CS 189 Introduction to Machine Learning
Spring 2023 Jonathan Shewchuk **DIS1**

In this discussion, we'll review linear classifiers and develop some intuition for the hard-margin support vector machine (SVM) optimization problem:

$$\boxed{\min_{w,\alpha} \ \|w\|^2 \ \text{ subject to } \ y_i(X_i \cdot w + \alpha) \geq 1, \ \ \forall i \in \{1, \ldots, n\}.}$$

1 Linear Decision Rules

A *decision rule* is a function $r : \mathbb{R}^d \to \pm 1$ that maps a feature vector (test point) to +1 ("in class") or −1 ("not in class"). Many classifiers compute a *decision function*, f, which is also known as a *predictor function* or *discriminant function*. The decision rule for the classifier is then defined as

$$\boxed{r(x) = \begin{cases} +1 & \text{if } f(x) \geq 0 \\ -1 & \text{otherwise} \end{cases}.}$$

The *decision boundary* is the boundary chosen by the classifier to separate items in different classes. For a decision function, f, the decision boundary is

$$\boxed{\{x \in \mathbb{R}^d : f(x) = 0\}.}$$

For a d-dimensional feature space, linear decision functions have the form

$$\boxed{f(x) = x \cdot w + \alpha,}$$

where $w \in \mathbb{R}^d$ is a weight vector and $\alpha \in \mathbb{R}$ is a bias term. For this linear decision function, the decision boundary is the hyperplane

$$\boxed{\mathcal{H} = \{x \in \mathbb{R}^d : x \cdot w + \alpha = 0\}.}$$

(a) Draw a figure depicting the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^2 : x \cdot w + \alpha = 0\}$ with $w = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ and $\alpha = -4$. Include in your figure the vector w, drawn relative to $\mathcal{H}$.

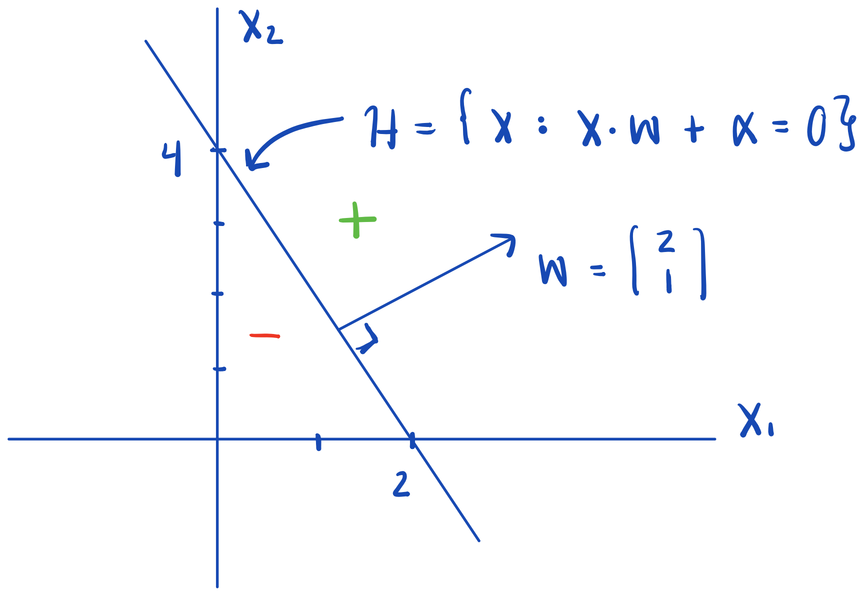
(b) Indicate in your figure the region in which data points would be classified as +1 (in class). Do the same for data points that would be classified as −1 (not in class).

$$X \cdot W + \alpha = 0$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 \\ 1 \end{bmatrix} - 4 = 0$$

$$2x_1 + x_2 - 4 = 0$$

$$x_2 = -2x_1 + 4$$



2 Maximum Margin Classifier

Consider a data set of n d-dimensional sample points, $\{X_1, \ldots, X_n\}$. Each sample point, $X_i \in \mathbb{R}^d$, has a corresponding label, y_i, indicating to which class that point belongs. For now, we will assume that there are only two classes and that every point is either in the given class ($y_i = 1$) or not in the class ($y_i = -1$). Consider the linear decision boundary defined by the hyperplane

$$\mathcal{H} = \{x \in \mathbb{R}^d : x \cdot w + \alpha = 0\}.$$

The *maximum margin classifier* maximizes the distance from the linear decision boundary to the closest training point on either side of the boundary, while correctly classifying all training points.

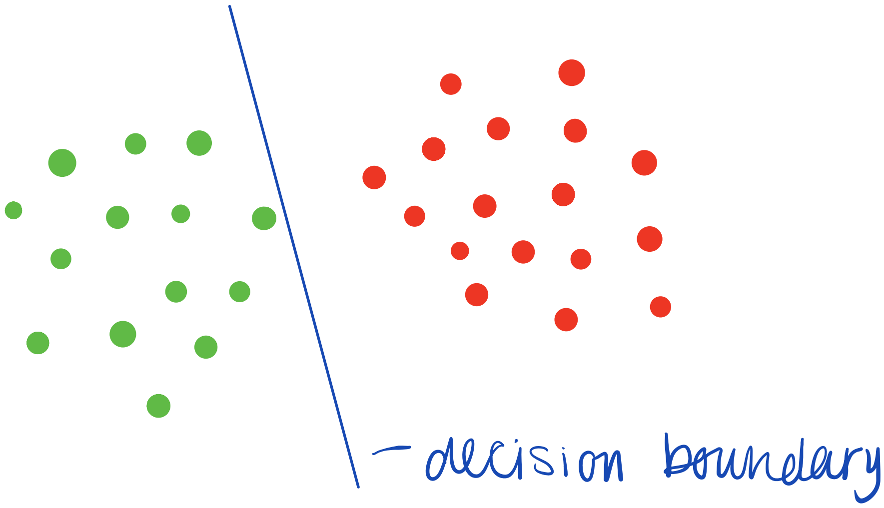
(a) An in-class sample point is correctly classified if it is on the positive side of the decision boundary, and an out-of-class sample is correctly classified if it is on the negative side. Write a set of n constraints to ensure that all n points are correctly classified.

(b) The maximum margin classifier aims to maximize the distance from the training points to the decision boundary. Derive the distance from a point X_i to the hyperplane $\mathcal{H}$.

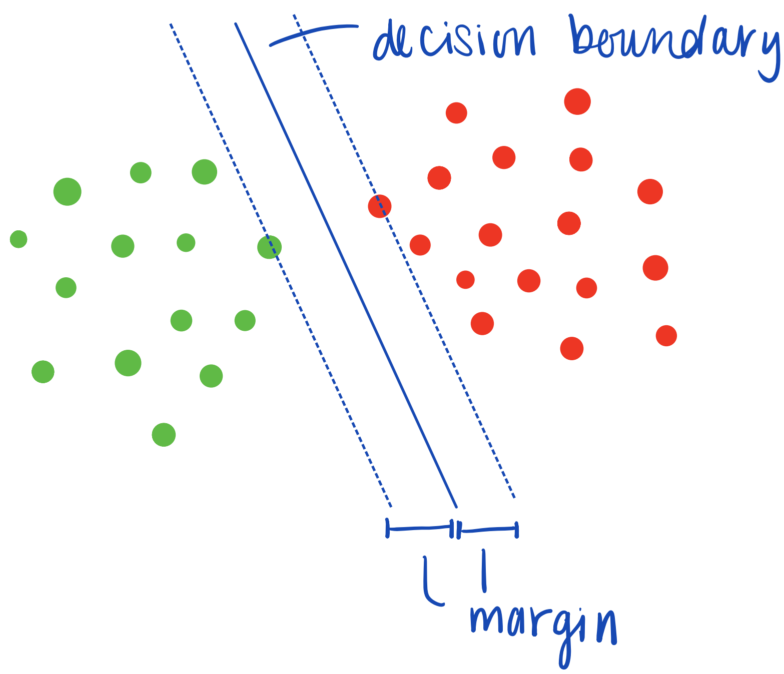
(c) Assuming all the points are correctly classified, write an inequality that relates the distance of sample point X_i to the hyperplane $\mathcal{H}$ in terms of only the normal vector w.

(d) For the maximum margin classifier, the training points closest to the decision boundary on either side of the boundary are referred to as *support vectors*. What is the distance from any support vector to the decision boundary?

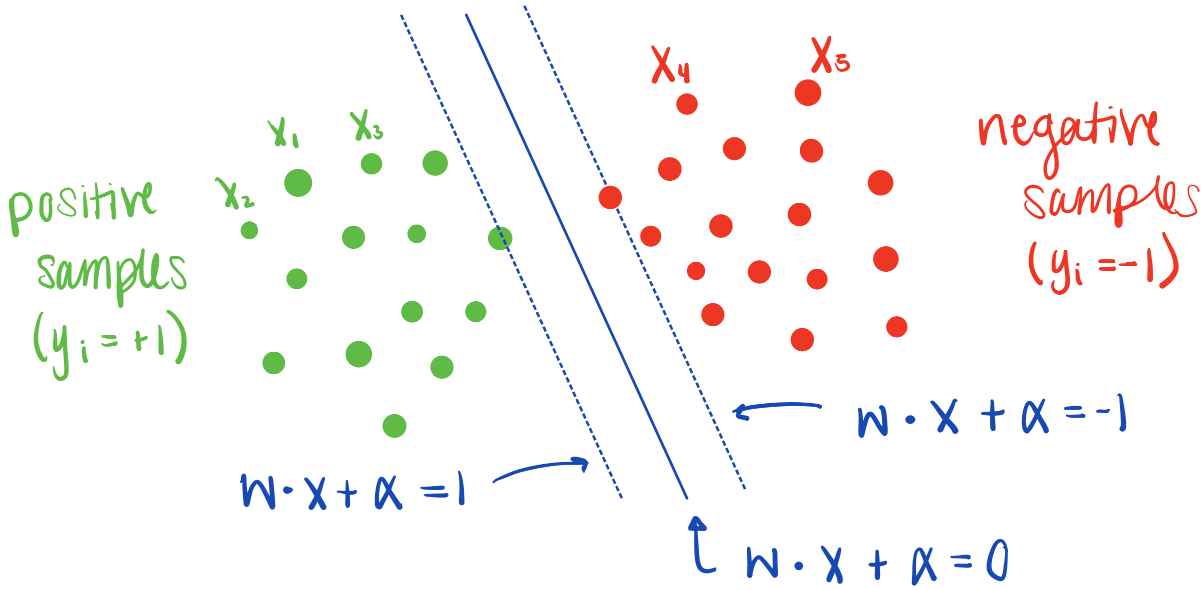
(e) Using the previous parts, write an optimization problem for the maximum margin classifier.



This linear decision boundary correctly classifies all the pts but is <u>not</u> a maximum margin classifier.



This is a maximum margin classifier!



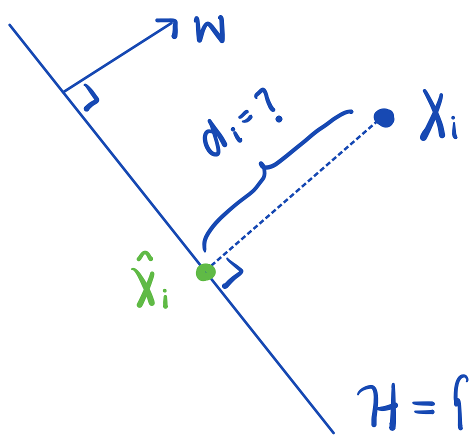
a) If $y_i = 1$, we want $x_i \cdot w + \alpha \geq 1$
If $y_i = -1$, we want $x_i \cdot w + \alpha \leq -1$

We can combine these constraints:

$y_i(x_i \cdot w + \alpha) \geq 1, \quad i = 1, \ldots, n$

└ # of sample pts

b)



$$d_i := \| x_i - \hat{x}_i \|_2$$

$$\mathcal{H} = \{ x : x \cdot w + \alpha = 0 \}$$

1) $\hat{x}_i$ lies on the hyperplane :

$$\hat{x}_i \cdot w + \alpha = 0$$

2) $(x_i - \hat{x}_i) \perp \mathcal{H}$:

$$(x_i - \hat{x}_i) = \eta w \quad \text{for some } \eta \in \mathbb{R}$$

$$d_i = \| \eta w \|_2 = |\eta| \, \| w \|_2 \, , \qquad \eta = ?$$

$$(x_i - \hat{x}_i) \cdot w = (\eta w) \cdot w$$

$$x_i \cdot w - \hat{x}_i \cdot w = \eta \| w \|_2^2$$

$$x_i \cdot w + \alpha = \eta \| w \|_2^2$$

$$\eta = \frac{|x_i \cdot w + \alpha|}{\| w \|_2^2}$$

$$d_i = \frac{|x_i \cdot w + \alpha|}{\| w \|_2^2} \| w \|_2 = \frac{|x_i \cdot w + \alpha|}{\| w \|_2}$$

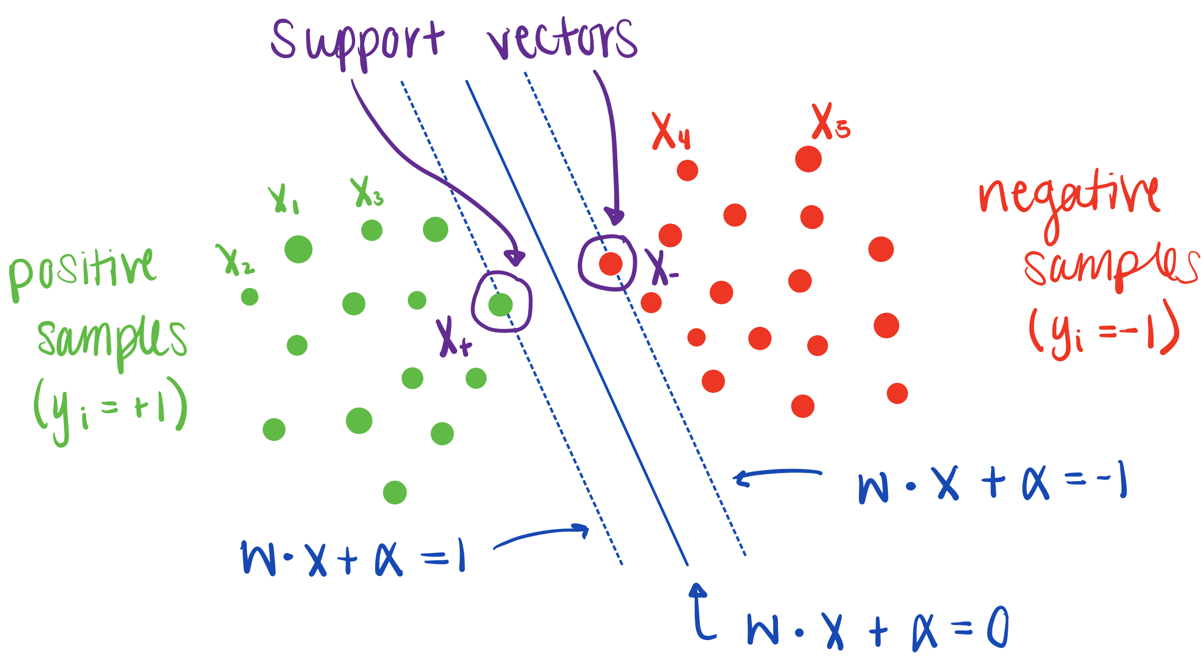
c) (a) $\rightarrow$ $y_i(X_i \cdot w + \alpha) \geq 1$

(b) $\rightarrow$ $d_i = \dfrac{|X_i \cdot w + \alpha|}{\|w\|_2}$

If $y_i \in \{-1, 1\}$, then (a) $\rightarrow$ $|X_i \cdot w + \alpha| \geq 1$

$$d_i \geq \frac{1}{\|w\|_2}$$

d)



$$\begin{aligned} X_+ \cdot w + \alpha &= 1 \\ X_- \cdot w + \alpha &= -1 \end{aligned} \rightarrow |X_i \cdot w + \alpha| = 1 \text{ if } X_i \text{ is a support vector}$$

$$\downarrow$$

$$d_i = \frac{1}{\|w\|_2}$$

e) <u>Goal</u>: Choose w & α to maximize d_i across all training points, while ensuring all pts are correctly classified

$$\max \sum_i d_i \equiv \max \frac{1}{\|w\|_2} \equiv \min \|w\|_2 \equiv \min \|w\|_2^2$$

└ part (b), (c) └ convex opt. ┘

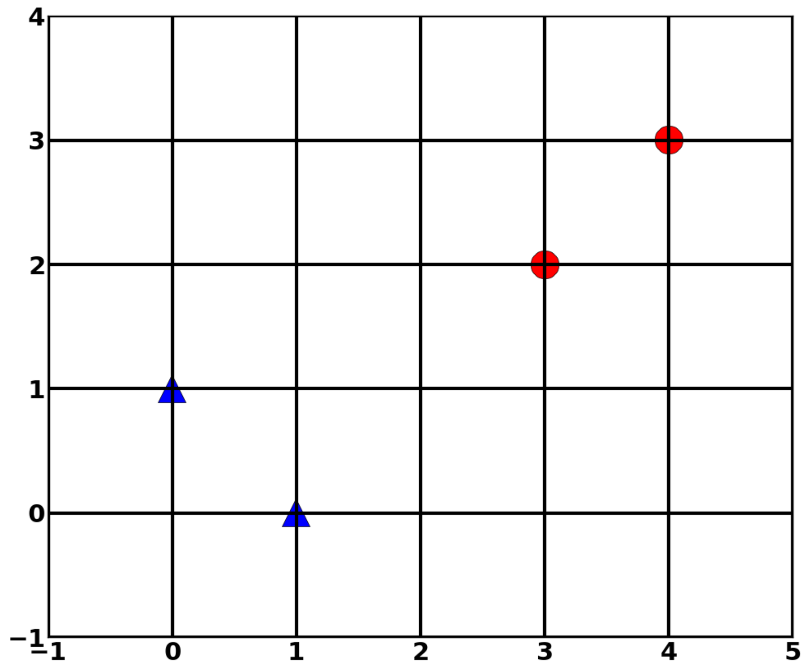
$$\min_{w,\alpha} \|w\|_2^2$$

$$\text{s.t. } y_i(X_i \cdot w + \alpha) \geq 1, \quad i = 1, \ldots, n$$

└ part (a)

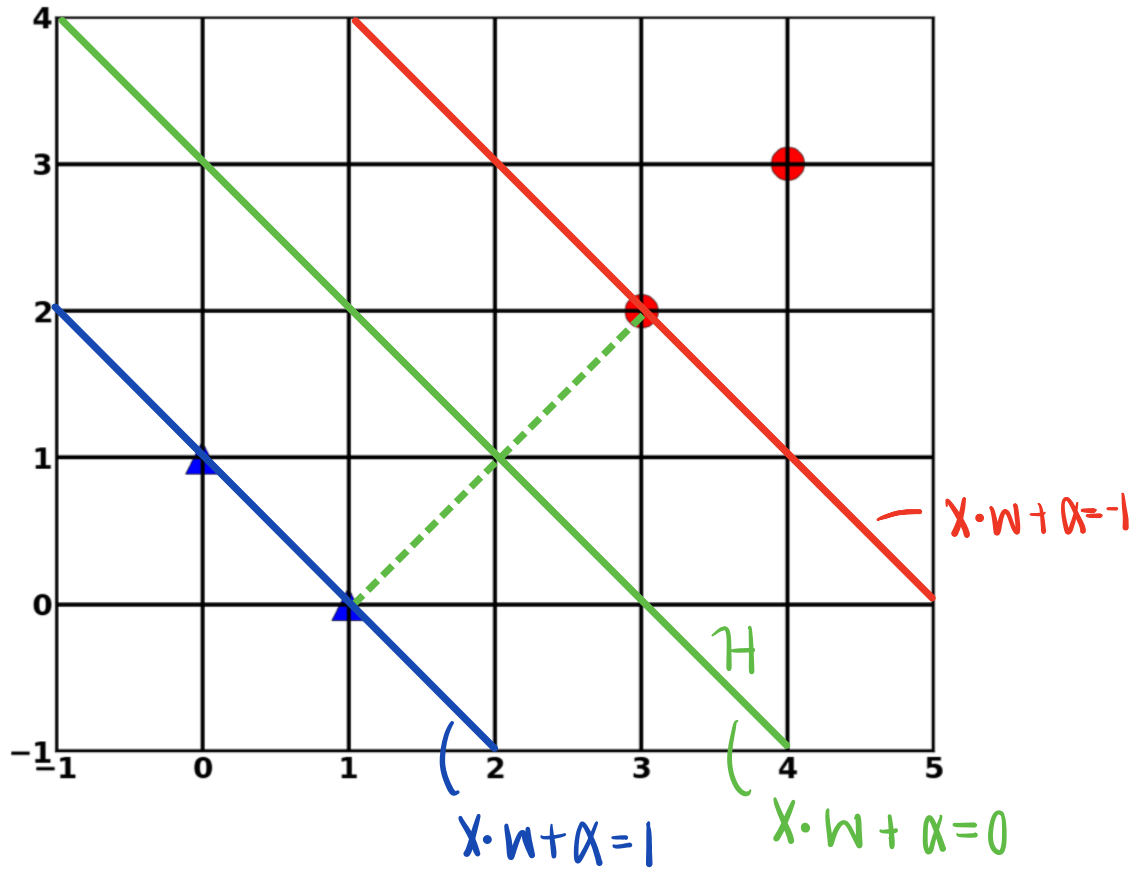
3 Hard-Margin SVM by Hand

You are given the sample points shown in the figure below. The blue triangles are positive samples (in the given class), and the red circles are negative examples (not in the given class).



Find (by hand) the equation of the hyperplane $\mathcal{H} = \{x \in \mathbb{R}^2 : x \cdot w + \alpha = 0\}$ that a hard-margin SVM classifier would learn. Draw the decision boundary and its margins.

The hard-margin SVM classifier is a maximum margin classifier!



$d = 2$ so we have 3 unknowns: w_1, w_2, α

$\therefore$ we need to obtain 3 lin. ind. eq'ns

1) $(2,1)$ is on the hyperplane $\rightarrow$

$$2w_1 + 1w_2 + \alpha = 0$$

2) $(0,1)$ is a positive support vector $\rightarrow$

$$0w_1 + 1w_2 + \alpha = 1$$

3) $(3,2)$ is a negative support vector $\rightarrow$

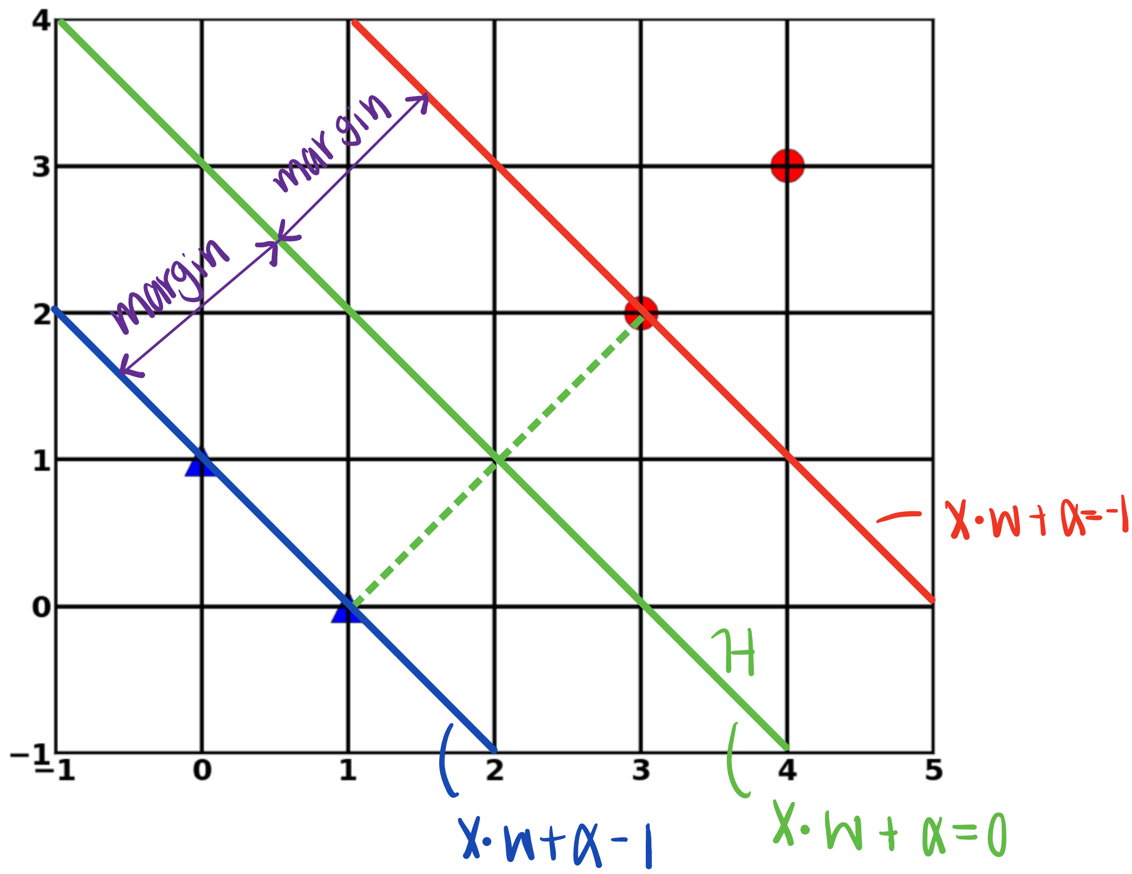
$$3w_1 + 2w_2 + \alpha = -1$$

$$\alpha = 1 - w_2 \rightarrow \alpha = \frac{3}{2}$$

$$2w_1 + w_2 + (1 - w_2) = 0 \rightarrow w_1 = -\frac{1}{2}$$

$$3\left(-\frac{1}{2}\right) + 2w_2 + (1 - w_2) = -1 \rightarrow w_2 = -\frac{1}{2}$$

$$w = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \quad \alpha = \frac{3}{2}$$



$$\text{margin} = \frac{1}{\|w\|_2}$$

$$\|w\|_2^2 = w \cdot w = \begin{bmatrix} -\frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2} \end{bmatrix} = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

$$\|w\|_2 = 1/\sqrt{2} \quad , \quad 1/\|w\|_2 = \sqrt{2}$$