Answer 1

The technical paper titled On Convergence Proofs for Perceptrons by Albert Novikoff states that no matter what assignment of weights we begin with the process of recursively readjusting the weights by the method known as 'error correction' will terminate after a finite number of correction in a satisfactory assignment, provided any such satisfactory assignment exists.

In simple terms we can also say that, the perceptron learning rule is guaranteed to converge to a weight vector that correctly classifies the examples provided the training examples are linearly separable.

Re-Statement of Proof:

Pre Conditions / Assumptions -

1. We assume that the data-set is linearly separable
2. It can be classified into two classes

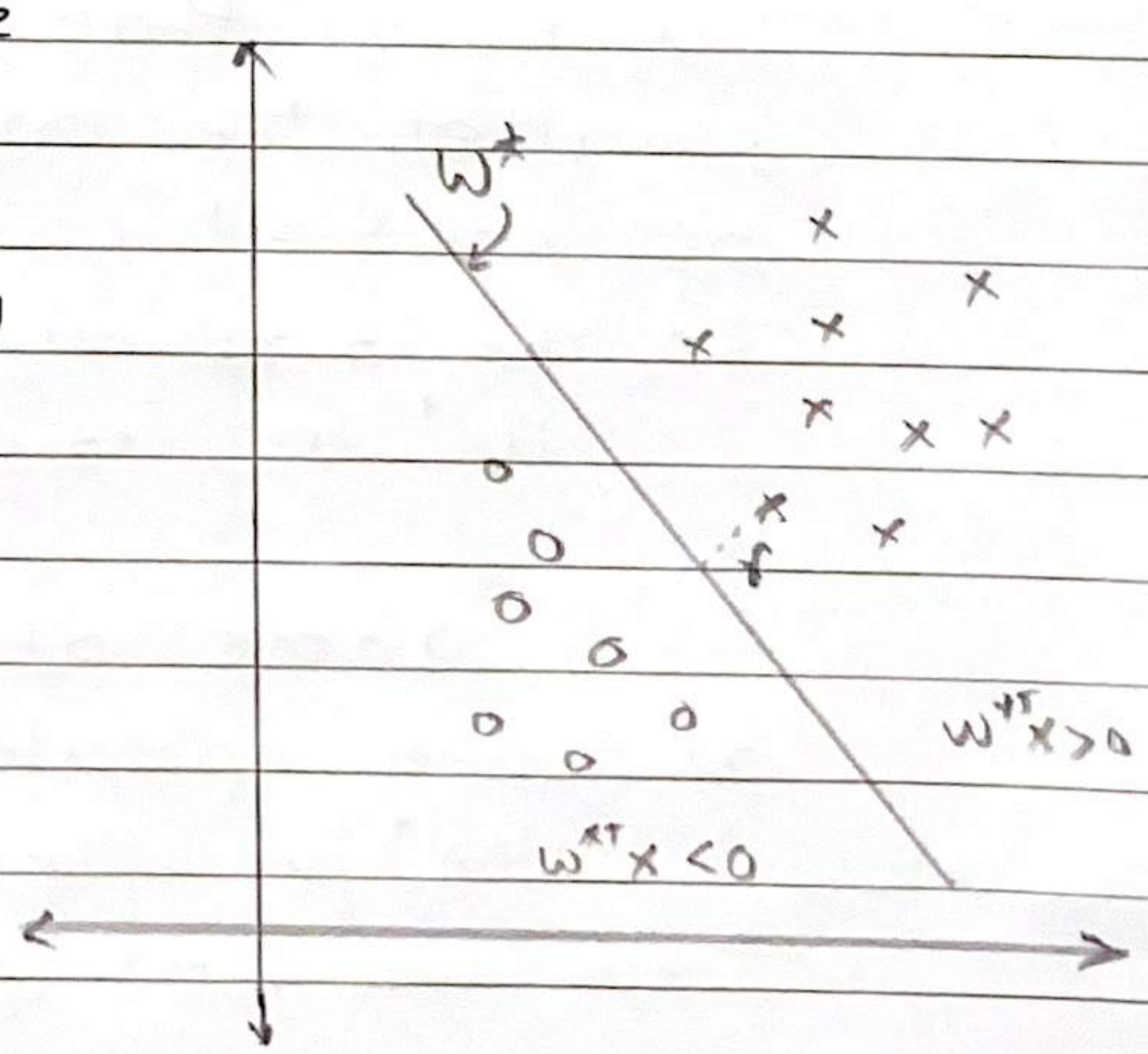
   Class I

   $w^{*T} x > 0$

   &

   Class II

   $w^{*T} x < 0$



3. We will assume that the length of every input vector $\|x\| \leq 1$
4. We know that there is some weight vector w^* that separates the data into two class with $\|w^*\| = 1$ i.e. w^* lies exactly on the unit sphere
5. Given set of weight vectors $w_1, w_2, \ldots w_N$

6. Given data set as $x_1, x_2, \ldots x_N$

As per the statement of convergence theorem Perceptron learning algorithm aims to find some weight vector 'w' that is parallel to w^* or as close as possible.

We know that when two vectors are parallel the angle between them is $\theta = 0°$
so $\cos\theta = 1$
so the size of the inner product $w^* w$ is a maximum

From above statement, if we show that at each update $w^* w$ increases
then we have nearly shown that the

FOR EDUCATIONAL USE

algorithm will converge.
However we need to go further and check the length of w does not increase too much

Hence we need to check two things when we consider a weight update:
i) The value of $w^* w$
ii) length of w

Proof:
Now suppose that at t^{th} iteration a particular input x get a wrong output y
so

$$y\, w^{(t-1)} \cdot x < 0$$

from this we can say

1] $$y\,(w^T x) \leq 0$$

This holds because x is misclassified by w, otherwise we couldn't make the update

2] $$y\, w^{*T} x > 0$$

This holds because w^* is a correctly separating hyperplane & classifies all the point correctly

The weight update will be

$$w^{(t)} = w^{(t-1)} + yx$$

where $t-1$ index means weights at $(t-1)$ step
here we have considered $\eta = 1$ for simplicity (η = learning rate)

To see how this changes the two values in which we are interested we compute

$$w^* w^{(t)} = w^* \cdot (w^{(t-1)} + yx)$$

$$= w^* \cdot w^{(t-1)} + yw^* \cdot x$$

$$w^* w^{(t)} \geq w^* \cdot w^{(t-1)} + \gamma \quad \ldots \text{(A)}$$

The inequality follows from the fact that, for w^*, the distance from the hyperplane defined by w^* to x must be at least γ

i.e. $y(w^* \cdot x) = |w^* \cdot x| \geq \gamma \quad \ldots$

γ is the smallest distance between the optimal hyperplane defined by w^* and any data point x. Also referred as 'margin'

From equation A we conclude that with each update the inner product increases by at least γ

So after t updates of weights

$$w^* w^{(t)} \geq t\gamma \quad \ldots\ (i)$$

We can use this to put a lower bound on the length of $\|w^{(t)}\|$ by using the cauchy-schwartz inequality

$$w^* . w^{(t)} \leq \|w^*\| \, \|w^{(t)}\|$$

so

$$\|w^{(t)}\| \geq t\gamma \quad \ldots\ (B)$$

$\because \|w^*\| = 1$ from point 4 in assumptions

The length of the vector after t steps is

$$\|w^{(t)}\|^2 = \|w^{(t-1)} + yx\|^2$$

$$= \|w^{(t-1)}\|^2 + y^2\|x\|^2 + 2y\, w^{(t-1)} . x$$

$$\|w^{(t)}\|^2 \leq \|w^{(t-1)}\|^2 + 1$$

Above line follows because.

i) $y^2 = 1$ $\because$ y conbe either 1 or -1

ii) $\|x\| \leq 1$ from point 3 in assumptions

iii) $2y(w^{(t-1)} \cdot x) < 0$ as we made an update because x was misclassified

$w^{(t-1)} \cdot x$ are perpendicular to each other

This means that for each update w grows by at most 1

Therefore after t steps

$$\|w^{(t)}\|^2 \leq \cancel{\|w^{(t-1)}\|^2} + t \quad \ldots$$

$$\|w^{(t)}\|^2 \leq t \quad \ldots \text{(ii)}$$

we can put these two inequalities together to get

$$t\gamma \leq \|w^{(t-1)}\| \leq \sqrt{t}$$

Solving for t we can conclude the proof.

From ~~B~~ (i) we know

$$t\gamma \leq w^{*} w^{(t)}$$

$$= \|w^{(t)}\| \cos\theta$$ by definition of inner product

where θ is the angle between w & w^{*}

$$\leq \|w^{(t)}\|$$ by definition of cos, we must have $\cos(\theta) \leq 1$

$$t\gamma \leq \sqrt{t}$$ from equation (ii)

$$\therefore \quad t^2\gamma^2 \leq t$$

$$t \leq \frac{1}{\gamma^2} \quad \ldots \text{(D)}$$

from D we conclude that the number of updates t is bounded by a constant i.e. after that many updates the algorithm must have converged.

From the above proof we showed that if the weights are linearly separable then the algorithm will converge, and the time that it takes is function of distance between the separating hyperplane & nearest datapoint.