Update Rule

$w^{t+1} = w^t + x_i y_i$ → if mistake $sign(x_i^T w) \neq sign(y_i)$

Now, we need to understood how by convergence we will get a bet for w^t!!

→ If we think about it, there are two different ways this update rule can go wrong.

mistake ①

Predicted = 1 $sign(w^{t^T} x_i) \geq 0$

Actual = -1 $y_i = -1$

mistake ②

predicted = -1 $sign(w^{t^T} x_i) < 0$

Actual = +1 $y_i = +1$

→ Say, we are using the update rule as $w^{t+1} = w^t + x_i y_i$

→ So, if we take update rule as a good one, then it should not make mistake for the input for which previous weight had made a mistake.

→ Let's find the new weight dot product with the point where old weight made a mistake.

→ $(w^{t+1})^T x_i = (w^t + x_i y_i)^T x_i = w^{tT} x_i + y_i \|x_i\|^2$

→ Now, considering type 1 mistake,

we have, $w^{tT} x_i \geq 0$

$y_i = -1$

Substitute them in new weight,

And we get,

$$(w^{t+1})^T x_i = \underbrace{w^{tT} x_i}_{\geq 0} + \underbrace{y_i}_{-1} \underbrace{\|x_i\|^2}_{\geq 0}$$

This whole thing is negative.

→ From this, what we can say is that the new weight's dot product with x_i is equal to,

old weight's dot product with x_i − Something.

→ So, we are subracting out something from old w's dot product.

→ well, the old w's was positive, and by subtracting something, we can say that the value is reducing.

→ ofcourse, it doesn't immediately become negative, but however it is moving in the right direction as we are subtracting!!

→ Similarly if we consider Type 2,

where,

$$(w^{t^T} x_i) < 0$$
$$y_i = +1$$

substitute them in new weight,

And we get,

$$(w^{t+1})^T x_i = \underbrace{w^{t^T} x_i}_{<0} + \underbrace{y_i \|x_i\|^2}_{>0}$$

→ and here we are adding something.

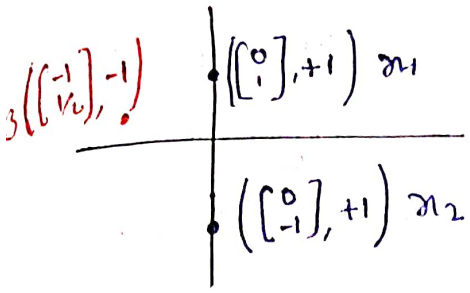
→ overall we can say that,

update Rule pushes w in the right direction for x_i.
i.e; if dotproduct is -ve, when it should have been +ve, the update rule is subracting something.

→ But then, the question raised that, okay we fixed the previous mistake, but what's the guarantee that this newweight didn't breaks anything that was previously correct!!

→ So Problem is – fixing w for one x_i, might affect decision for other data points. So we need more careful argument for convergence.

Let's solve an example..



Is this a linear seperable data or?

yes, there is $w \in \mathbb{R}^2$

s.t. $w^T x_i \geq 0 \Rightarrow y_i = +1$

$w^T x_i < 0 \Rightarrow y_i = -1$

Solving with perceptron as it's linearly seperable.

Initially we can take the weight $[0, 0]$

And that given us, $w^0 = [0, 0]$

$w^{0T} x_1 = 0$ | $w^{0T} x_2 = 0$ | $w^{0T} x_3 = 0$ → This is a mistake.

$\hat{y}_1 = +1$ | $\hat{y}_2 = +1$ | $\hat{y}_3 = +1$

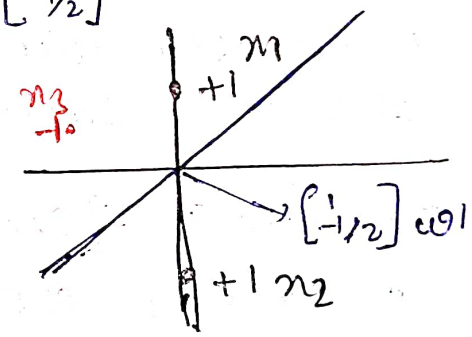
$y_3 = -1$

So we find w^{t+1}

$w^1 = w^0 + x_3 y_3 = \begin{bmatrix}0\\0\end{bmatrix} + \begin{bmatrix}-1\\1/2\end{bmatrix} \times -1 = \begin{bmatrix}1\\-1/2\end{bmatrix}$

→ Now with the new w_1 line, we can see that it made mistake for x_1.

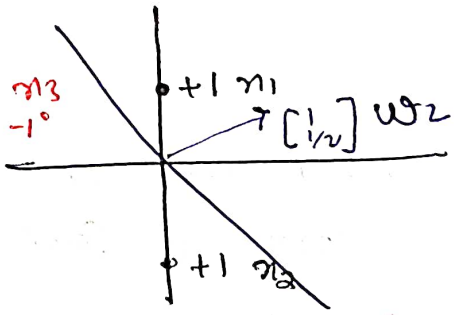
So we find w_2



$w^2 = w^1 + x_1 y_1 = \begin{bmatrix}1\\-1/2\end{bmatrix} + \begin{bmatrix}0\\1\end{bmatrix} \cdot 1 = \begin{bmatrix}1\\1/2\end{bmatrix}$

→ Now with the new w_2 line, we can see that it made mistake for x_2.

So we find w^3.



$w^3 = w^2 + x_2 y_2 = \begin{bmatrix}1\\-1/2\end{bmatrix} = w^1$

And now w^3 has become w_1, and then w^4 will become w_2 & it will keep looping like this & it will not have a convergence!!

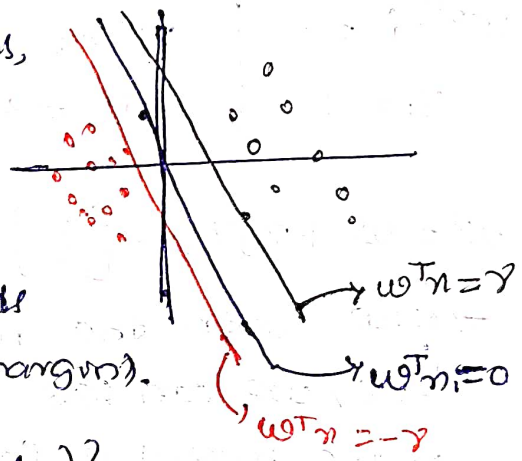
→ Now if we wonder is perceptron wrong or data wrong. And if we closely observe the data, the points actually lie on the decision boundary, we are saying $w^T x_i \geq 0$ → it can also be equal to "0".

i.e, $w^T x_i$ can be "0". so we need to make sure it can't be zero.

So the data Assumption has to be changed.

① Linear seperability with "γ" margin.

→ If we take the data like this, where there is a region with no datapoints. Then data can be modelled well with perceptron. It needs atleast a small region (γ-margin).



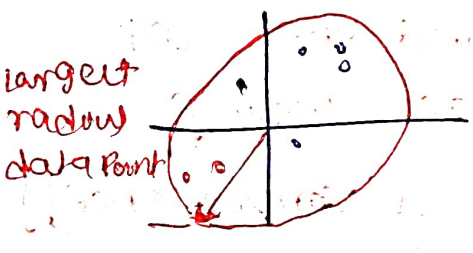
→ A dataset $\{(x_1, y_1), \ldots (x_n, y_n)\}$ is linearly seperable with γ-margin.

→ If $\exists w^* \in \mathbb{R}^d$ s.t, $(w^{*T} x_i) y_i \geq \gamma \ \forall i$ for some $\gamma > 0$.

→ For ease or simplification or Proof, we can consider few harmless assumptions such as

② Radius Assumption

$\forall i$ (every point) $\in D$, $\|x_i\|_2 \leq R$ for some $R > 0$



→ For a largest point, we are away, all other points we could that circle.

3) without loss of generality, assume $\|w^*\| = 1$,

$w^* \rightarrow$ The actual best line that seperates the data.

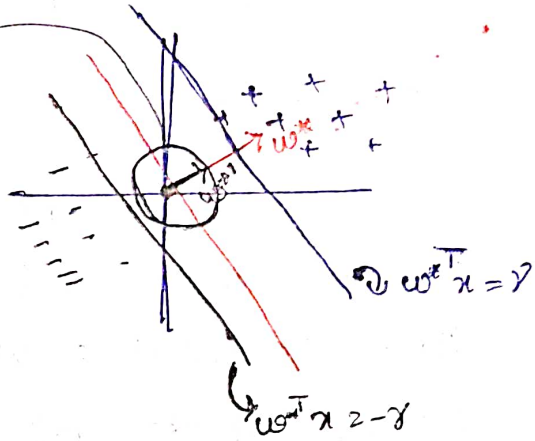
we know that there will be w^*, but we can't be sure will norm ($\|w^*\|$) will be equal to 1.

However, we can say that there will exist a w^{*1}, which will have norm = 1.

we have,

$w^{*T} x = \gamma$

considering $\|w^*\| = 1$



divide by $\|w^*\|$

$$\frac{w^{*T} x}{\|w^*\|} = \frac{\gamma}{\|w^*\|}$$

$$\left(\frac{w^{*T}}{\|w^*\|}\right) \cdot x = \gamma'$$

$$\therefore \gamma' = \frac{\gamma}{\|w^*\|}$$

↳ considered w^{*1}

$$(w^{*1})^T x = \gamma'$$

↳ Here we just scaled the w^*.

But the advantage is that, the new w^{*1} we have now has norm = 1.

i.e, If you give me a dataset, with w^* & some γ, say $\gamma = 10$.

Now I can create another w^* by rescaling the w^*, to make sure it has length 1, and the γ will rescale accordingly. Then I will have a w^{*1} which also linearly seperates data with a different γ', so this assumption holds ✓