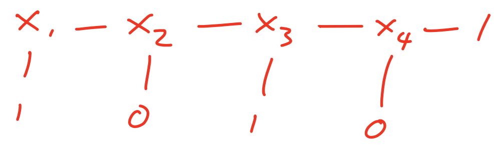
Last time: • $O(rn)$ MB alg. for length-r 1-DL's
• $O(k \log n)$ MB alg. (Winnow1) for length-k monotone disj.



Today: Algs for LTFs:
- Winnow2: alg for certain LTFs with positive weights over $\{0,1\}^n$
- Perceptron: alg for LTFs "with a margin" over $\mathbb{R}^n$
- start Dual Perceptron, "Kernelization"

Questions?

Winnow 2 motiv: uses LTFs - can it learn LTFs?

Recall W1: • sets wts to 0 - extreme, for general LTFs
• only gives wts 2^i: bad for some LTFs.

Ex: $x_1 + 2x_2 + 3x_3 + \dots + nx_n \geq n.$ $\{0,1\}^n$ inputs

Natural to use diff update param. than 2.

With these issues in mind, here's Winnow2.
Soon... first, here's a subclass of LTFs.

Fix param $0 < \delta < 1$.
Define $F(\delta)$, the class of "δ-separable monotone LTFs" over $\{0,1\}^n$, as:

$c \in F(\delta)$ iff exist $u_1, \dots, u_n \geq 0$ s.t. $\forall\, x \in \{0,1\}^n$, have

$$\begin{aligned} c(x) = 1 &\iff u \cdot x \geq 1 \\ c(x) = 0 &\iff u \cdot x \leq 1 - \delta \end{aligned}$$

Ex 1: r-out-of-k thr fn:

$c(x) = \mathbb{1}\{x_{i_1} + \dots + x_{i_k} \geq r\}$: $c(x) = 1$ iff $x_{i_1} + \dots + x_{i_k} \geq r$,

equiv to $\frac{1}{r}(x_{i_1} + \dots + x_{i_k}) \geq 1$

≥ 1 or $\leq 1 - \frac{1}{r}$:

is in $F(1/r)$.

Ex 2:

$x_1 + 2x_2 + 3x_3 + \dots + n x_n \geq n$. Rescale by $\frac{1}{n}$:

$x_i \in \{0,1\}$

$\frac{1}{n} \sum_{i=1}^{n} i x_i \geq 1$ either ≥ 1, or $\leq 1 - \frac{1}{n}$

in $F(1/n)$.

Here's W2.

(Initially $w_i = 1$, like W1)

Takes "update parameter" $\alpha > 1$ as input.

always n

Predicts using $h(x) = \mathbb{I}[w_1 x_1 + \dots + w_n x_n \geq n]$.

Updates:

- Promotions: multiply each w_i s.t. $x_i = 1$ by α.

- Demotions: divide each w_i s.t. $x_i = 1$ by α.

Thm: For $0 < \delta < 1$, let $c : \{0,1\}^n \to \{0,1\}$ be in $F(\delta)$, & let $u_1, \dots, u_n$ be its wts as above.
If W2 run with $\alpha = 1 + \frac{\delta}{2}$, then it makes at most

$$\frac{8}{\delta^2} + \left(\frac{5}{\delta} + \frac{14 \ln n}{\delta^2}\right) \sum_{i=1}^{n} u_i$$

$0 \leq u_i \leq 1$

$$= O\left(\frac{\log n}{\delta^2} \cdot \sum_{i=1}^{n} u_i\right)$$

many mistakes on c.

Ex: r-out-of-k thr fns:

$$\delta = \frac{1}{r}, \quad \sum_{i=1}^{n} u_i = \frac{k}{r}$$

+ bd is $O(kr \log n)$. Can be $\ll n$ if $k, r \ll n$.

Ex: Can show that any 1-DL of length r can be expressed as (using new vars $y_1, \dots, y_n$ for $\bar{x}_1, \dots, \bar{x}_n$) a δ-sep. "mon" LTF with $\delta = \frac{1}{2^{\Theta(r)}}$.

(idea: wts like $2^r, 2^{r-1}, 2^{r-2}, \dots$)

So can use W2 to learn length-r DL with m.b. of $\boxed{2^{\Theta(r)} \cdot \log n}$.

$\Theta(n^2)$: $O(n^2) + \Omega(n^2)$

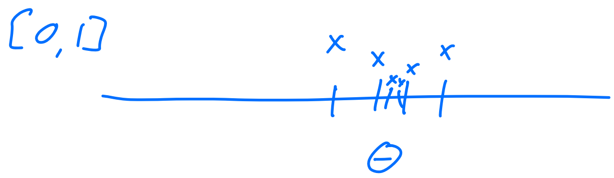
Can extend W2 to $[0,1]^n$ (need more assumpt...), nonmonotone LTFs, etc.

Noise-tolerant. Good in practice.

W2: 80s. ×

Perceptron: 50s +

Perceptron: alg for learning LTFs over $\mathbb{R}^n$.

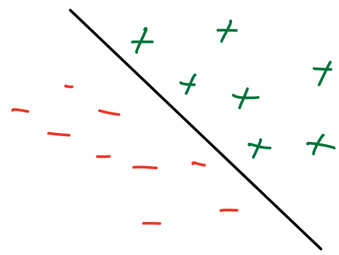


(?)

A: use a margin assumption.

Consider target LTF

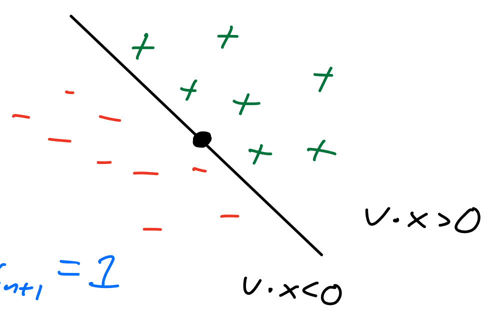
$\text{sign}(v \cdot x - \theta)$, i.e. $\mathbb{1}[v_1 x_1 + \dots + v_n x_n - \theta]$



Simplifing assumptions:

Can assume

① $\theta = 0$. i.e. pic is



WLOG!

↓

can add new coord. $x_{n+1} = 1$

to every ex, we get.

Now we're learning $\mathbb{1}[v' \cdot x' \geq 0]$

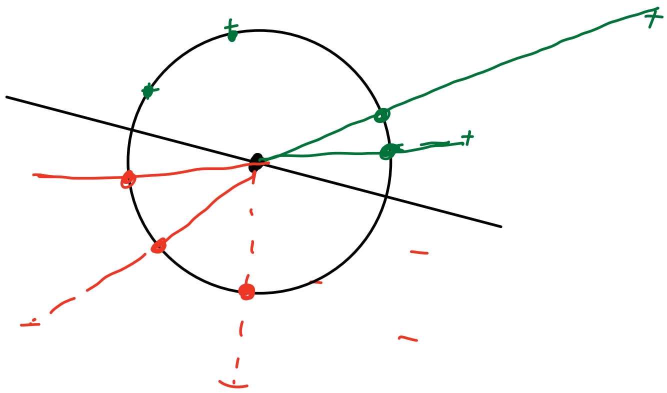
where $v' = (v_1, \dots, v_n, -\theta)$

$+ \; x' = (x_1, \dots, x_n, 1)$.

② Can assume each ex. $x \in \mathbb{R}^n$ (for $\underline{\text{origin-centered}}$ $v \cdot x \geq 0$)

is a unit vector: $\|x\|_2 = \sqrt{x_1^2 + \dots + x_n^2}$

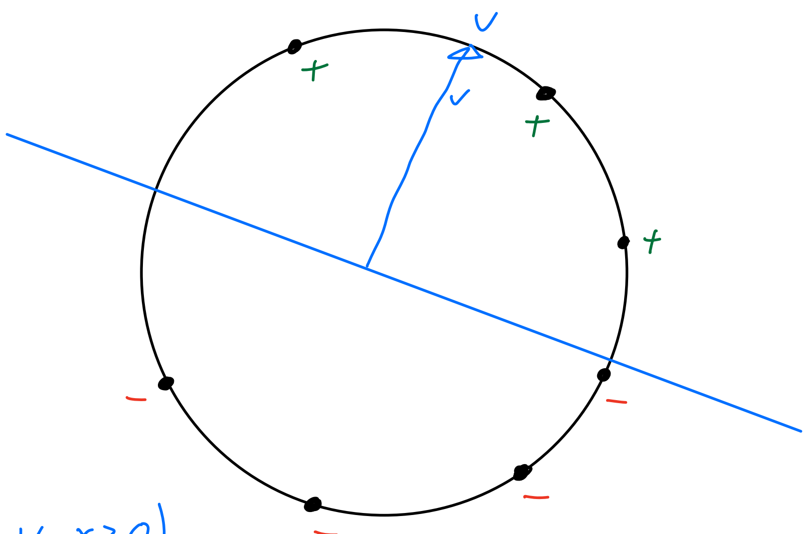
WLOG! For origin-centered LTF data, rescaling doesn't change labels



$v \cdot x \geq 0$

(3) Can assume wlog target vector $v \in \mathbb{R}^n$ has $\|v\| = 1$.

$$3x_1 + 4x_2 \geq 0 \iff \frac{3}{5}x_1 + \frac{4}{5}x_2 \geq 0$$



(to learn $v \cdot x \geq 0$)

Perc. alg: Maintain hyp. vector $w \in \mathbb{R}^n$, initially 0^n

On ex $x \in \mathbb{R}^n$: predict $\mathrm{sign}(w \cdot x)$ $\begin{cases} 1 & w \cdot x \geq 0 \\ -1 & w \cdot x < 0 \end{cases}$

Updates: • right: no update

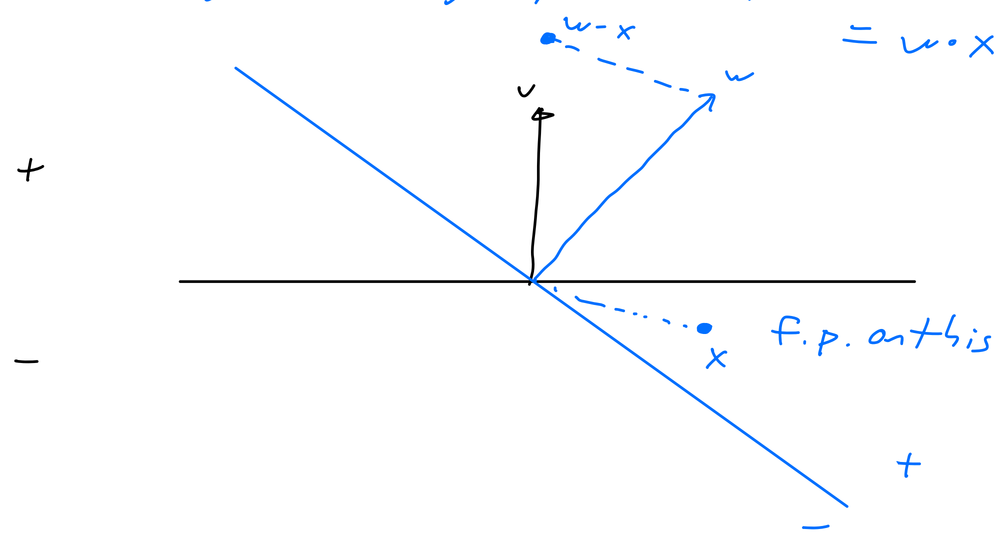
Mistake:

- f.p. $(\overset{+}{w \cdot x \geq 0} \quad \overset{-}{v \cdot x < 0})$:
  set $w \leftarrow w - x$
- f.n. $(w \cdot x < 0 \quad v \cdot x > 0)$:
  set $w \leftarrow w + x$.

$$w \leftarrow w + \underbrace{\mathrm{sign}(v \cdot x)}_{\pm 1} \cdot x$$

Sensible: consider a f.p.. New w is $w - x$

if got x again, $(w-x)\cdot x = w \cdot x - \overbrace{x \cdot x}^{\text{unit vector}}$
$= w \cdot x - 1$



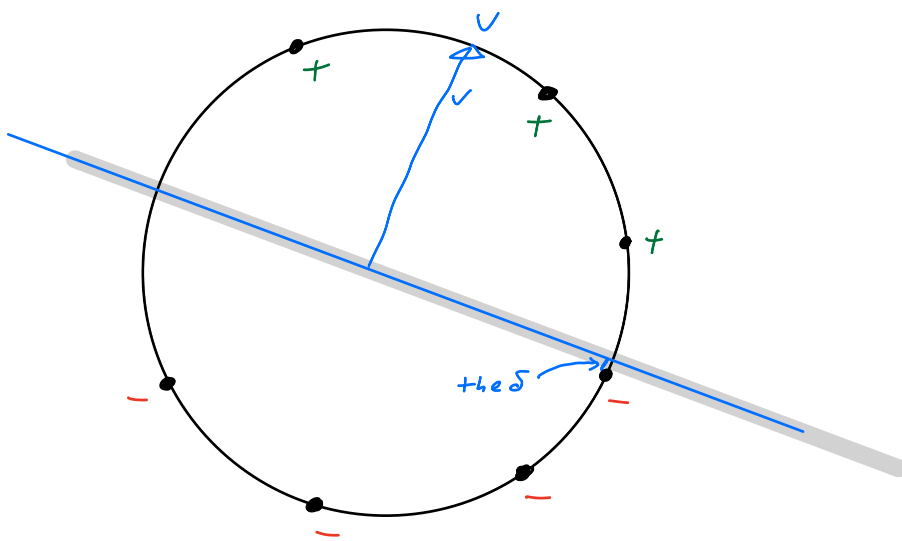
Thm (Perceptron Convergence Theorem)

Sps run Perc alg on seq of ex in $\mathbb{R}^n$ labeled by $v \cdot x \geq 0$ where v, all examples satisfy our assump. above.

since x, v unit vectors, this is dist from x to hyperplane defined by v.

Let $\delta = \min |v \cdot x|$, where min is over all examples x the alg. ever receives.

Then Perc. makes $\leq 1/\delta^2$ mistakes.



Pf: 2 key quantities:

$w \cdot v$ (length of hyp vector w's proj. in "right" direction, v)

+

$\|w\|^2$ (length of hyp vector).

L1: After M mistakes, have
$$w \cdot v \geq \delta \cdot M$$

Pf: w init $\vec{0}$. $w \cdot v = 0$. $M = 0$ ✓

Each mist. causes $w \cdot v$ to go up by $\geq \delta$:

w = old hyp vector before mistake,

$w + \text{sign}(v \cdot x)x$ = new hyp vector after mist. on x.

$$
\begin{aligned}
(w + \text{sign}(v\cdot x)x)\cdot v &= w\cdot v + \text{sign}(v\cdot x)\, v\cdot x \\
&= w \cdot v + |v \cdot x| \\
&\geq w \cdot v + \delta \quad \text{by our def of } \delta.
\end{aligned}
$$

$\blacksquare$

($w = \vec{0}$ initially).

L2: After M mist., have $\|w\|^2 \leq M$.

Pf: Each mist causes $\|w\|^2$ to incr. by ≤ 1.

$$\|w + \text{sign}(v\cdot x)\cdot x\|^2 = (w + \text{sign}(v\cdot x)x)\cdot(w + \text{sign}(v\cdot x)x)$$

$$= w\cdot w + \underbrace{2\,\text{sign}(v\cdot x)\, w\cdot x}_{\text{negative b/c we made mist. on } x.} + \underbrace{\text{sign}(v\cdot x)^2}_{=1}\,\underbrace{x\cdot x}_{=1}$$

$\leq \|w\|^2 + 1.$

$\blacksquare$

Pf: $\forall$ vector w, recall (v unit vector:)

$$w \cdot v = \|w\| \cdot \cos(\text{angle between } w + v)$$

so $w \cdot v \le \|w\|$.

So if M = # of mist. Perc. makes, (1) + (2)

$$\underbrace{\gamma M \le w \cdot v}_{(1)} \le \underbrace{\|w\| \le \sqrt{M}}_{(2)} .$$

$$\gamma M \le \sqrt{M} \implies M \le \frac{1}{\gamma^2} \quad \blacksquare$$

Next: discussion, example, extension to "kernel methods".
