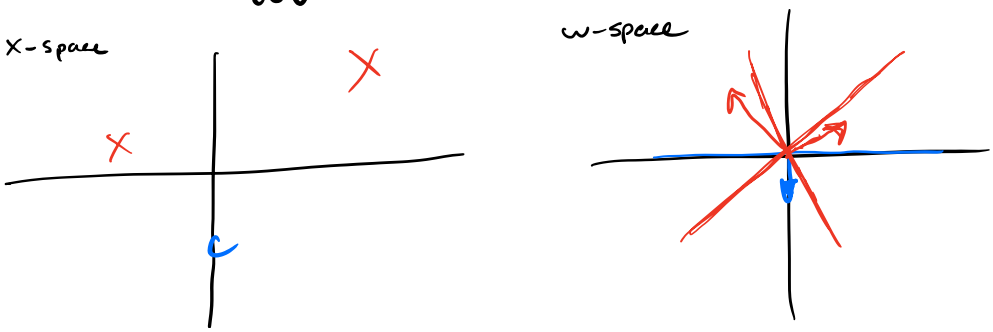
Gradient Descent, stochastic gradient descent, Perceptron learning algorithm. Feature vs. weight space. Max margin classifier, aka hard margin SVM.

Perceptron continued

$$R(w) = \sum_{i \in V} -y_i x_i^T w \qquad V \quad \{x_i : y_i x^T w < 0\}$$



Optimization algorithm: Gradient descent on R:

- Starting point w (not origin), find gradient w.r.t. w
  ↳ step in opposite direction to steepest ascent

$$\nabla R(w) = \sum_{i \in V} \nabla -y_i x_i w = -\sum_{i \in V} y_i x_i$$

$w \leftarrow$ non zero starting point (use a sample point)
while $R(w) > 0$:
  $V \leftarrow$ i's for $y_i x_i^T w \leq 0$
  $w \leftarrow w + \varepsilon \sum_{i \in V} y_i x_i$

$\varepsilon > 0$ is the step size aka learning rate

Slow! $O(nd)$ time!!!

Optimization alg. 2: stochastic gradient descent

- each step, pick one misclassified X_i
- do gradient descent on loss function of $L(X_i^T w, y_i)$
- "perceptron alg" $O(d)$ time

while some $y_i X_i^T w < 0$

$w \leftarrow w + \varepsilon y_i X_i$

return w

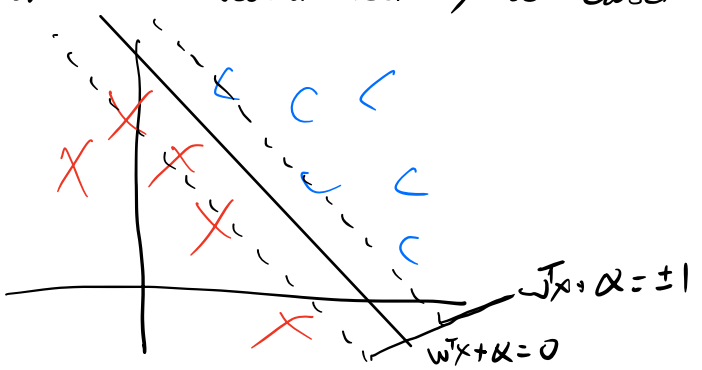
what if $\alpha \neq 0$

Add a fictitious dimension: $\mathbb{R}^{d+1}$ $[X_i]_{d+1} = 1$

$$f(x) = w^T x + \alpha = [w \ \alpha] \begin{bmatrix} x \\ 1 \end{bmatrix}$$

MAX MARGIN CLASSIFIER

margin = dist from decision boundary to closet point



$$y_i (w^T x_i + \alpha) \geq 1 \quad \text{for } i \in [1, n]$$

$\|w\|_2 = 1$, signed dist to X_i is $w^T X_i + \alpha$

otherwise, $\frac{w^T}{\|w\|_2} X_i + \frac{\alpha}{\|w\|_2}$

margin is $\min_i \frac{1}{\|w\|_2} \left| w^T x_i + \alpha \right| \geq \frac{1}{\|w\|_2}$

To max margin, min w:

QP $\rightarrow$ $\min_w \|w\|_2^2$: $y_i (x_i^T w + \alpha) \geq 1$ for $i = 1 \ldots n$

"hard margin SVM" margin $\frac{1}{\|w\|_2}$ so $\frac{2}{\|w\|_2}$