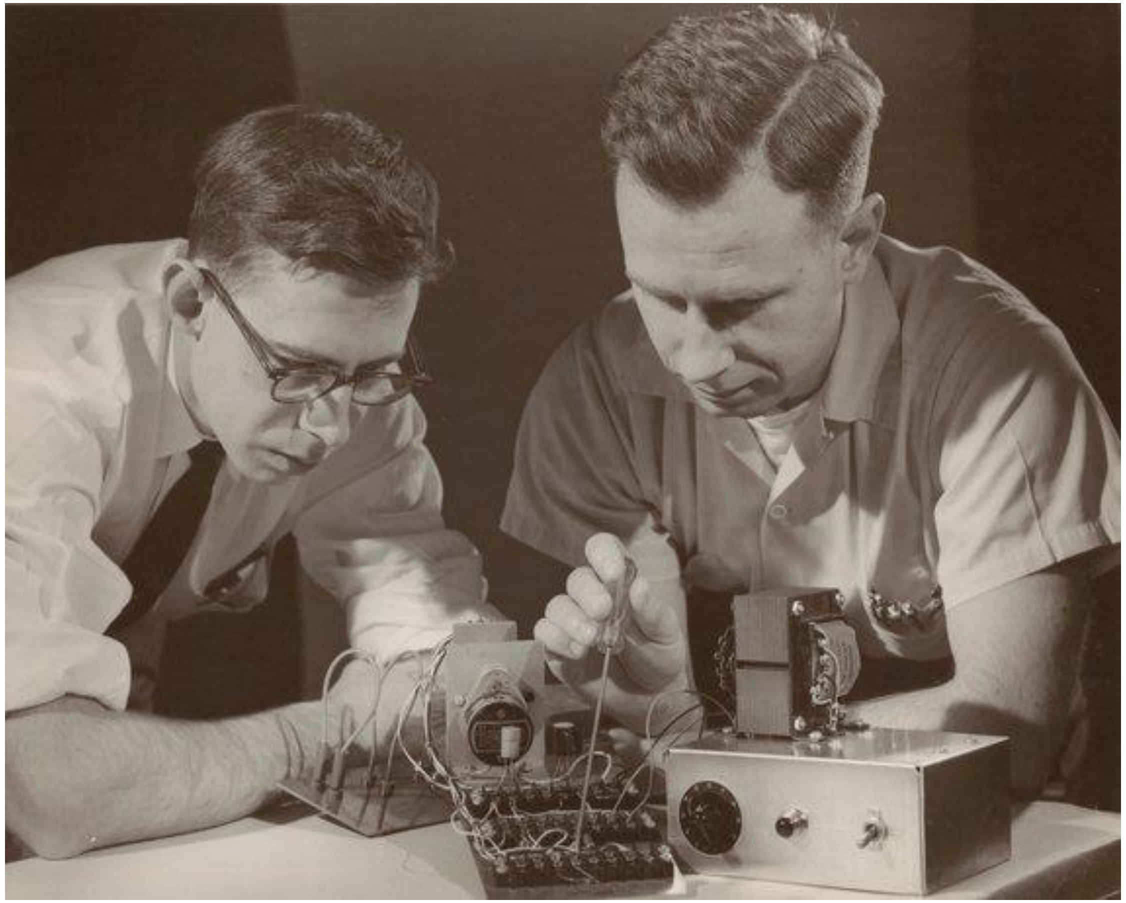
Perceptron

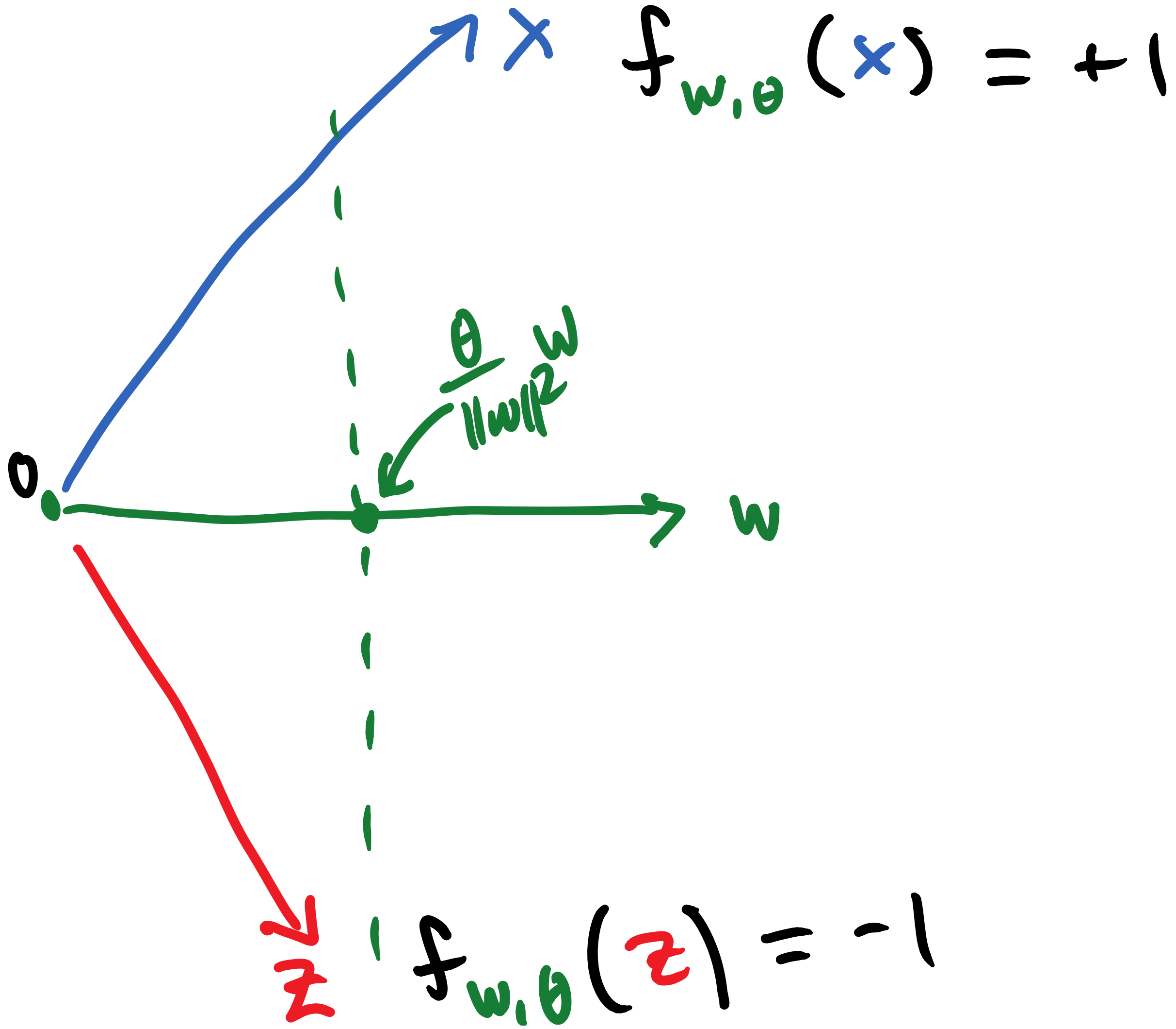


Linear classifiers

- Describe inputs using "feature vectors" in $\mathbb{R}^d$.
- Linear classifier: $w \in \mathbb{R}^d$ (*weight vector*) and $\theta \in \mathbb{R}$ (*threshold*)

$$f_{w,\theta}(x) = \begin{cases} +1, & \langle x, w \rangle > \theta \\ -1, & \langle x, w \rangle \leq \theta \end{cases}$$

Linear classifiers



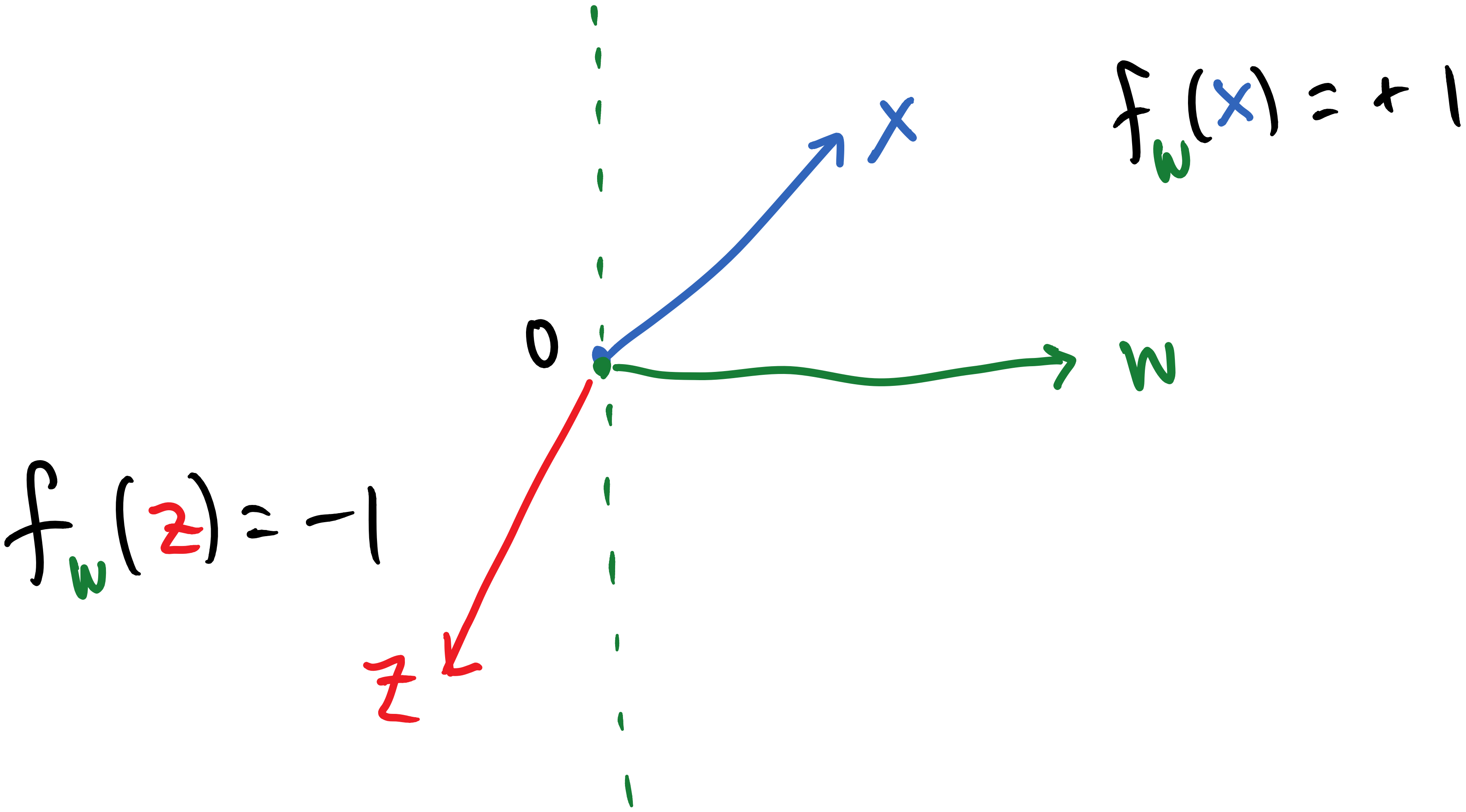
Homogeneous linear classifiers

- Homogeneous linear classifier: $w \in \mathbb{R}^d$ (*weight vector*)

$$f_w(x) = f_{w,0}(x) = \begin{cases} +1, & \langle x, w \rangle > 0 \\ -1, & \langle x, w \rangle \leq 0 \end{cases}$$

w must at positive side!!!

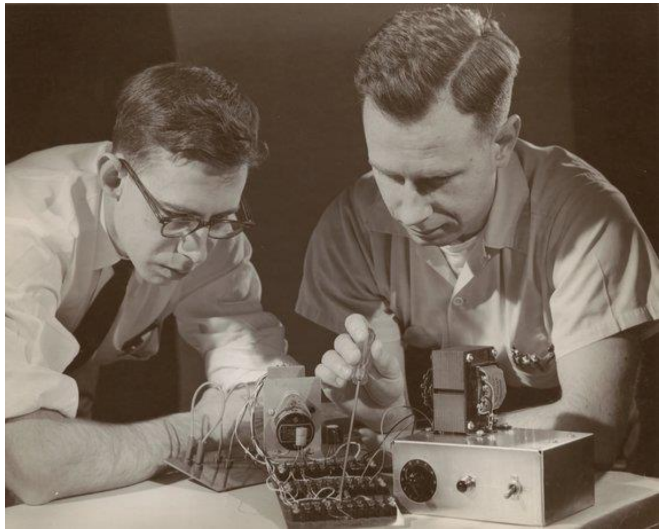
Homogeneous linear classifiers



Lifting non-homogeneous linear classifiers

- Suppose $f_{w,\theta}$ is a non-homogeneous linear classifier in $\mathbb{R}^d$
- Map weight vector and threshold to $\tilde{w} := (w, -\theta) \in \mathbb{R}^{d+1}$
- Map feature vectors x to $(x, 1) \in \mathbb{R}^{d+1}$
- $f_{\tilde{w},0}$ is equivalent homogeneous linear classifier in $\mathbb{R}^{d+1}$

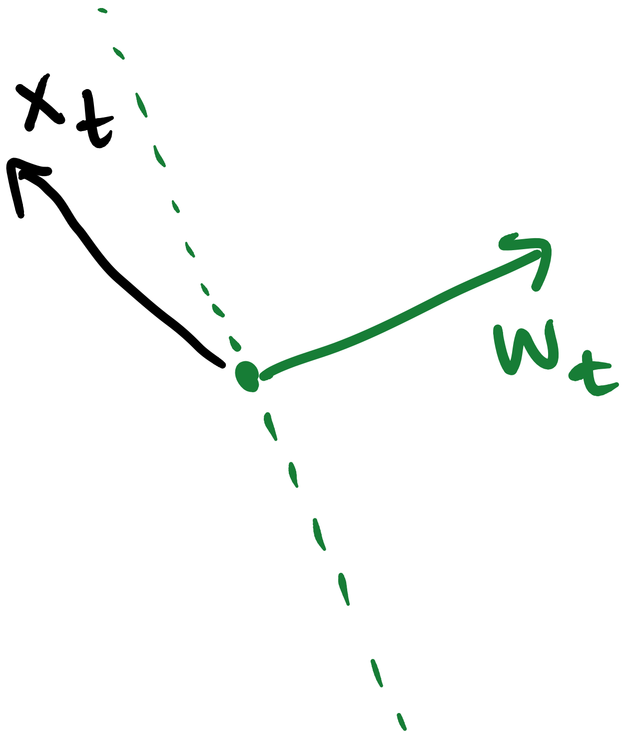
Perceptron (Rosenblatt, '58)



Input: training data S

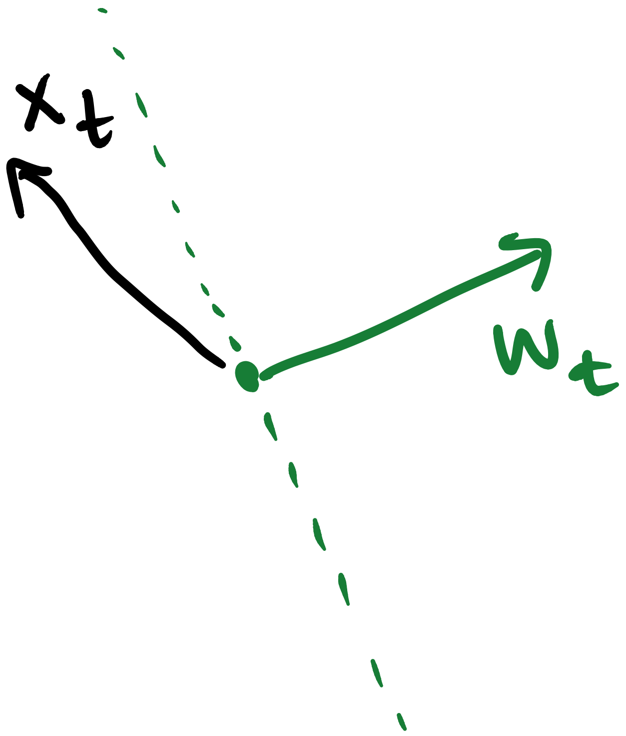
- **Let** $w_1 = \vec{0}$.
- **For** $t = 1,2, \dots$:
  - **If** there is $(x_t, y_t) \in S$ such that $f_{w_t}(x_t) \neq y_t$, **then**:
    - **Update**: $w_{t+1} \coloneqq w_t + y_t x_t$
  - **Else**: **return** w_t

Perceptron



predict $a_t = -1$

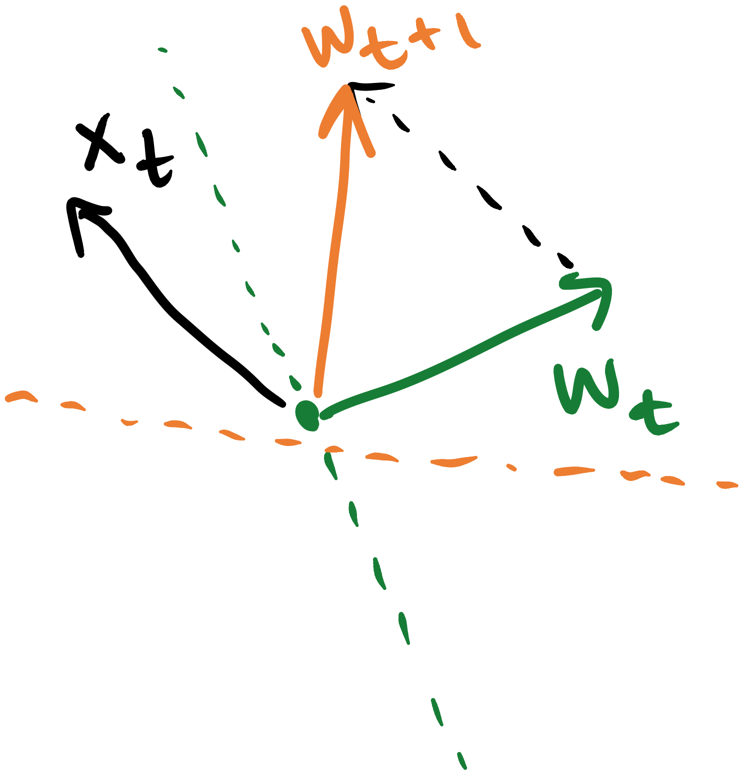
Perceptron



predict $a_t = -1$

correct label $y_t = +1$

Perceptron

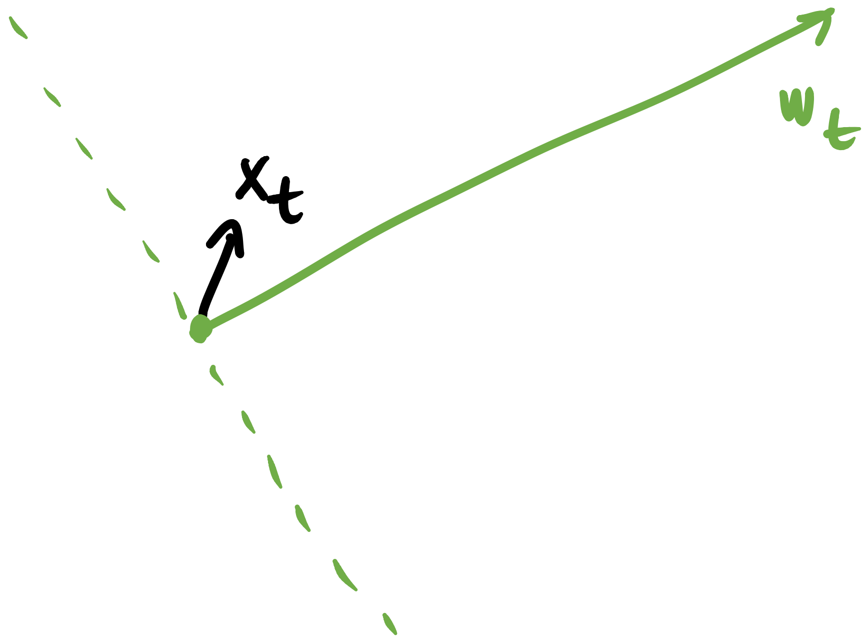


predict $a_t = -1$

correct label $y_t = +1$

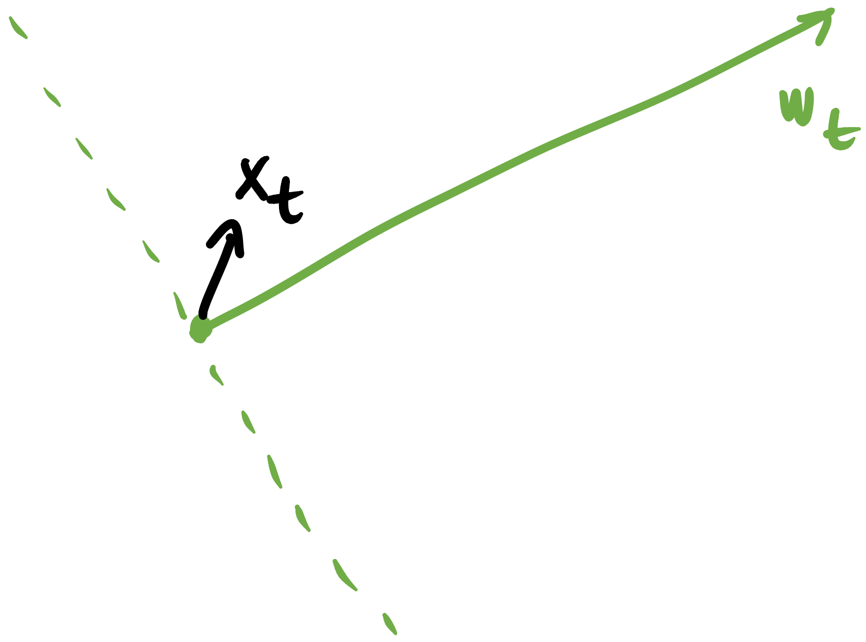
$w_{t+1} := w_t + y_t x_t$

Perceptron



predict $a_t = +1$

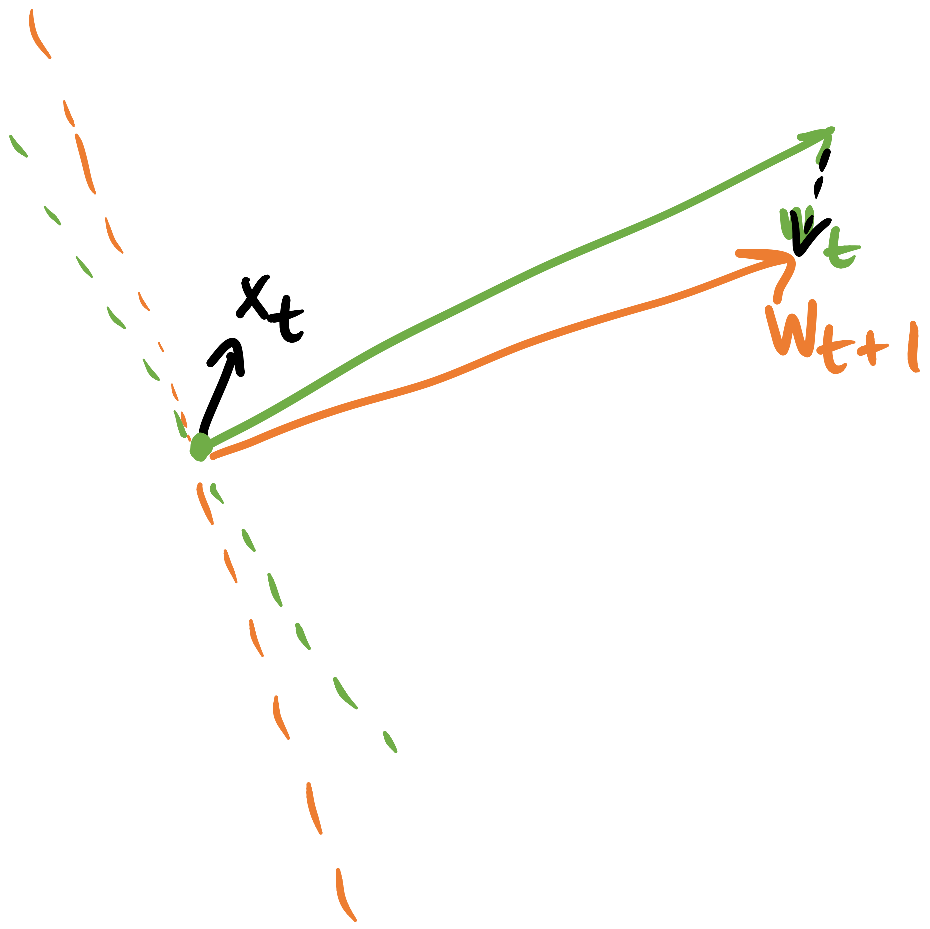
Perceptron



predict $a_t = +1$

Correct label $y_t = -1$

Perceptron



predict $a_t = +1$

Correct label $y_t = -1$

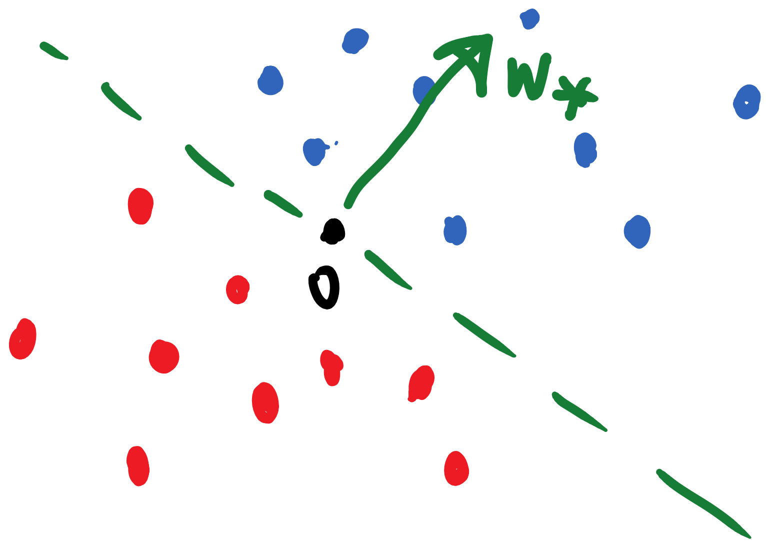
$$w_{t+1} := w_t + y_t x_t$$

Separable data

- Training data S from $\mathbb{R}^d \times \{-1, +1\}$
- Assume some $w_\star \in \mathbb{R}^d$ satisfies

$$y\langle x, w_\star \rangle > 0$$

  for all $(x, y) \in S$.



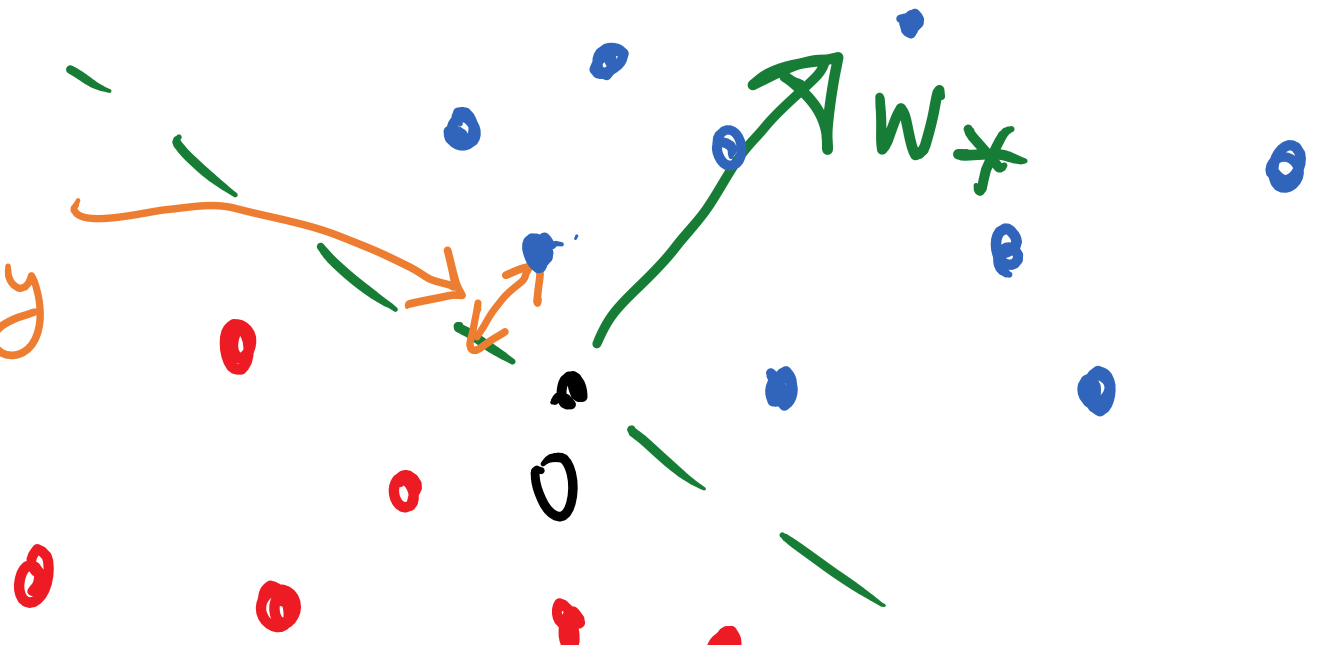
Separable data

- Training data S from $\mathbb{R}^d \times \{-1, +1\}$
- Assume some $\boxed{w_\star \in \mathbb{R}^d}$ satisfies

$$y\langle x, w_\star \rangle > 0$$

for all $(x, y) \in S$.

distance to boundary



Margins

find the w* with maximum margin(margin is the shortest distance from hyperplane to any point)

- Training data S from $\mathbb{R}^d \times \{-1, +1\}$

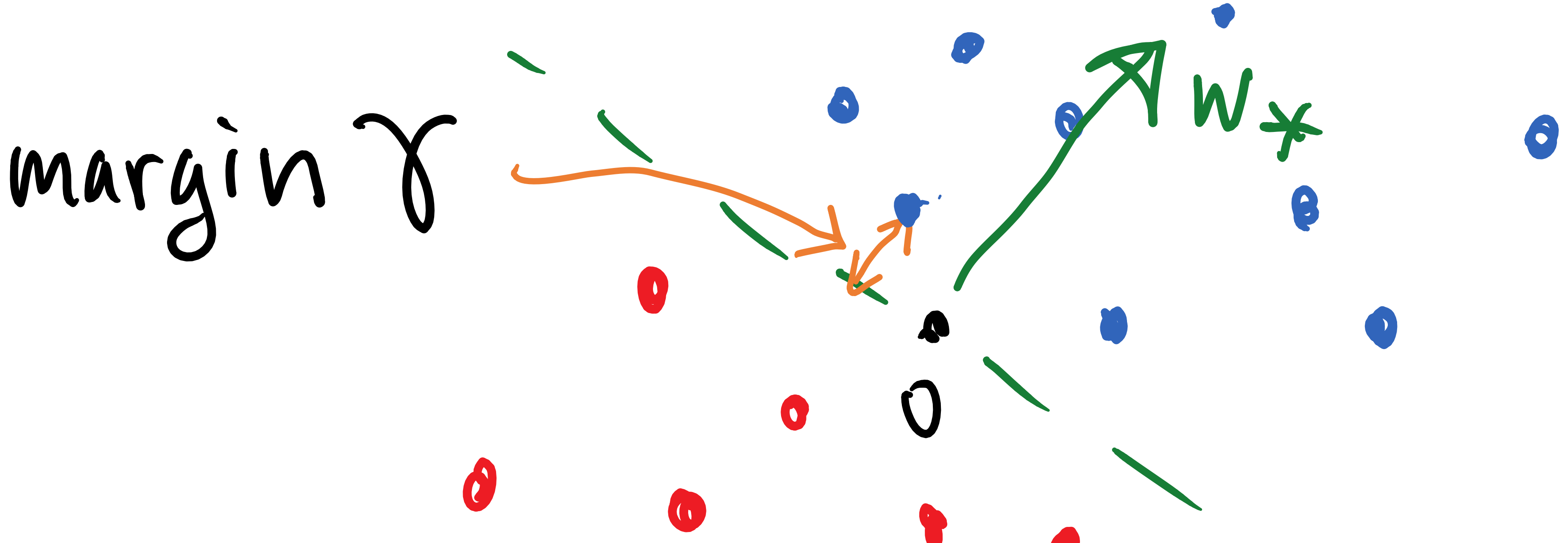
the margin is also decided by llw*ll

- Define the margin of S to be

find out the cloest point, and then adjust the llw*ll to get the max margin.

$$\gamma = \gamma(S) := \max_{\|w_\star\| \leq 1} \min_{(x,y) \in S} y\langle x, \boxed{w_\star \rangle.}$$

there is no wrong classification case!!!



Margins-based analysis of Perceptron

- Training data S from $\mathbb{R}^d \times \{-1, +1\}$
- Assume S is separable with margin $\gamma > 0$ (as witnessed by $w_\star$).
- Also, let $R \coloneqq \max_{(x,y) \in S} \|x\|$.

- Does Perceptron terminate?
- After how many updates?

Main idea of analysis

- Track the (cosine of the) angle between w_t and $w_\star$:
$$\frac{\langle w_\star, w_t \rangle}{\|w_\star\| \|w_t\|}$$

- With each update from w_t to w_{t+1}, how does this quantity change?

Margins-based analysis of Perceptron

Suppose Perceptron makes an update in iteration t.

This is positive

$$\langle w_\star, w_{t+1}\rangle = \langle w_\star, w_t + y_t x_t\rangle \geq \langle w_\star, w_t\rangle + \gamma$$

W* is a linear classifier could correctly classify all samples in the set.

$$\|w_{t+1}\|^2 = \|w_t\|^2 + 2\langle w_t, y_t x_t\rangle + \|y_t x_t\|^2 \leq \boxed{\|w_t\|^2} + R^2$$

R is the max length of any feature vector.

negative cause it's a false classification~~

Could be very small~~

Interesting

Margins-based analysis of Perceptron

Suppose Perceptron makes T updates.

$$\langle w_\star, w_{T+1} \rangle \geq T\boxed{\gamma}$$

$$\langle w_\star, w_{T+1} \rangle \leq \|w_\star\| \cdot \|w_{T+1}\| \leq R\sqrt{T}$$

Conclusion: number of updates must satisfy

no matter the sequence,
must satisfy this.

$$T \leq \left(\frac{R}{\gamma}\right)^2.$$

If the sample is separable, we have the upper bound over T~~could be calculated!