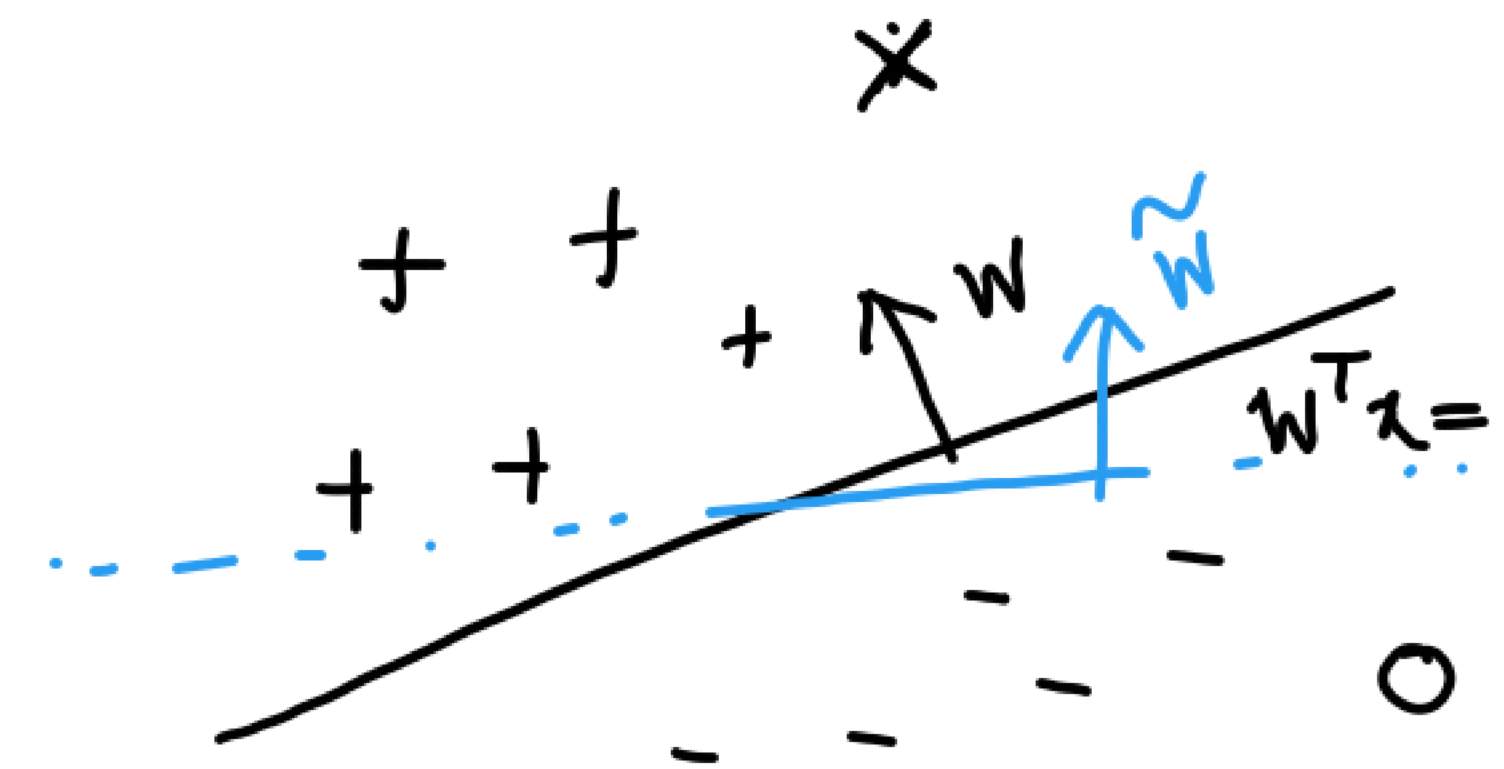
Lec 09: Perceptron : linear deterministic classification method

$$y_i = f(x_i) \in \{+1, -1\}$$



Goal: learn a weight vector w that linearly separates the training data points

$$\boxed{\begin{bmatrix} 1 \\ x_i \end{bmatrix} \quad w^T x}$$

$$w^T x_i + w_0 \geq 0 \quad \text{for } y_i = 1$$
$$< 0 \quad \text{o/w.}$$

Remark: The decision boundary is not unique.

$$f_W(x) = \mathrm{sgn}(W^T x)$$

Q: How does perceptron find W?

1. Goes over the training examples (x_i, y_i) one by one W_0
2. Check if the current classifier W_t, i.e.
$$y_i = \mathrm{sgn}(W_t^T x_i)$$
3. If correct – no update
4. If not – correct W_{t+1}
$$W_{t+1} \leftarrow W_t + \underline{\underline{y_i}}\, x_i$$
vector (d+1)
5. STOP if no update for a certain number of iteration.

Remarks: 1 Online algorithms

The data points are used one-by-one

$\leftrightarrow$ batch algorithms (LR, NB) – takes all data

Algorithm: Perceptron

- Initialize w_0
- for $t = 0, 1, 2, \ldots, \text{maxRounds}$: $y_i = \text{sgn}(w^T x_i)$

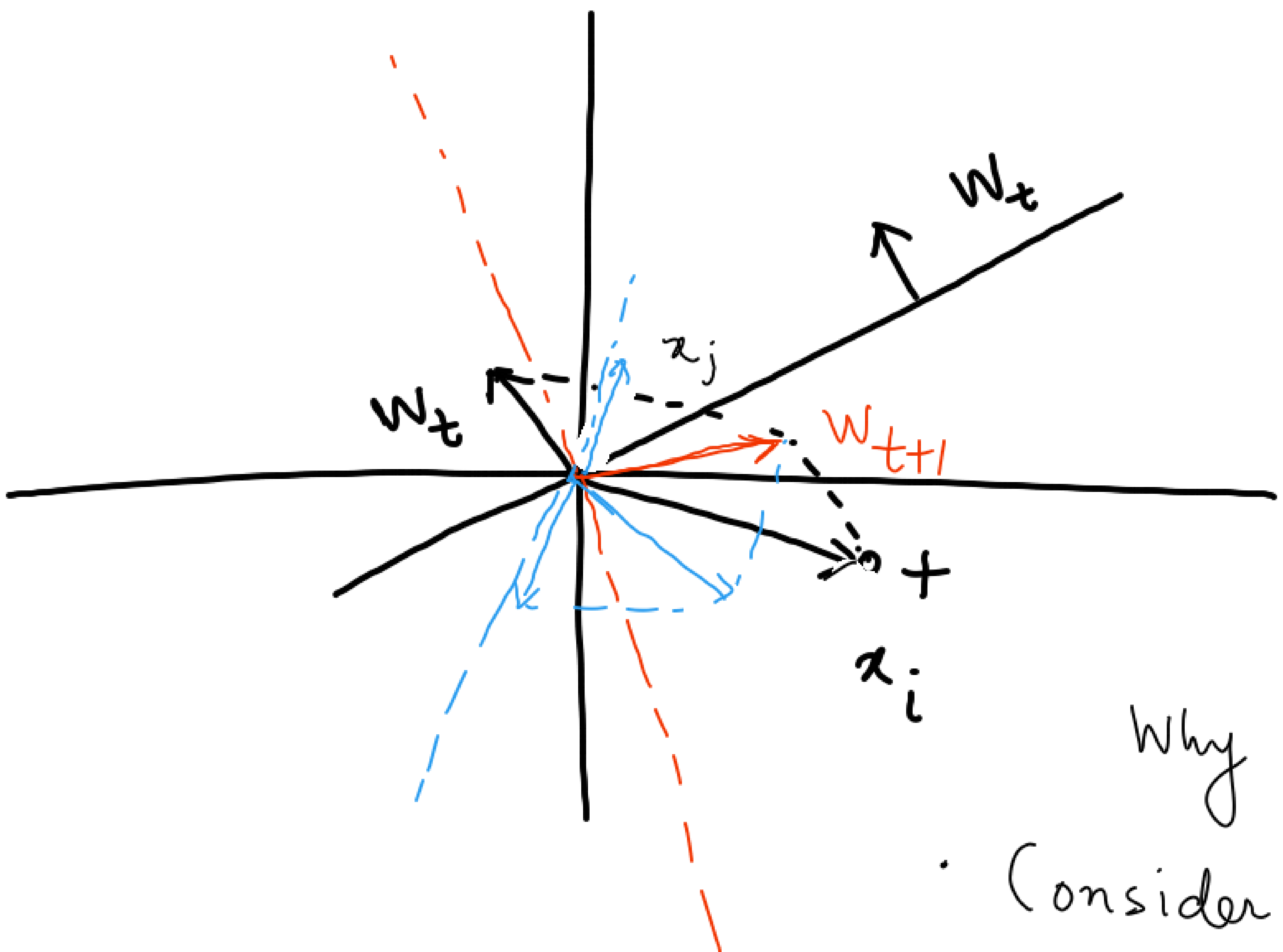
  Randomly choose a training example (x_i, y_i)

  if $y_i(w_t^T x_i) < 0$, then

  $$w_{t+1} \leftarrow w_t + y_i x_i$$

- STOP as before.

$$w_{t+1} \leftarrow w_t + \frac{1}{2}\left(y_i - \text{sgn}(w_t^T x_i)\right) \times x_i$$



$$w_{t+1} \leftarrow w_t + x_i$$

$$w_{t+2} \leftarrow w_{t+1} - x_j$$

$$w_{new} = w_{old} + y_i x_i$$

Why is perceptron doing a meaningful update?

- Consider a misclassified example (x_i, y_i)

  i.e. $\operatorname{sgn}\left(w_{old}^T x_i\right) \neq y_i$

$$y_i\left(w_{new}^T x_i\right) = y_i\left(w_{old} + y_i x_i\right)^T x_i = y_i w_{old}^T x_i + \underline{\|x_i\|^2}$$

$$> y_i\left(w_{old}^T x_i\right)$$

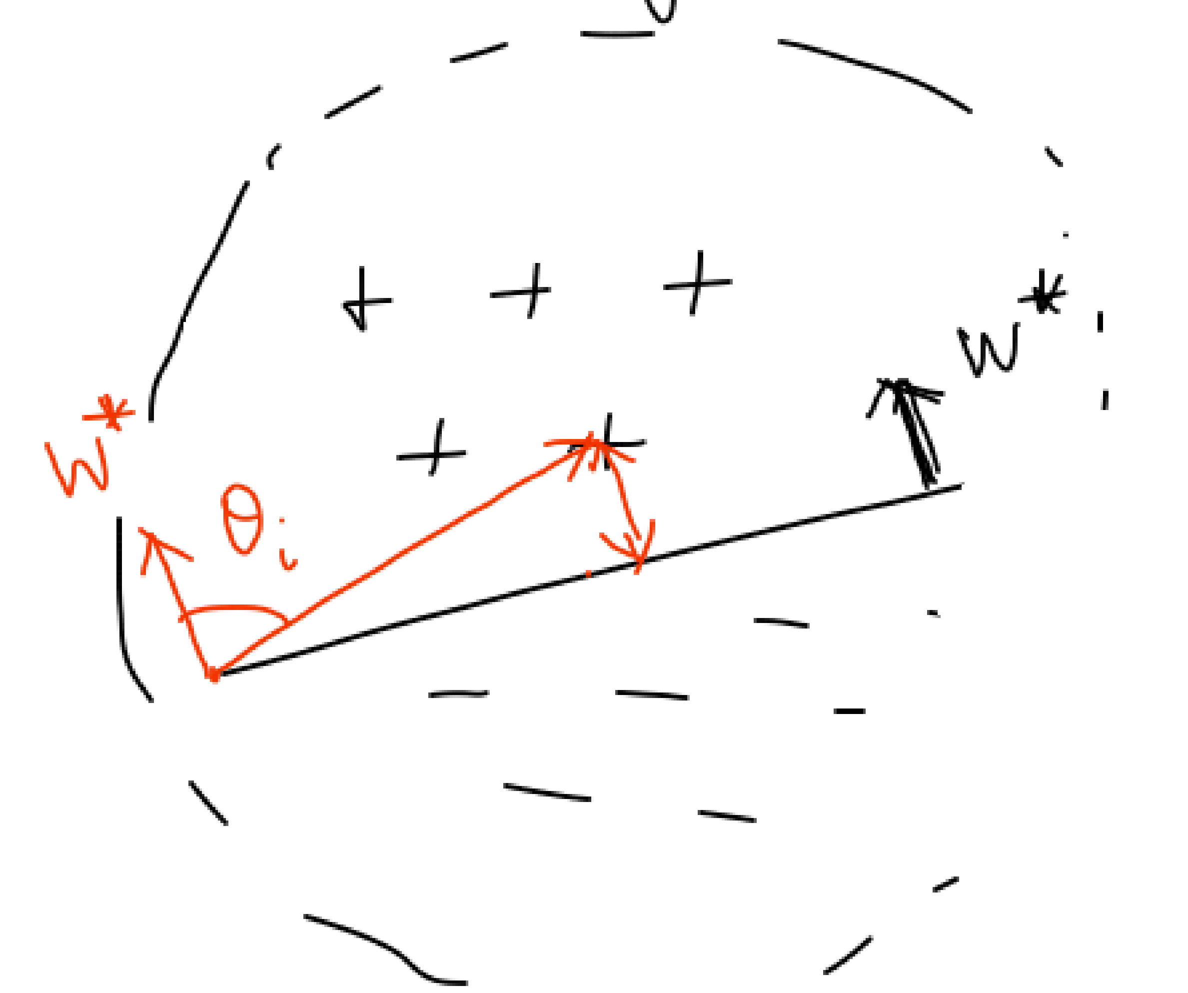
Note: still not guaranteed to be correctly classified.

Summary:

- Perceptron is a mistake-driven online learning algo.
- Guaranteed to converge for linearly separable training examples.

If the data is linearly separable,

$\exists\, w^*$ s.t. $\quad y_i({w^*}^T x_i) \geq 0 \quad \forall\, i = 1, \ldots, n.$



Assume $\|w^*\| = 1$, $\quad \|x_i\| \leq 1$

Define, margin of separation $\quad \gamma = \min_i |{w^*}^T x_i| = \|w^*\| \|x_i\| \cos\theta_i$

Theorem: If $\exists$ a unit vector w^* s.t. $y_i {w^*}^T x_i \geq \gamma \quad \forall (x_i, y_i) \in D$

Then the # of weight updates by perceptron is at most $\dfrac{1}{\gamma^2}$.

Proof: track two quantities (1) $w_t^T w^*$, (2) $\|w_t\|_2^2$

(1) Claim: $w_t^T w^*$ on every update increases by at least γ

$$w_{t+1}^T w^* = (w_t + y_i x_i)^T w^*$$

$$= w_t^T w^* + \underbrace{y_i (w^{*T} x_i)}_{\geq \gamma} \geq w_t^T w^* + \gamma$$

(2) Claim: $\|w_t\|^2$ increases by at most 1.

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

Say $w_0 = 0$. After k updates

$$\left.\begin{aligned} w_{k+1}^T w^* &\geq k\gamma \\ \|w_{k+1}\|^2 &< k \end{aligned}\right\}$$

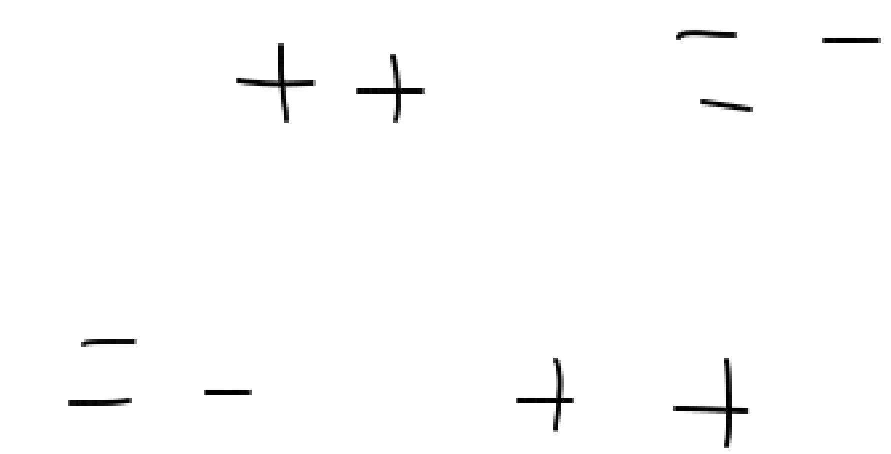
$$\sqrt{k} > \|W_{k+1}\| \geq W_{k+1}^T w^* \geq k\gamma \Rightarrow k < \frac{1}{\gamma^2}$$

$$W_{k+1}^T w^* = \|W_{k+1}\| \underbrace{\|w^*\|}_{=1} \underbrace{\cos\theta}_{\leq 1}$$

finite number of mistakes if data is linearly separable.

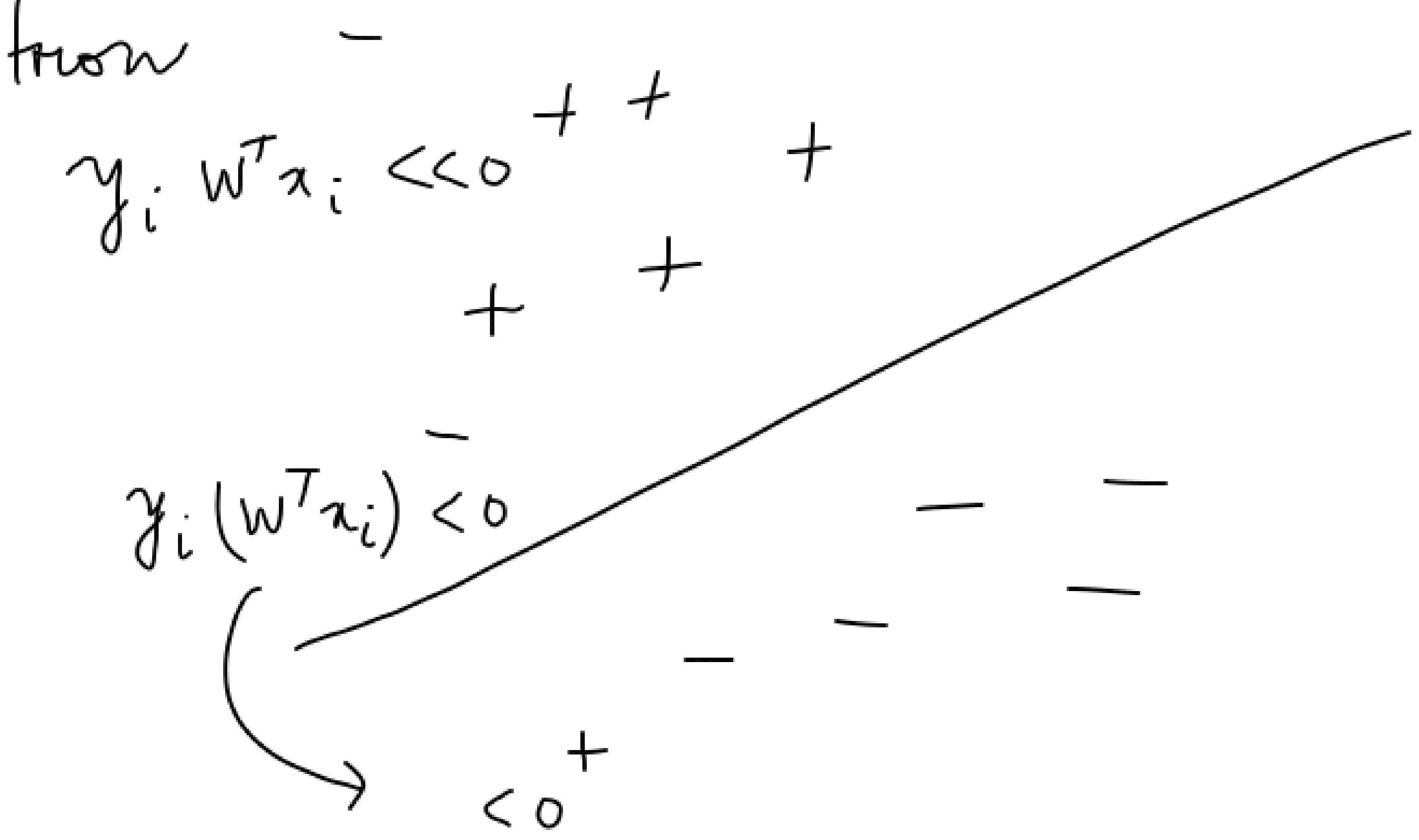
Limitations:

1. Not giving a rate of convergence
2. The # of iterations can be large if γ is small
3. May not converge if points are not linearly separable.



Find one such example of non convergence (HW)

The loss function view of perceptron



maximize $y_i (w^T x_i)$

$$\Rightarrow \min_{w} \sum_{i \in \text{misclassified examples}} \left(-y_i (w^T x_i)\right)$$

for any i

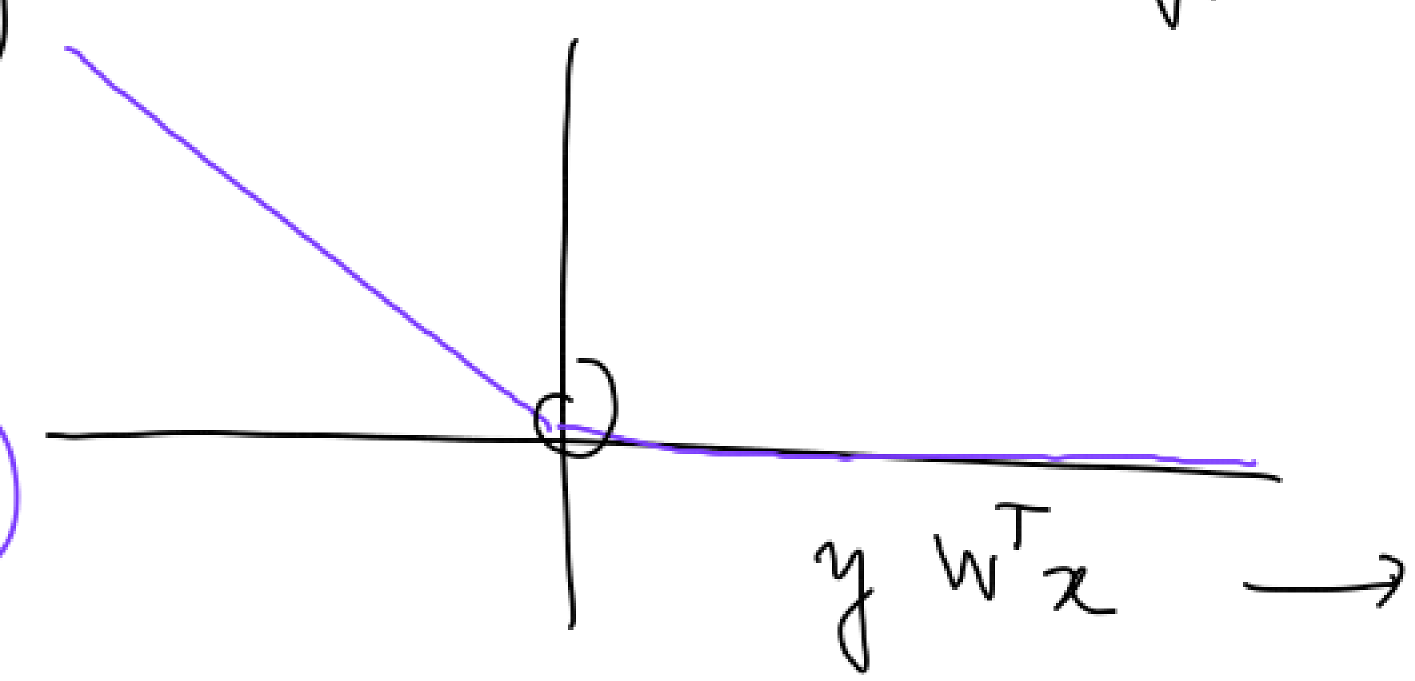
$y_i w^T x_i \geq 0 \rightarrow \text{loss} = 0$

$< 0 \rightarrow \text{loss} = -y_i w^T x_i$

$$L_i(w, D) = \max\{0, -y_i w^T x_i\}$$

Hinge loss

$$L(w,D) = \sum L_i(w,D)$$



Apply SGD : $\longrightarrow$ randomly pick i , compute $\nabla_w L_i(w, D)$

$$w_{t+1} \leftarrow w_t - \underbrace{\nabla_w L_i(w, D)}_{-y_i x_i}$$

$$-y_i w^T x_i$$

Hinge loss with SGD

is The perceptron algo.

$$w_t + y_i x_i$$

Decision Trees

Fuel efficiency	cyl	disp	Origin	Year
good	3	low		
	4	med		
bad	5			
	6	high