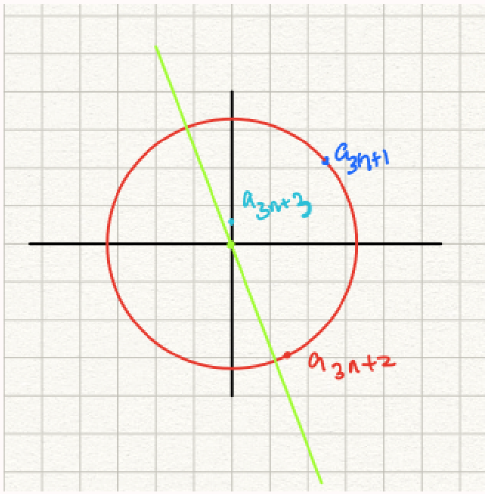

R=1
$\vec{a}_{3n+1}$
$\theta_{3n+2 \to 3n+3} > \frac{\pi}{2} \ \forall\, n \in \mathbb{Z}$
$\vec{a}_{3n+3}$
$\theta_{3n+2:3n+3}$
a_{3n+2}

By observation, this repeated series will satisfies three properties from Property I, Property II, and Property III. Now, let's formerly prove it does indeed meet all properties:

Proof 2.1: Property I - Strictly Linearly Separable

Since all labels are +1, and all data points are asymmetric, hence the data $\mathscr{D}$ is strictly linear separable with $b = 0$.

As shown in the graph below, there would exist some hyperplane, where all points on one side of the hyperplane through the origin with $\gamma > 0$.



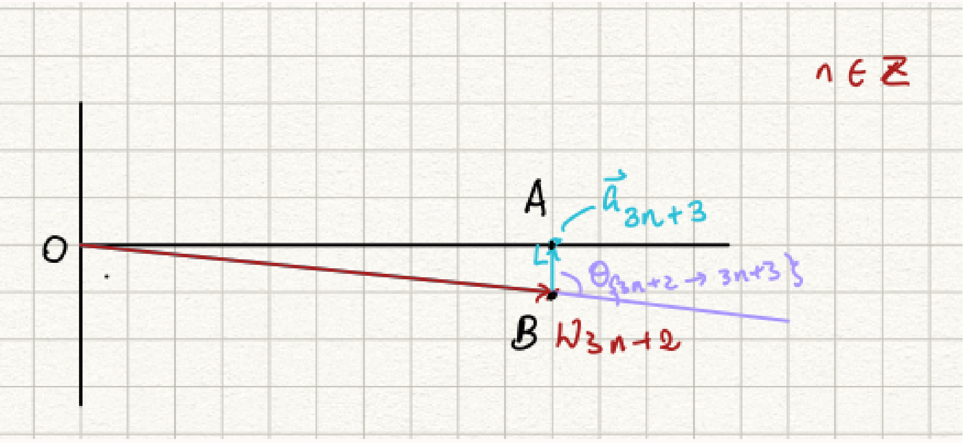
Proof 2.2: Property II - max $||x_i||_2 \leq 1$

$$\because \quad \|\mathbf{a}_{3n+1}\|_2 = 1 \quad \|\mathbf{a}_{3n+2}\|_2 = 1 \quad \|\mathbf{a}_{3n+3}\|_2 = \frac{\sqrt{3}}{2} - \frac{1}{\sqrt{2}} \quad \forall n \in \mathbb{Z} \tag{30}$$

$$\therefore \quad \|x_i\|_2 = \|a_i\|_2 \leq 1, \forall i \in \mathbb{Z} \tag{31}$$

Proof 2.3: Property III: infinite number of mistakes

We can guarantee the angle formed between every third point and prior weight vector would be larger or equal than 90 degree. As the figure below shown, the angle $\theta w_{3n+2} \rightarrow \mathbf{a}_{3n+3} \geq \frac{\pi}{2}, \forall n \in \mathbb{Z}$.



Q.E.D.