# Applied Machine Learning

Dr. Harikrishnan N B
Computer Science and Information Systems

BITS Pilani
Pilani Campus

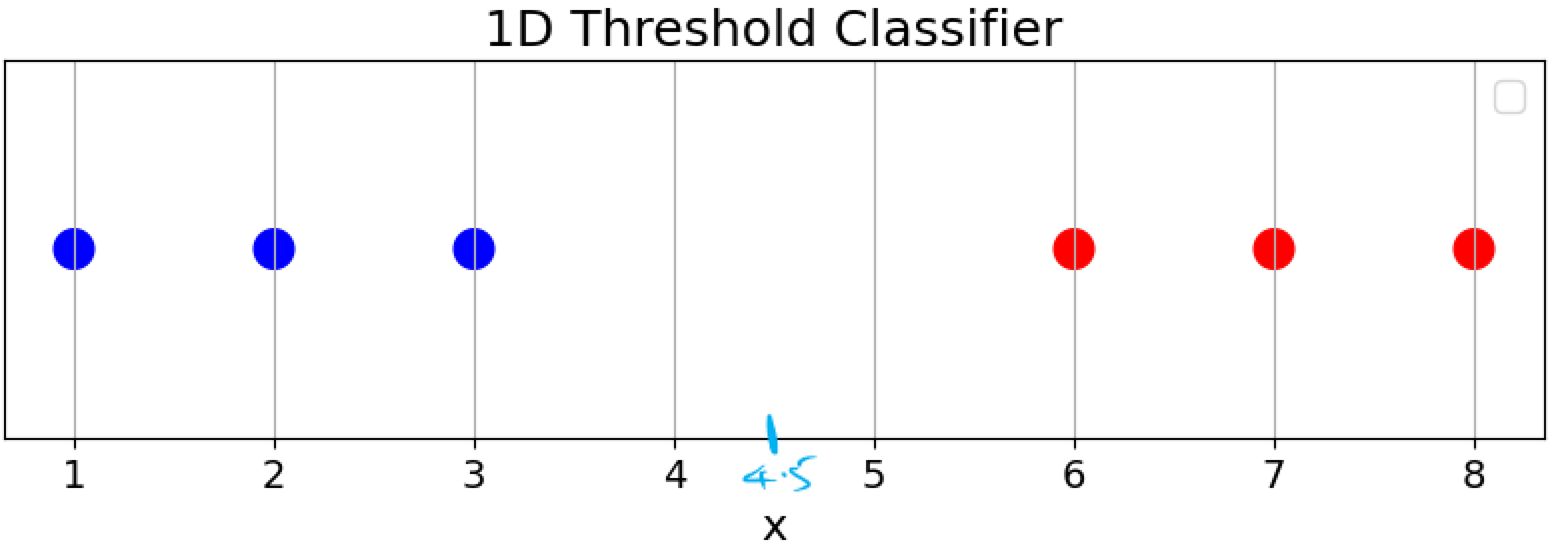
innovate achieve lead

# SE ZG568 / SS ZG568, Applied Machine Learning

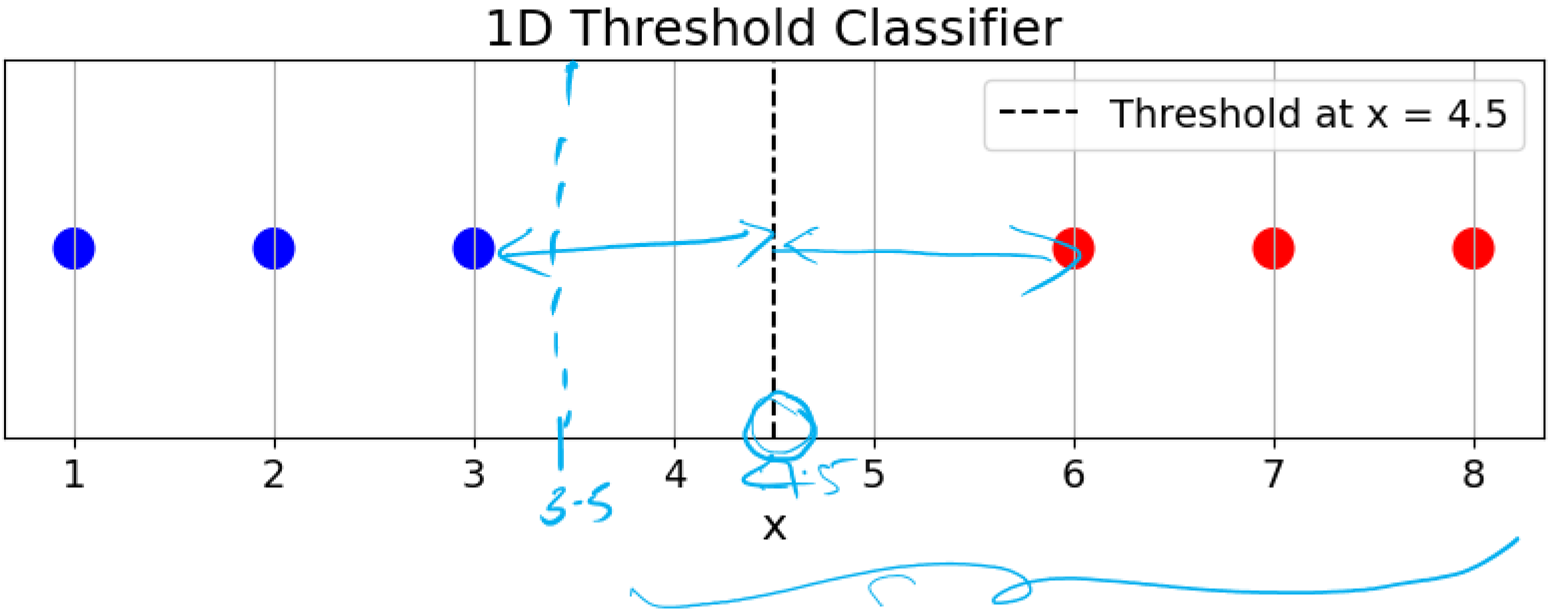
# Lecture No. 13 [27 April 2025]

# Support Vector Machine

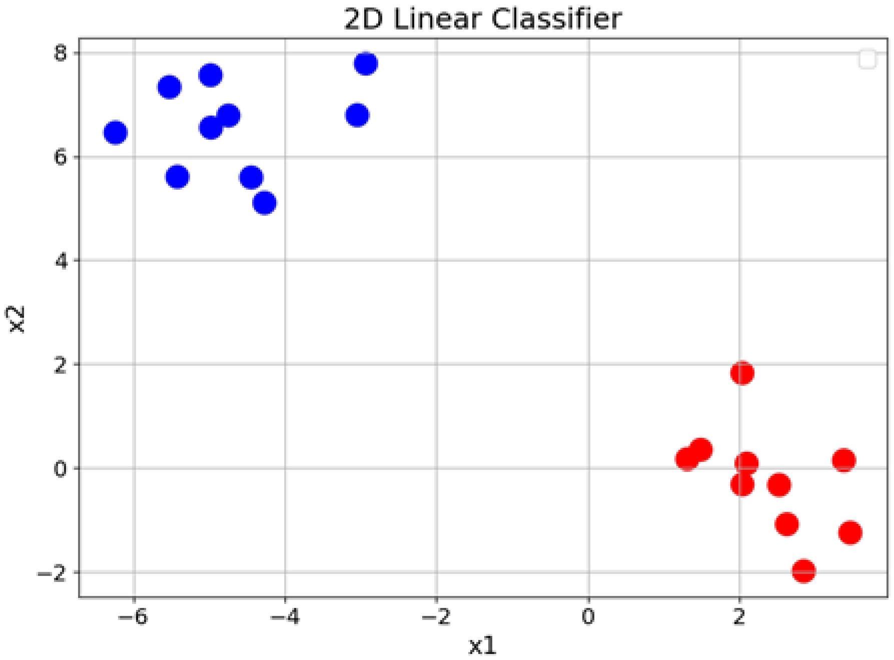
Points lying on x-axis: Come up with a classification rule!!!



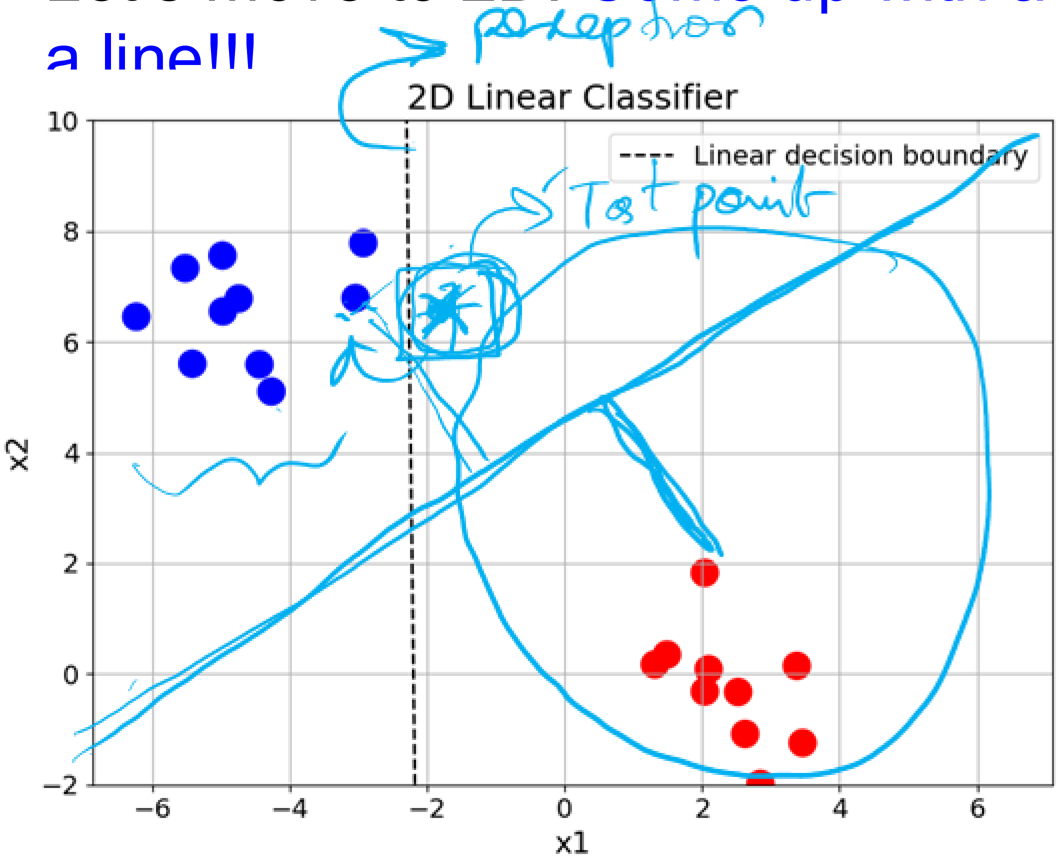
Points lying on x-axis: Come up with a classification rule!!!



Let’s move to 2D: Come up with a classification rule using a line!!!

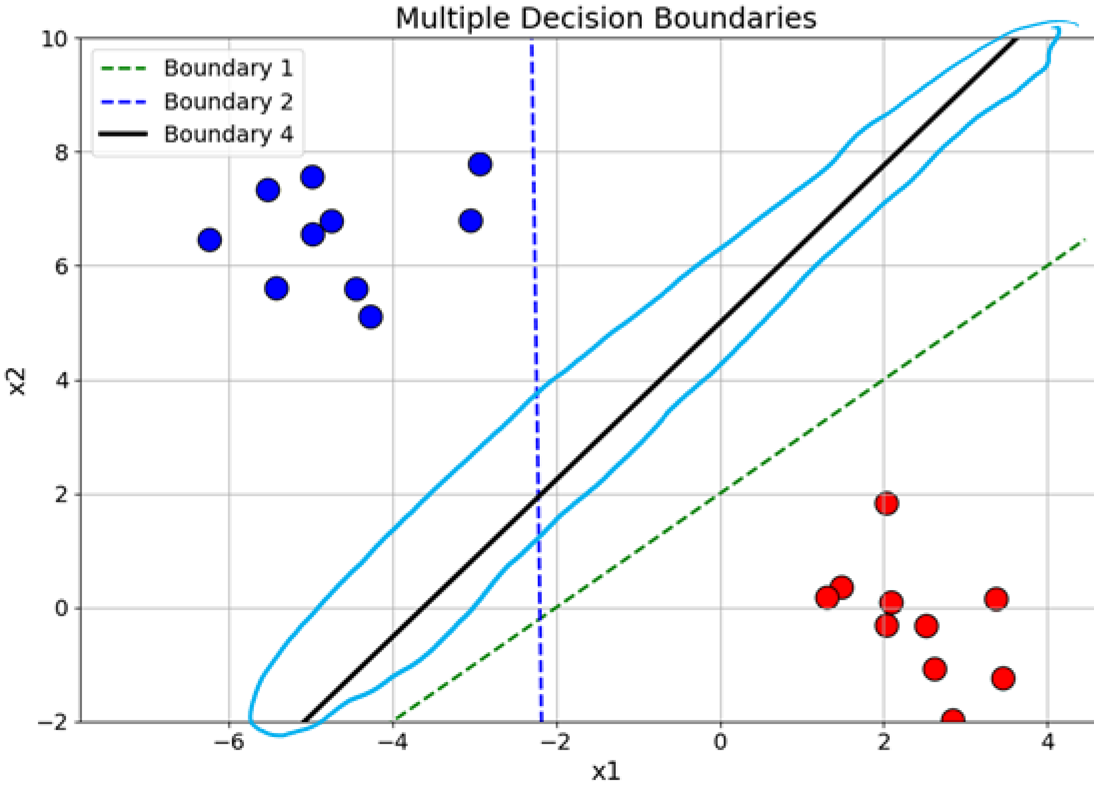


# Let's move to 2D: Come up with a classification rule using a line!!!



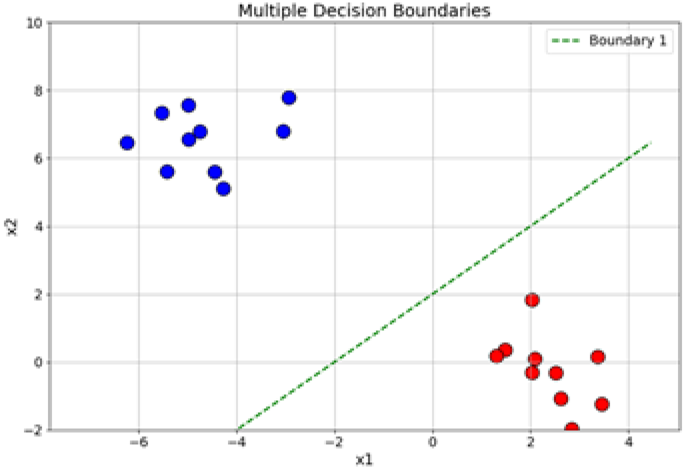

$x_2 = -100x_1 - 220$

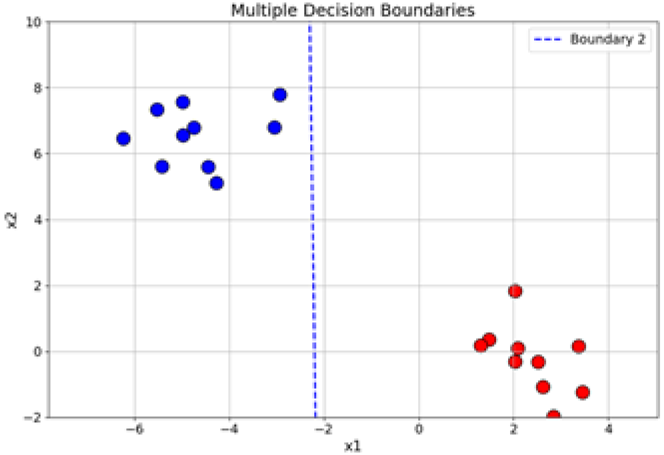
# This is not the only classification rule!!!

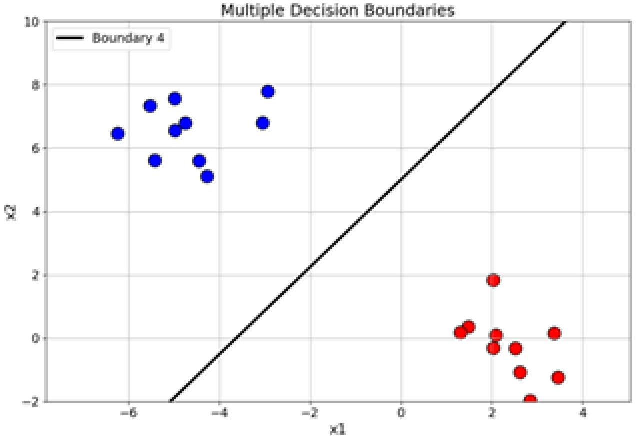



Which Decision Boundary will you choose?

# Which decision boundary would you choose?



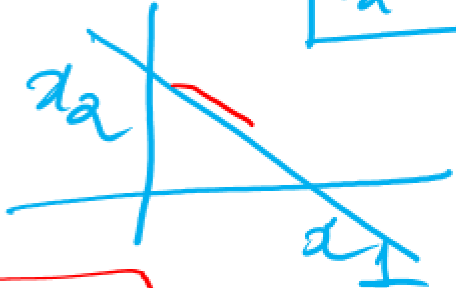



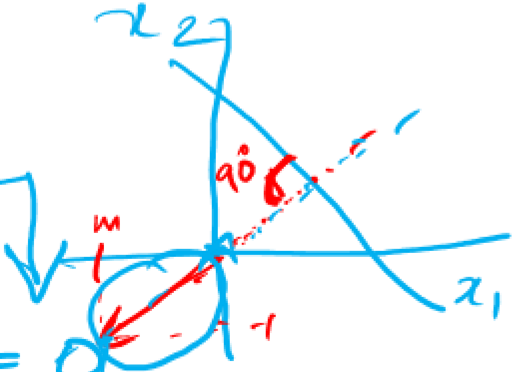
# Basics of Equation of Line

$x_2, x_1$

$$\boxed{x_2 = m x_1 + b}$$




$$m x_1 - x_2 + b = 0$$



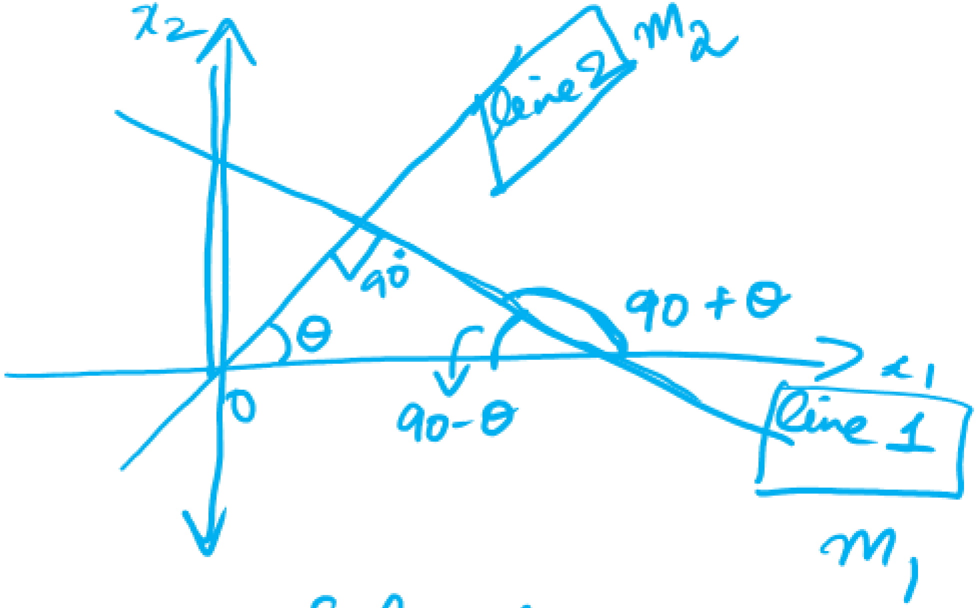

$(m, -1)$

$$\boxed{w = \begin{bmatrix} m \\ -1 \end{bmatrix}}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$\boxed{w}^T x + b = 0 \text{ ?}$$

What is the direction of $w$?




For line 1

$$m_1 = \tan(90+\theta)$$

$$m_2 = \tan\theta$$

$m_1$ : slope of line 1

$m_2$ : slope of line 2

$$\boxed{m_1 \times m_2}$$

$$m_1 \times m_2$$

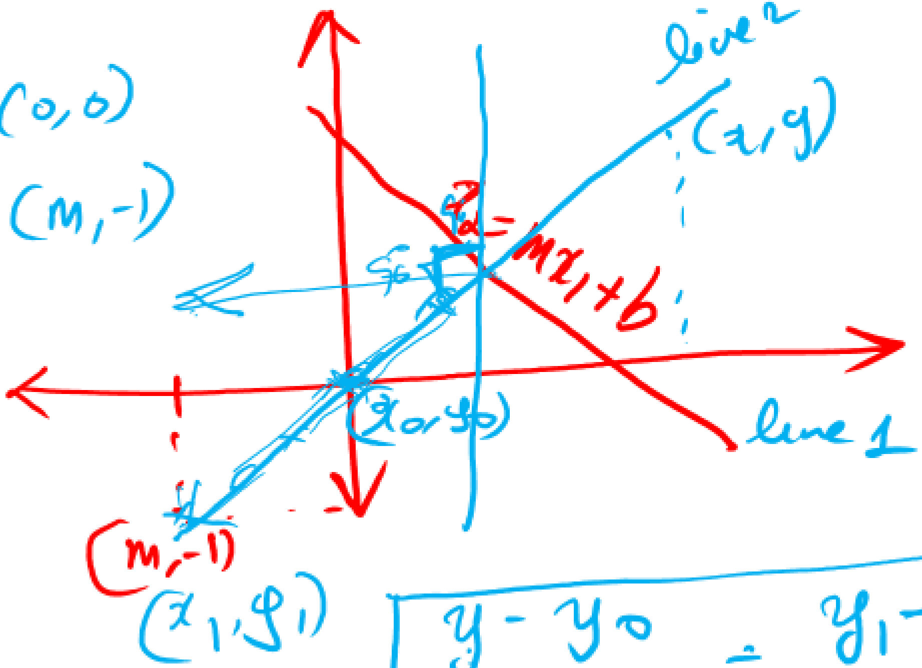
$$\tan(90+\theta) \times \tan\theta$$

$$\frac{\sin(90+\theta)}{\cos(90+\theta)} \times \frac{\sin\theta}{\cos\theta}$$

$$\frac{\cos\theta}{-\sin\theta} \times \frac{\sin\theta}{\cos\theta} = \underline{\underline{-1}}$$

$P_1\ (0,0)$

$P_2\ (m,-1)$




line 1 : $x_2 = m x_1 + b$

line 2 :

$$\boxed{x_2 = -\frac{x_1}{m}}$$

Slope of L1: $m$

Slope of L2 $= -\frac{1}{m}$

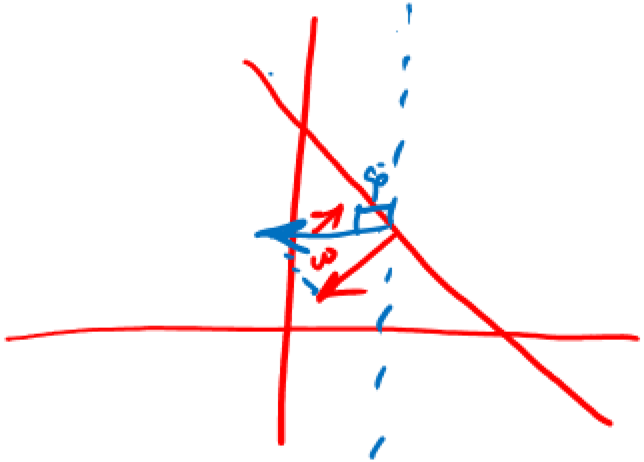
$m_1 m_2 = -1$

$$\boxed{\frac{y-y_0}{x-x_0} = \frac{y_1-y_0}{x_1-x_0}} \Big\} \text{ constant}$$

$$\frac{y-0}{x-0} = \frac{-1-0}{m-0} \Rightarrow \boxed{y = -\frac{x}{m}}$$

$$\boxed{\begin{aligned} y &= x_2 \\ x &= x_1 \end{aligned}}$$

what if I rotate my $\vec{\omega}$ ?



# Perceptron

| | |
|---|---|
| Initialize $\vec{w} = \vec{0}$ | // Initialize $\vec{w}$. $\vec{w} = \vec{0}$ misclassifies everything. |
| **while** TRUE **do** | // Keep looping |
|     $m = 0$ | // Count the number of misclassifications, $m$ |
|     **for** $(x_i, y_i) \in D$ **do** | // Loop over each (data, label) pair in the dataset, $D$ |
|         **if** $y_i(\vec{w}^T \cdot \vec{x_i}) \leq 0$ **then** | // If the pair $(\vec{x_i}, y_i)$ is misclassified |
|             $\vec{w} \leftarrow \vec{w} + y\vec{x}$ | // Update the weight vector $\vec{w}$ |
|             $m \leftarrow m + 1$ | // Counter the number of misclassification |
|         **end if** | |
|     **end for** | |
|     **if** $m = 0$ **then** | // If the most recent $\vec{w}$ gave 0 misclassifications |
|         break | // Break out of the while-loop |
|     **end if** | |
| **end while** | // Otherwise, keep looping! |

Initialize $\vec{w} = \vec{0}$
**while** TRUE **do**
    $m = 0$
    **for** $(x_i, y_i) \in D$ **do**
        **if** $y_i(\vec{w}^T \cdot \vec{x}_i) \leq 0$ **then**
            $\vec{w} \leftarrow \vec{w} + y\vec{x}$
            $m \leftarrow m + 1$
        **end if**
    **end for**
    **if** $m = 0$ **then**
        break
    **end if**
**end while**

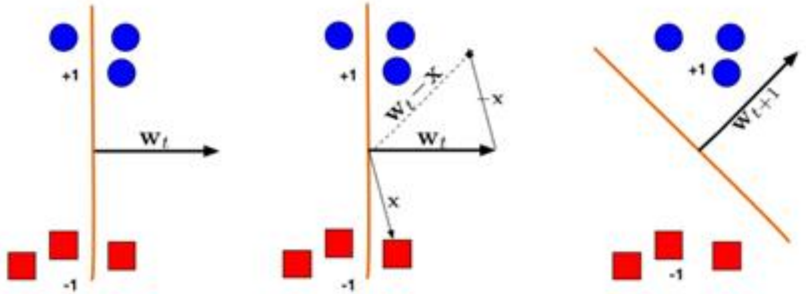
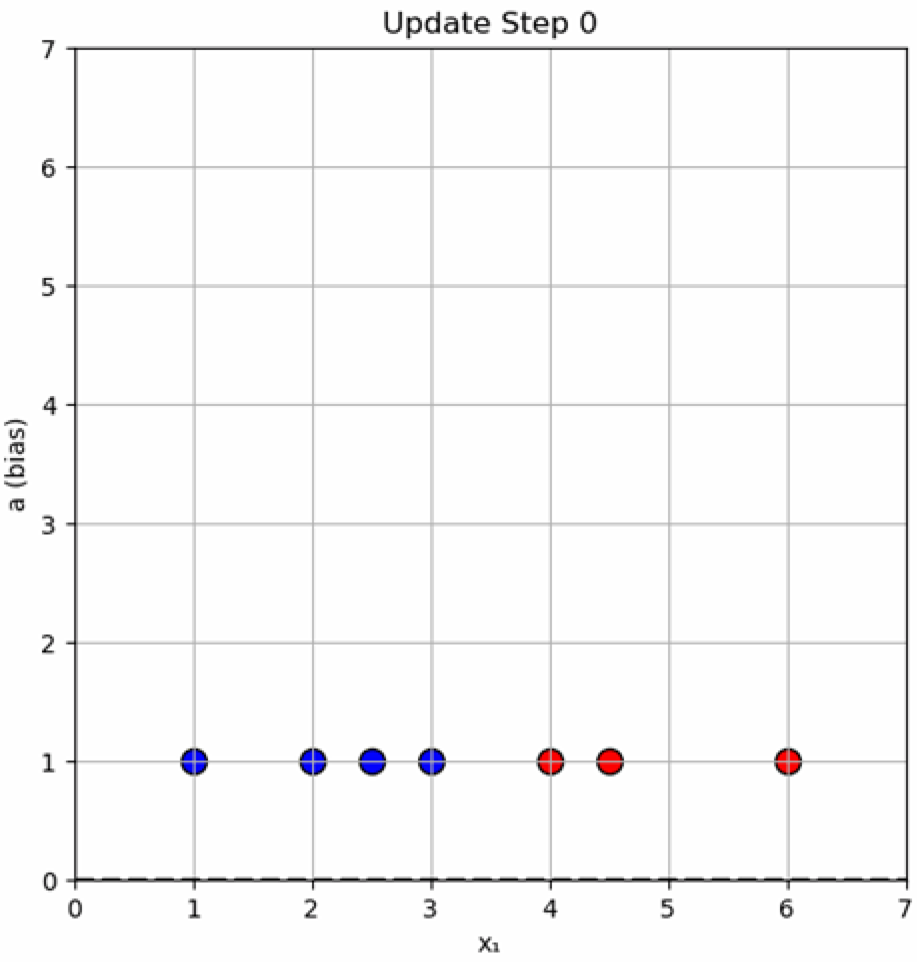
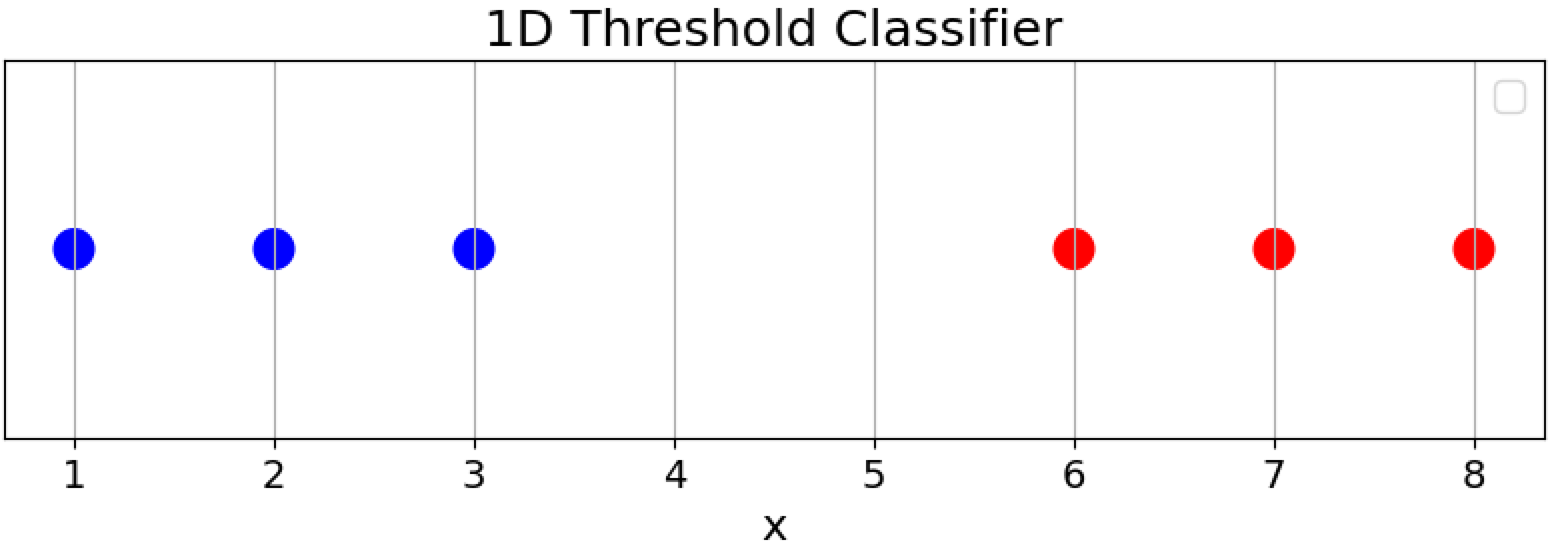



*Illustration of a Perceptron update. (Left:) The hyperplane defined by $\mathbf{w}_t$ misclassifies one red (-1) and one blue (+1) point. (Middle:) The red point $\mathbf{x}$ is chosen and used for an update. Because its label is -1 we need to* ***subtract*** *$\mathbf{x}$ from $\mathbf{w}_t$. (Right:) The udpated hyperplane $\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{x}$ separates the two classes and the Perceptron algorithm has converged.*



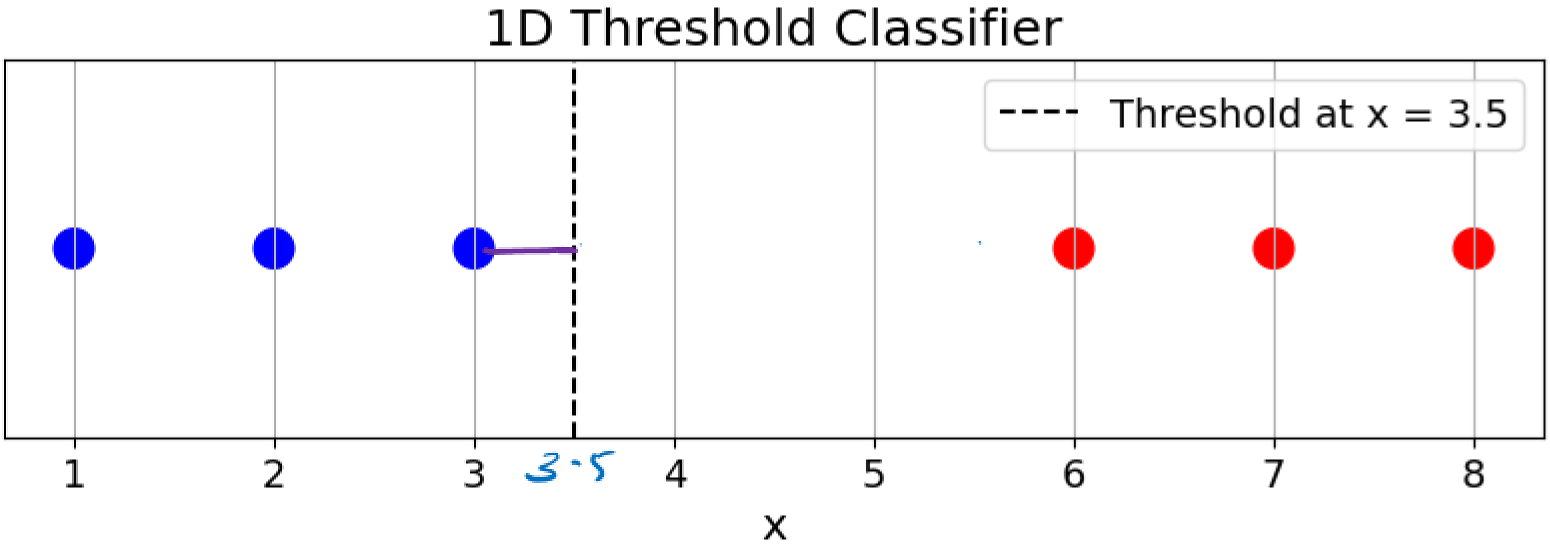
Update Step 0
a (bias)
x₁
0
1
2
3
4
5
6
7

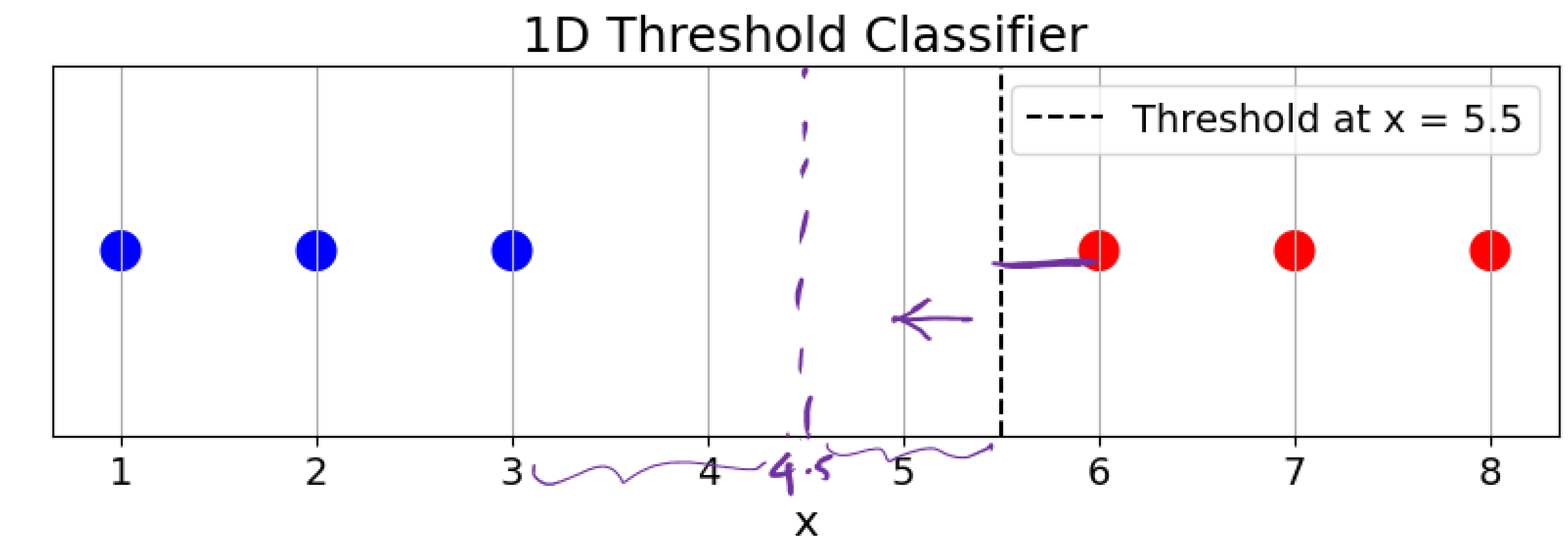
Points lying on x-axis: Come up with a classification rule!!!



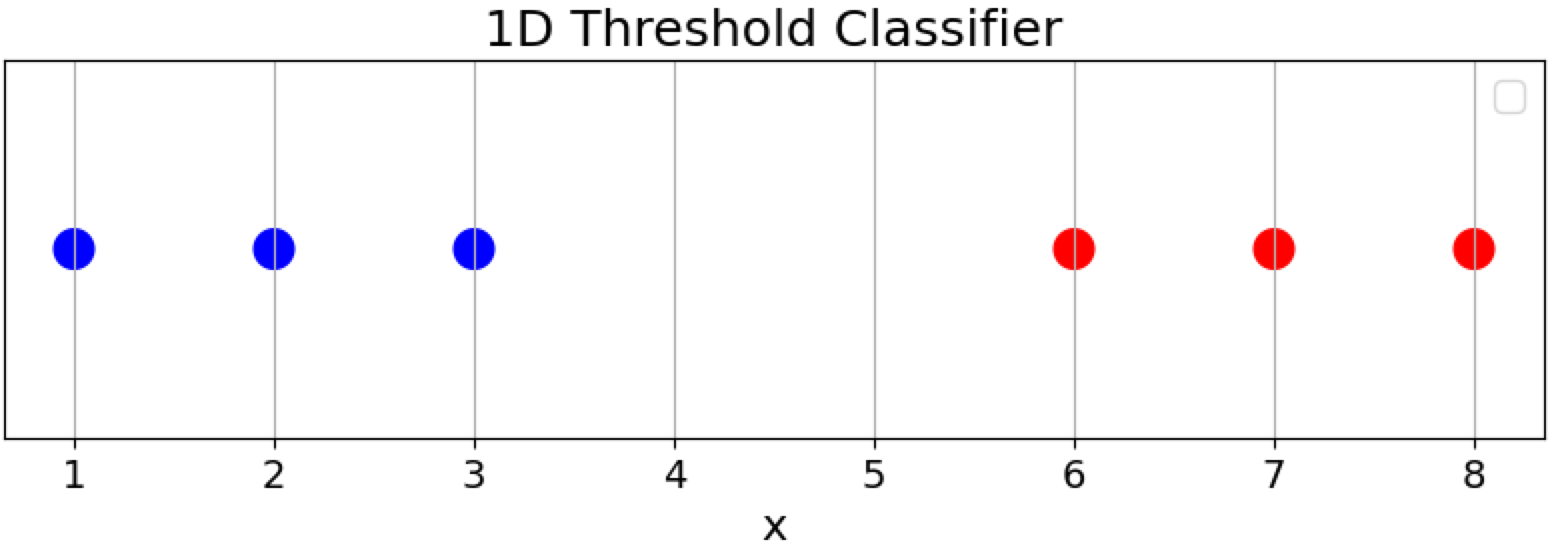
# What is Margin?

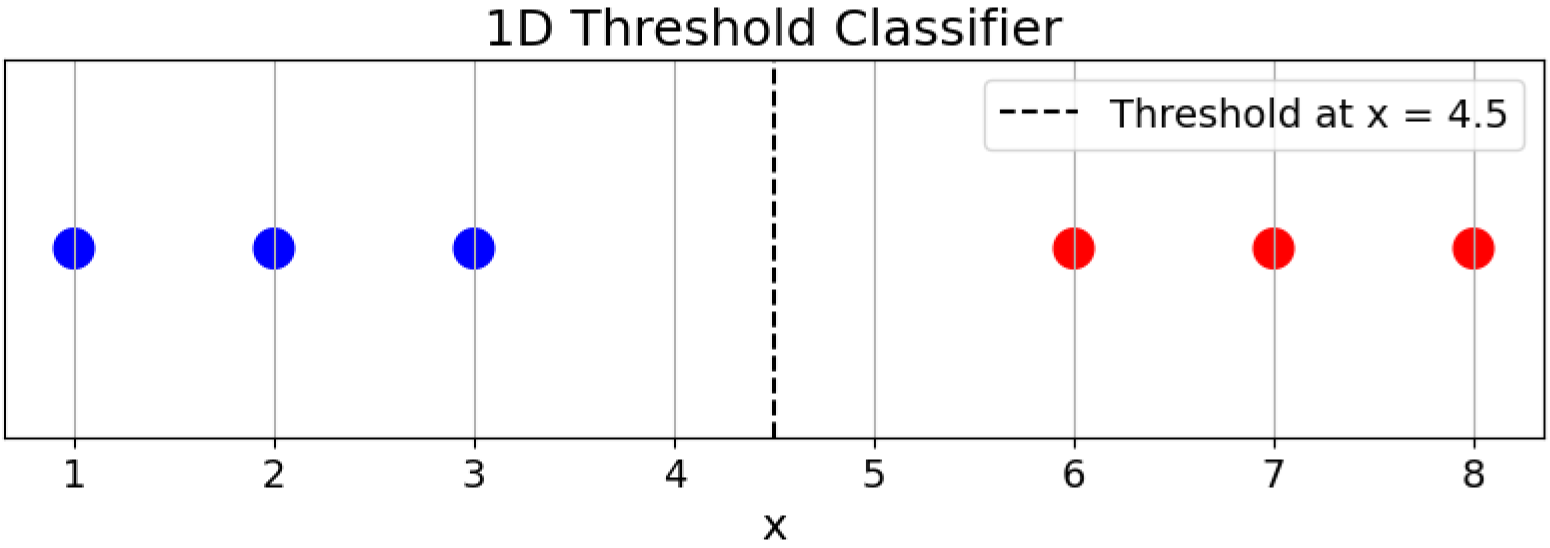
Shortest distance of the data observation from the line.



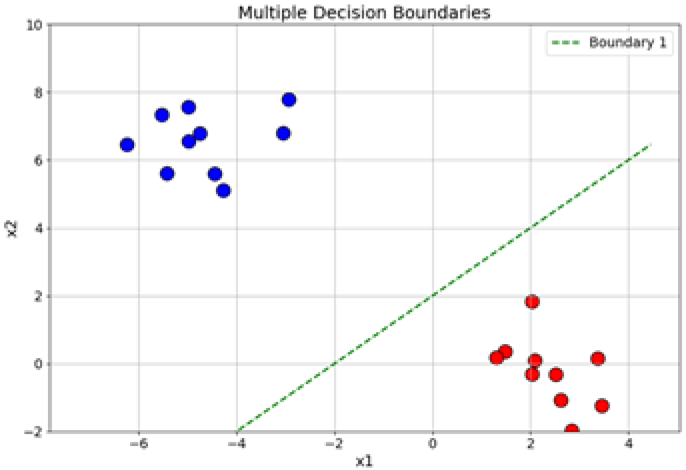

1D Threshold Classifier
Threshold at x = 5.5
1
2
3
4
4.5
5
6
7
8
x

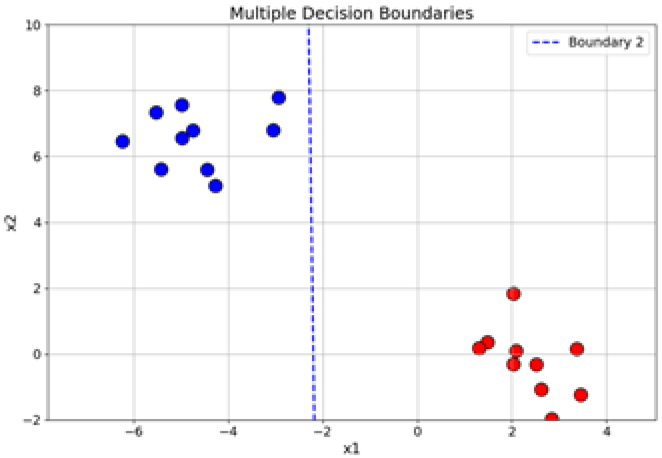
# Which threshold will give you maximum margin?

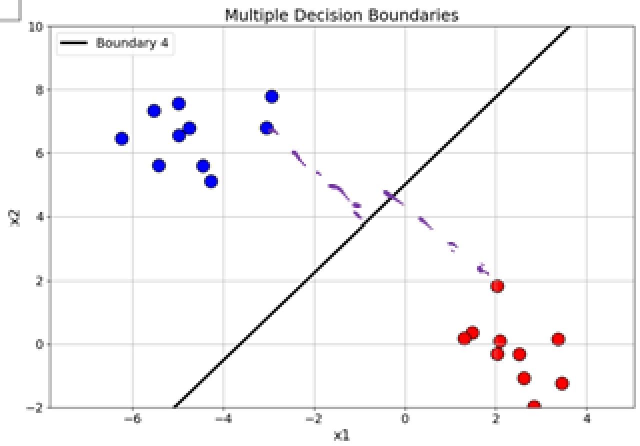



1D Threshold Classifier
Threshold at x = 4.5
1
2
3
4
5
6
7
8
x

Which decision boundary would give you **maximum margin?**



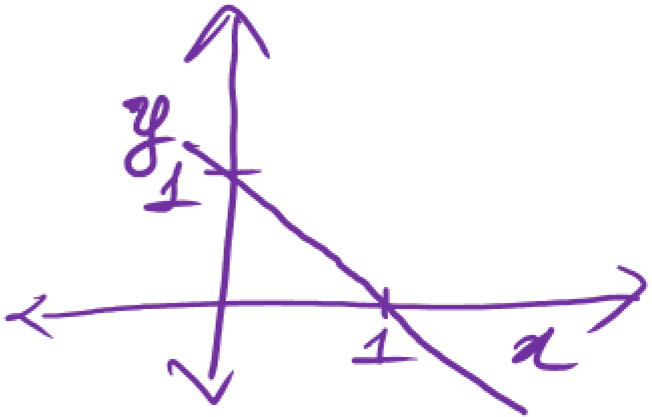
# SVM

- Maximum Margin Classifier
- Margin: is the shortest distance between threshold and data observation.

# Observations about Equation of Line

Consider x+y =1

And 5x +5y = 5




$5x+5y=5$

Divide by 5

$$\frac{5x+5y}{5}=\frac{5}{5}$$

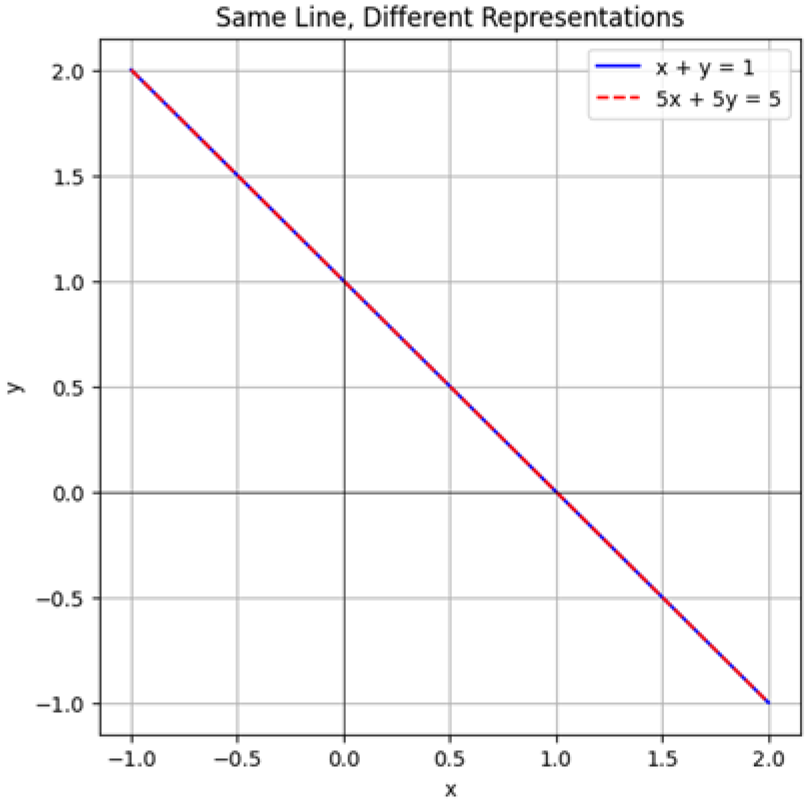
$$\frac{5(x+y)}{5}=1$$

$\Rightarrow \underline{x+y=1}$

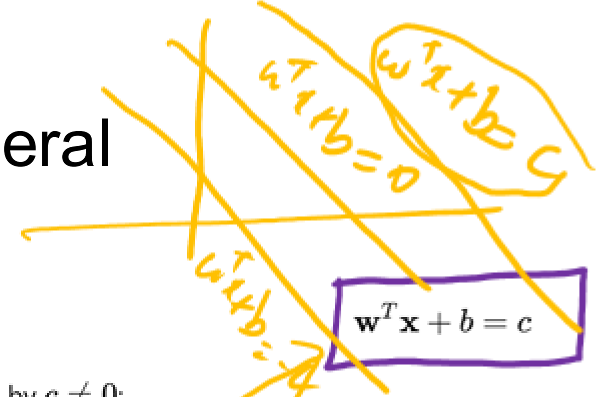
# Observations about Equation of Line

Consider x+y =1

And 5x +5y = 5



# In general




**1. When:**

$$\mathbf{w}^T\mathbf{x} + b = c$$

Divide both sides by $c \neq 0$:

$$\frac{\mathbf{w}^T\mathbf{x} + b}{c} = 1$$

So, **they represent the same line**, just written in a normalized or scaled form.

**2. Similarly, if:**

$$\mathbf{w}^T\mathbf{x} + b = -c$$

Divide by $c \neq 0$:

$$\frac{\mathbf{w}^T\mathbf{x} + b}{c} = -1$$

Again, it's the **same line**, just normalized.

$x_2 = \underbrace{m x_1 + b - c}$

$\boxed{m x_1 - x_2 + b = c}$

$w = \begin{bmatrix} m \\ -1 \end{bmatrix} \quad x \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$\boxed{w^T \vec{x} + b = c}$

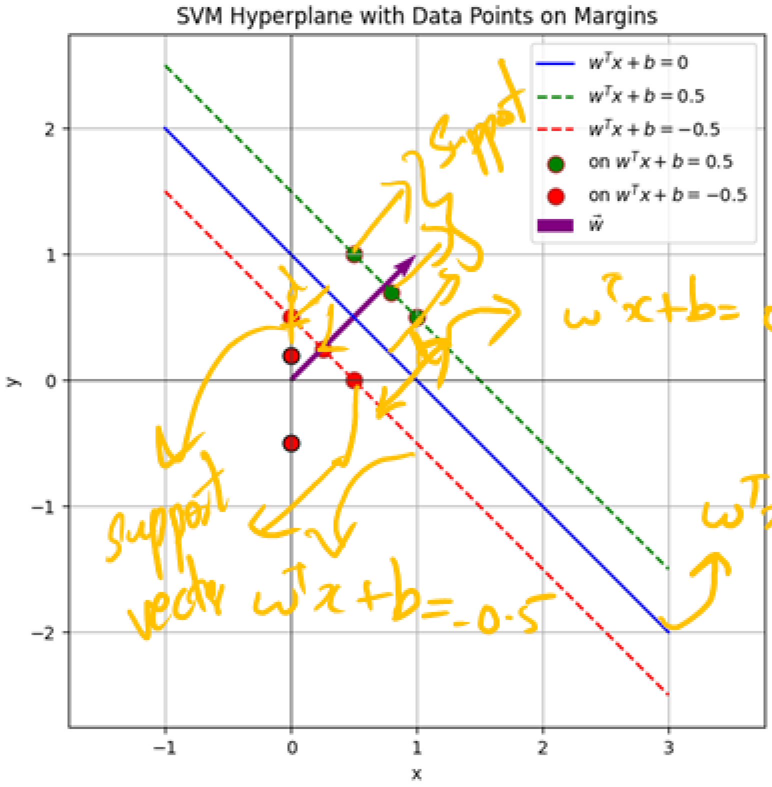
$y = 3x + 5$

$y = 3x + 2 + 3$

$\boxed{-3 = 3x - y + 2}$

# Observations

Bounding lines
Boundary Plane




$w^T x + b = 0.5$

$w^T x + b = 0$

**Decision Boundary:**

$$x + y = 1 \quad \Leftrightarrow \quad \mathbf{w}^T\mathbf{x} + b = 0 \quad \text{where } \mathbf{w} = [1, 1],\ b = -1$$

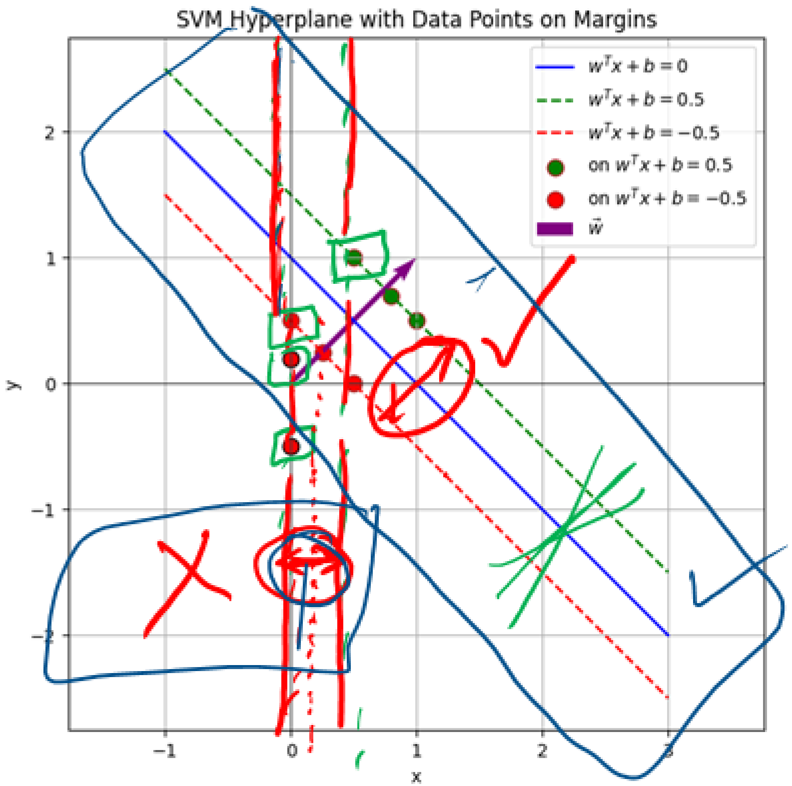
**Positive Margin:**

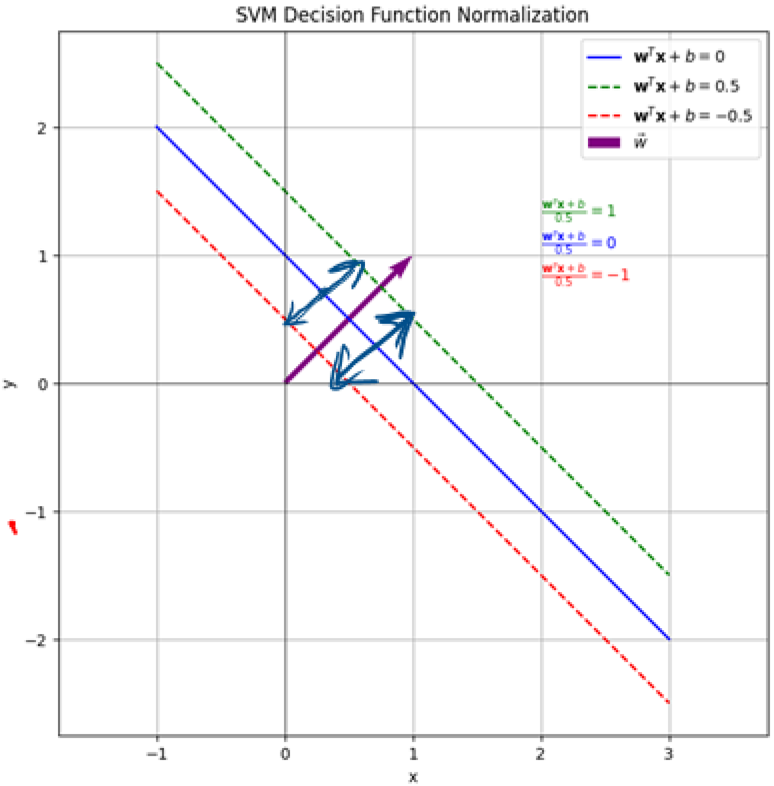
$$x + y = 1.5 \quad \Leftrightarrow \quad \mathbf{w}^T\mathbf{x} + b = 0.5 \quad \text{where } \mathbf{w} = [1, 1],\ b = -1$$

**Negative Margin:**

$$x + y = 0.5 \quad \Leftrightarrow \quad \mathbf{w}^T\mathbf{x} + b = -0.5 \quad \text{where } \mathbf{w} = [1, 1],\ b = -1$$

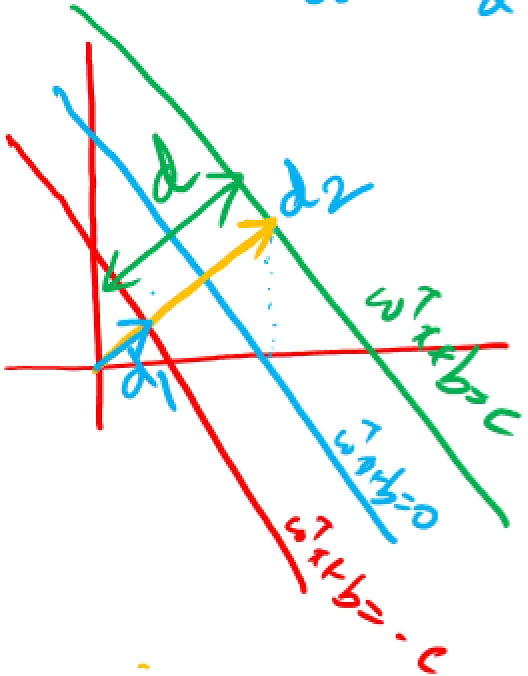
# Maximising the distance between the bounding plane

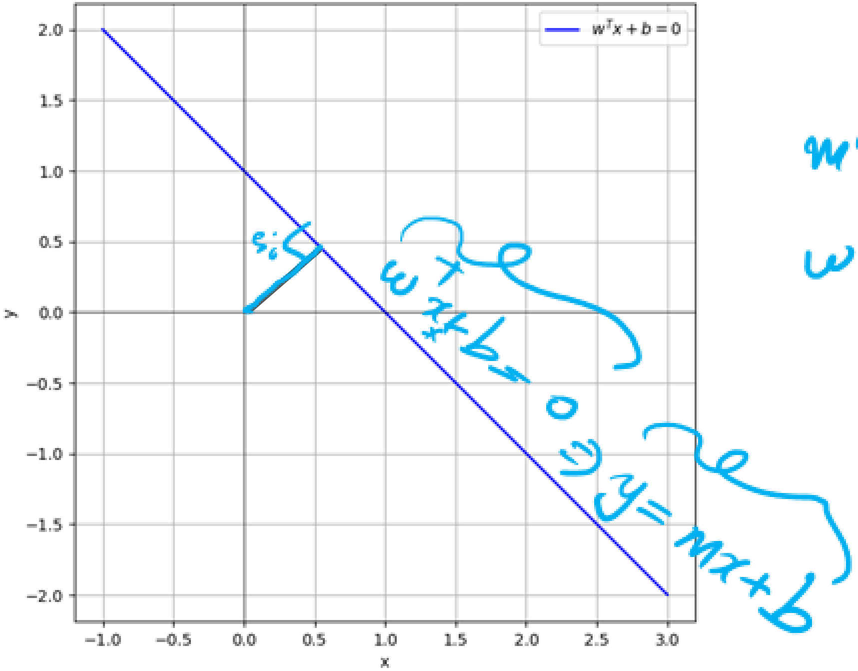


# Perpendicular Distance from origin to x+ y =1

$d = d_2 - d_1$




$mx - y + b = 0$

$w = \begin{bmatrix} m \\ -1 \end{bmatrix} \quad \vec{x_*} = \begin{bmatrix} x \\ y \end{bmatrix}$

$w^T x_* + b$

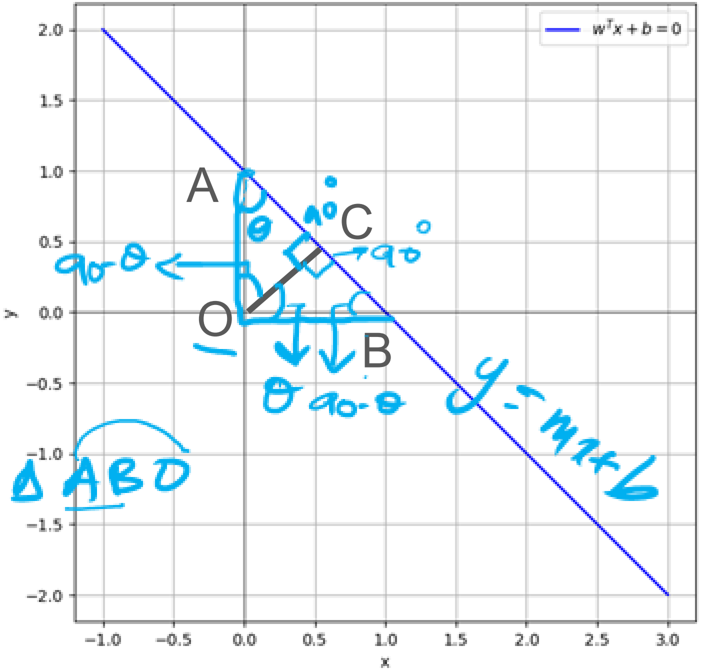
# Perpendicular Distance from origin to x+ y =1

Triangle OBC and

Triangle ABO

OB/OC = AB/OA

OC = OA. OB / AB




find OC?

$\Delta OBC$

$\angle COB = \theta$

$\angle CBO = 90 - \theta$

$\angle OCB = 90°$

$\Delta ABO$

$\angle OAB = \theta$

$\angle ABO = 90 - \theta$

$\angle AOB = 90°$

$\Delta OBC$ is similar to $\Delta ABO$

$$\frac{OB}{OC} = \frac{AB}{OA}$$

$\Delta OBC$ & $\Delta ABO$

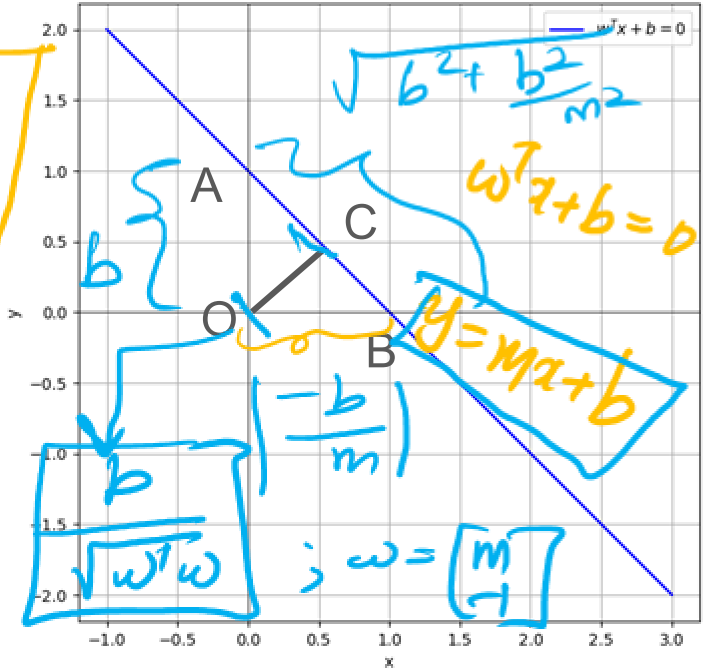
$$\frac{OB}{OC} = \frac{AB}{OA}$$

$$\boxed{OC = \frac{OB \times OA}{AB}}$$

$OB = ?$

$OA = ?$

$AB = ?$




$\sqrt{b^2 + \frac{b^2}{m^2}}$

$w^T x + b = 0$

$y = mx + b$

$b$

$\left|\frac{-b}{m}\right|$

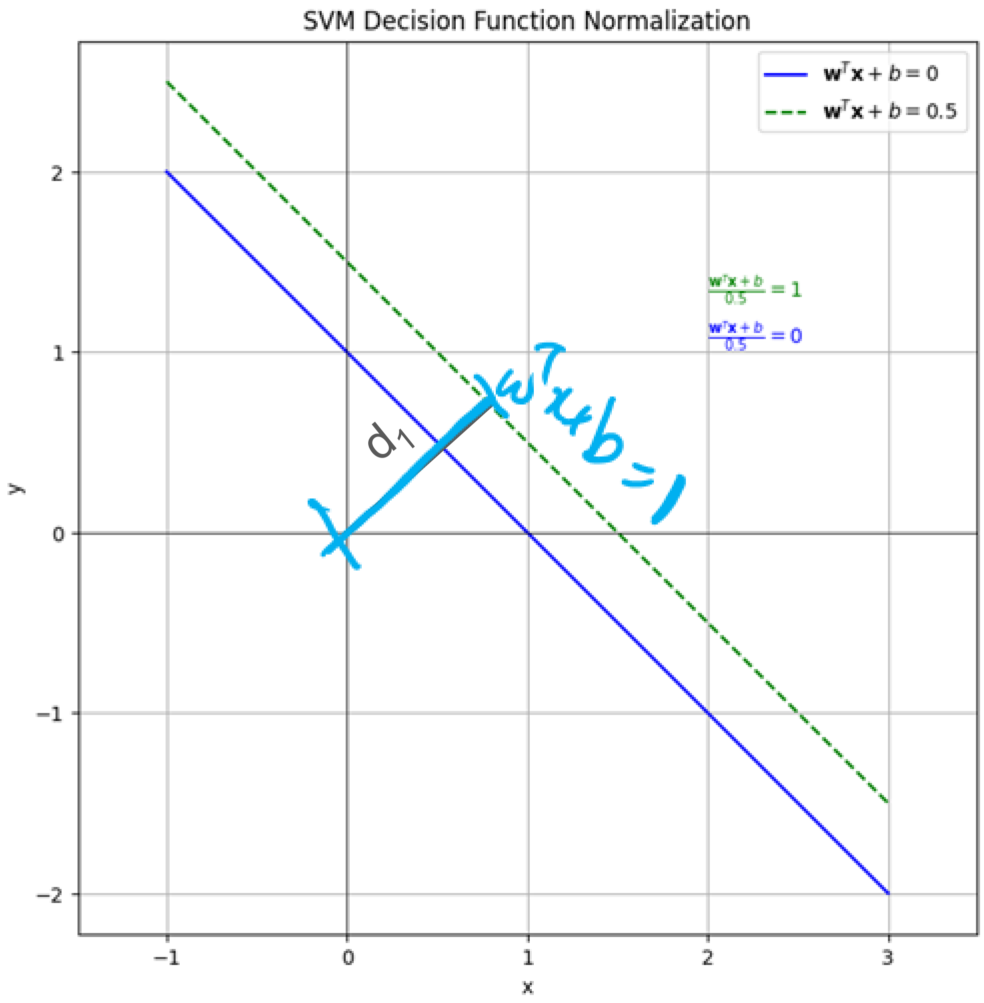
$$\boxed{\frac{b}{\sqrt{w^T w}}} \; ; \; w = \begin{bmatrix} m \\ -1 \end{bmatrix}$$

$$w = \begin{bmatrix} m \\ -1 \end{bmatrix}$$

$$w^T w = m^2 + 1$$

$$\frac{\frac{b}{m} \times b}{\sqrt{b^2 + \frac{b^2}{m^2}}}$$

$$\frac{b^2 \sqrt{m^2}}{m\, b \sqrt{m^2+1}} \Rightarrow \boxed{\frac{b}{\sqrt{m^2+1}}}$$



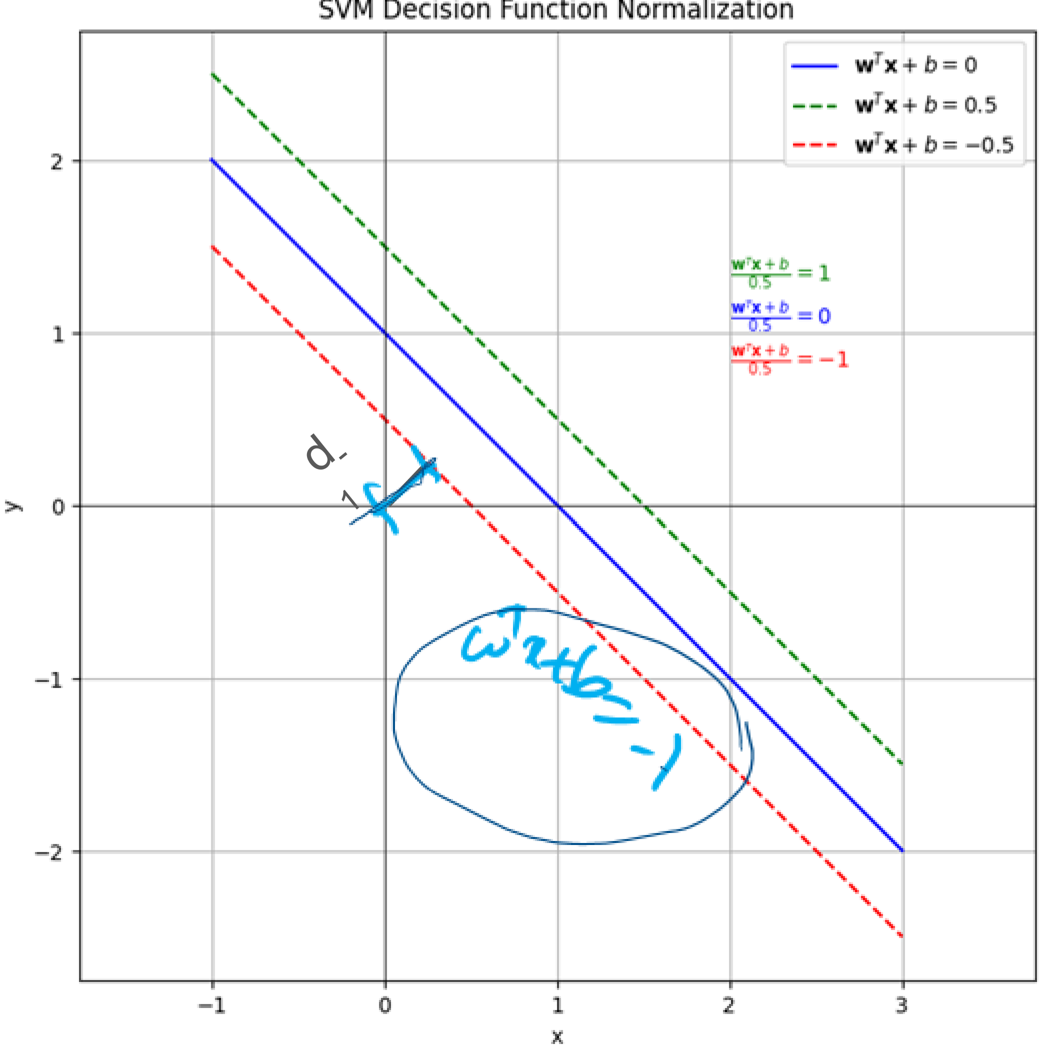

that for $w^Tx+b=0$
the ⊥ distance from origin is

$$\frac{b}{\sqrt{w^Tw}}$$

If $w^Tx+b=1$

$$w^Tx+b-1=0$$

$$\frac{|b-1|}{\sqrt{w^Tw}}$$



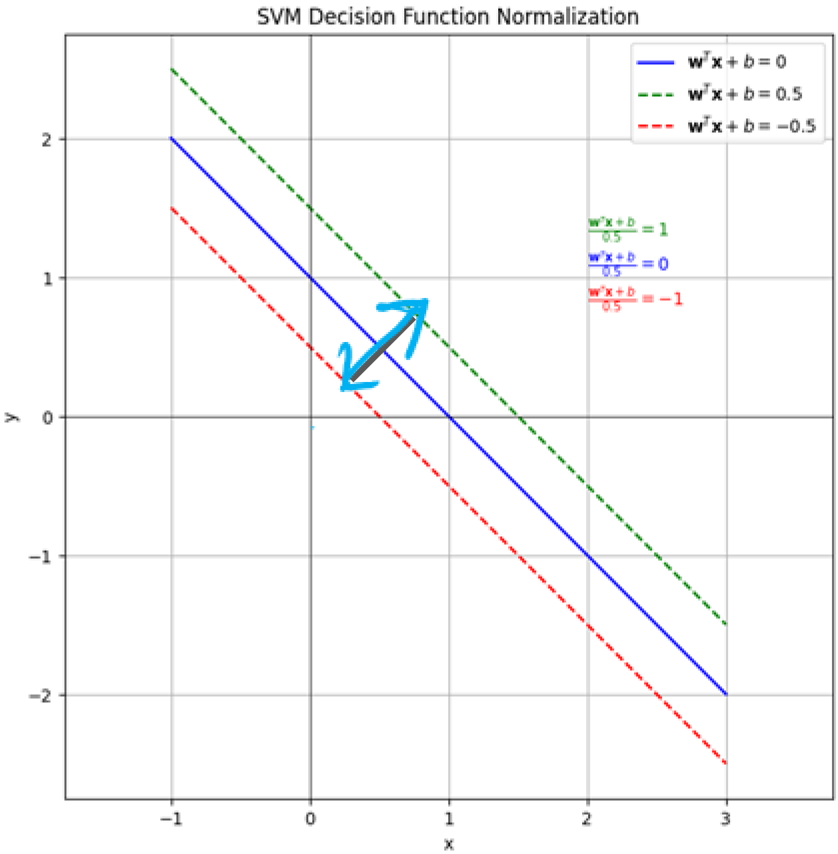

$$w^Tx+b=-1$$

$$\frac{|b+1|}{\sqrt{w^Tw}}$$

$$w^Tx+(b+1)=0$$

$$\frac{b+1}{\sqrt{w^Tw}}$$

# How to find the distance between the bounding planes?




$d = d_1 - d_{-1}$

$$\frac{|b-1|}{\sqrt{w^Tw}} - \frac{(|b+1|)}{\sqrt{w^Tw}}$$

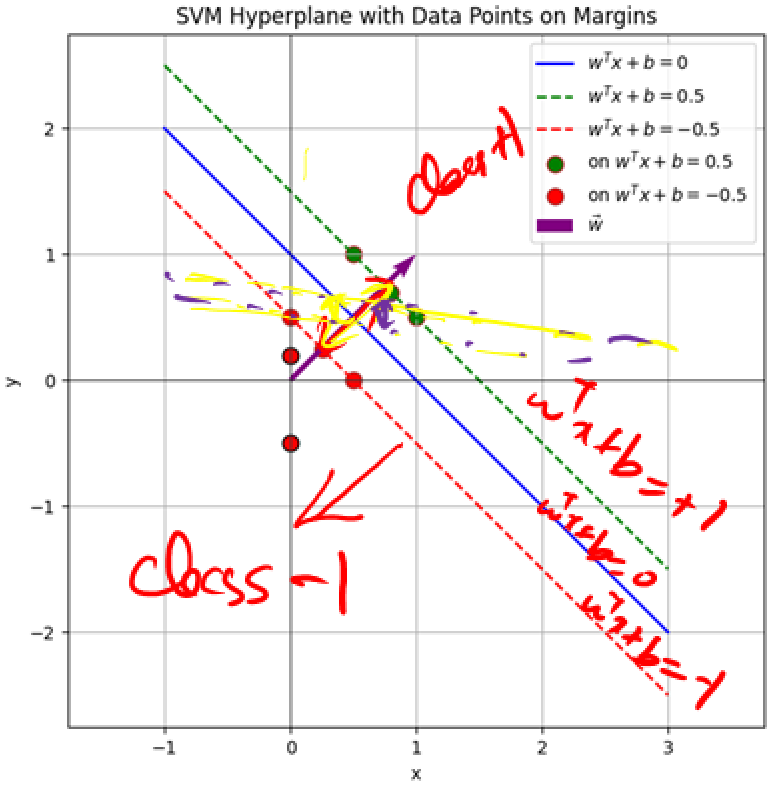
$$\Rightarrow \frac{|-2|}{\sqrt{w^Tw}} = \frac{2}{\sqrt{w^Tw}}$$

$x = \frac{2}{b}$

$\max \frac{2}{\sqrt{w^T w}}$ subject to

$\max \frac{1}{w^T w}$

$\max \frac{1}{\sqrt{w^T w}}$




**Geometric (Max-Margin) Formulation:**

$$\max_{\mathbf{w},b} \frac{2}{\sqrt{\mathbf{w}^T\mathbf{w}}}$$

subject to: $\mathbf{w}^T\mathbf{x}^i + b \geq 1 \quad \forall \mathbf{x}^i \in \mathcal{X}_+$ +ve class

$\mathbf{w}^T\mathbf{x}^i + b \leq -1 \quad \forall \mathbf{x}^i \in \mathcal{X}_-$

-ve class data informa

**Equivalent Convex Formulation (Minimization):**

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

subject to: $\mathbf{w}^T\mathbf{x}^i + b \geq 1 \quad \forall \mathbf{x}^i \in \mathcal{X}_+$

$\mathbf{w}^T\mathbf{x}^i + b \leq -1 \quad \forall \mathbf{x}^i \in \mathcal{X}_-$

$-5 < -1$

$5 > 1$

**Equivalent Convex Formulation (Minimization):**

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{subject to:} \quad \boxed{\begin{aligned} &\mathbf{w}^T\mathbf{x}^i + b \geq 1 \quad \forall \mathbf{x}^i \in \mathcal{X}_+ \\ &\mathbf{w}^T\mathbf{x}^i + b \leq -1 \quad \forall \mathbf{x}^i \in \mathcal{X}_- \end{aligned}}$$

Case I

what if $y^{(i)} = +1$ $\quad w^T x^i + b \geq 1$

Case II $\quad \boxed{\begin{aligned} &w^T x^{(i)} + b \leq -1 \\ &y^{(i)} = -1 \end{aligned}}$

We encode both constraints:

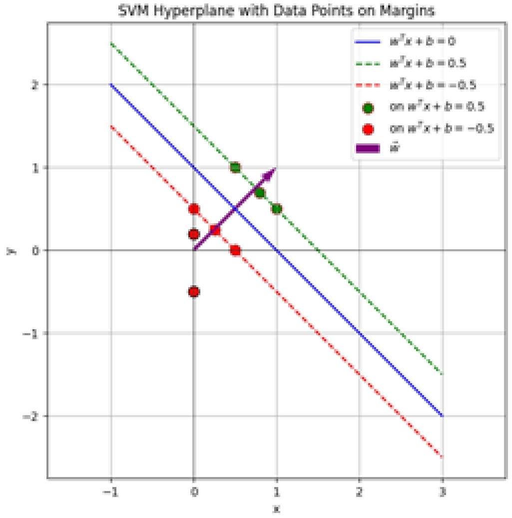
- $\mathbf{w}^T\mathbf{x}^i + b \geq 1$ for $y^i = +1$
- $\mathbf{w}^T\mathbf{x}^i + b \leq -1$ for $y^i = -1$

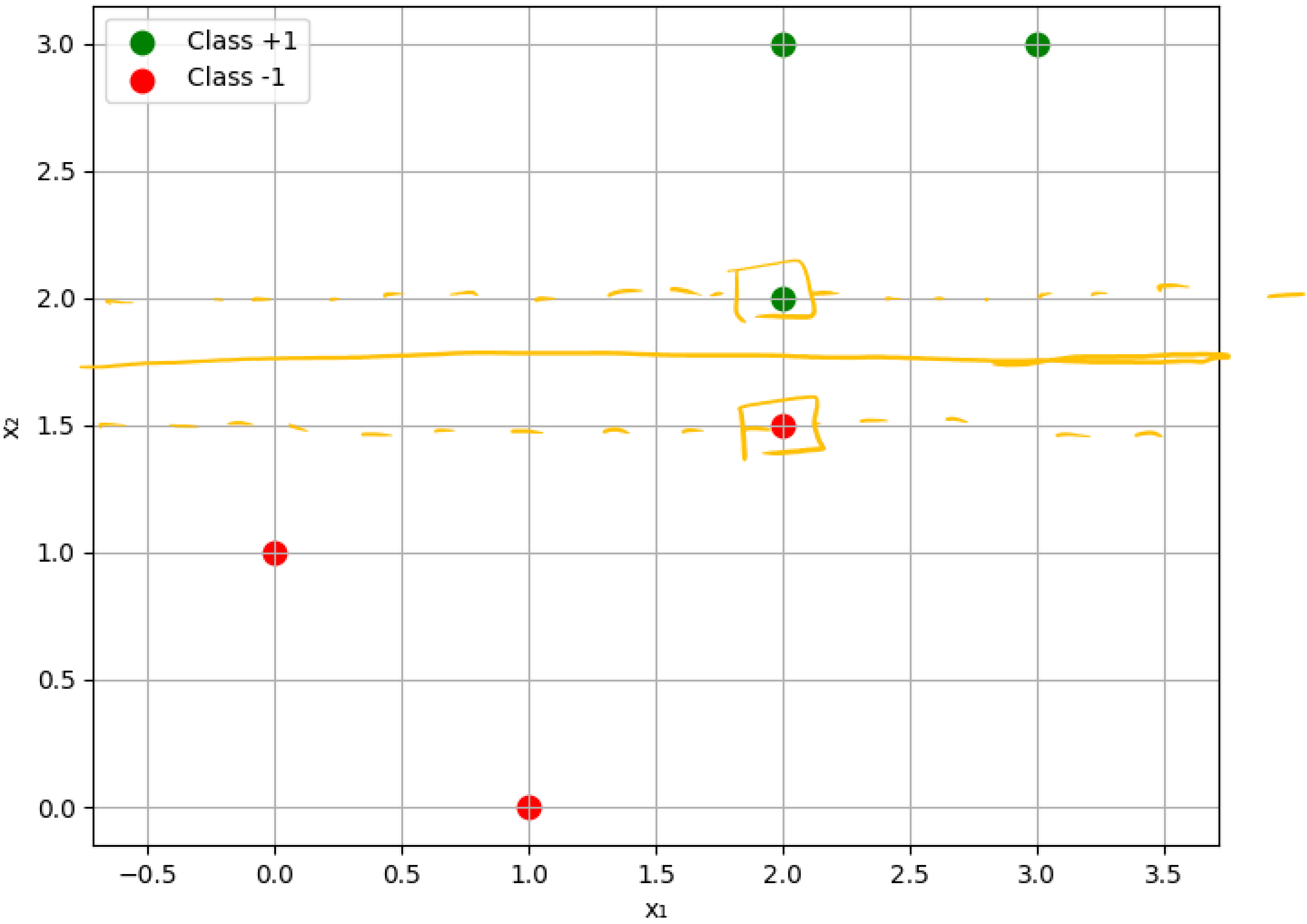
as:

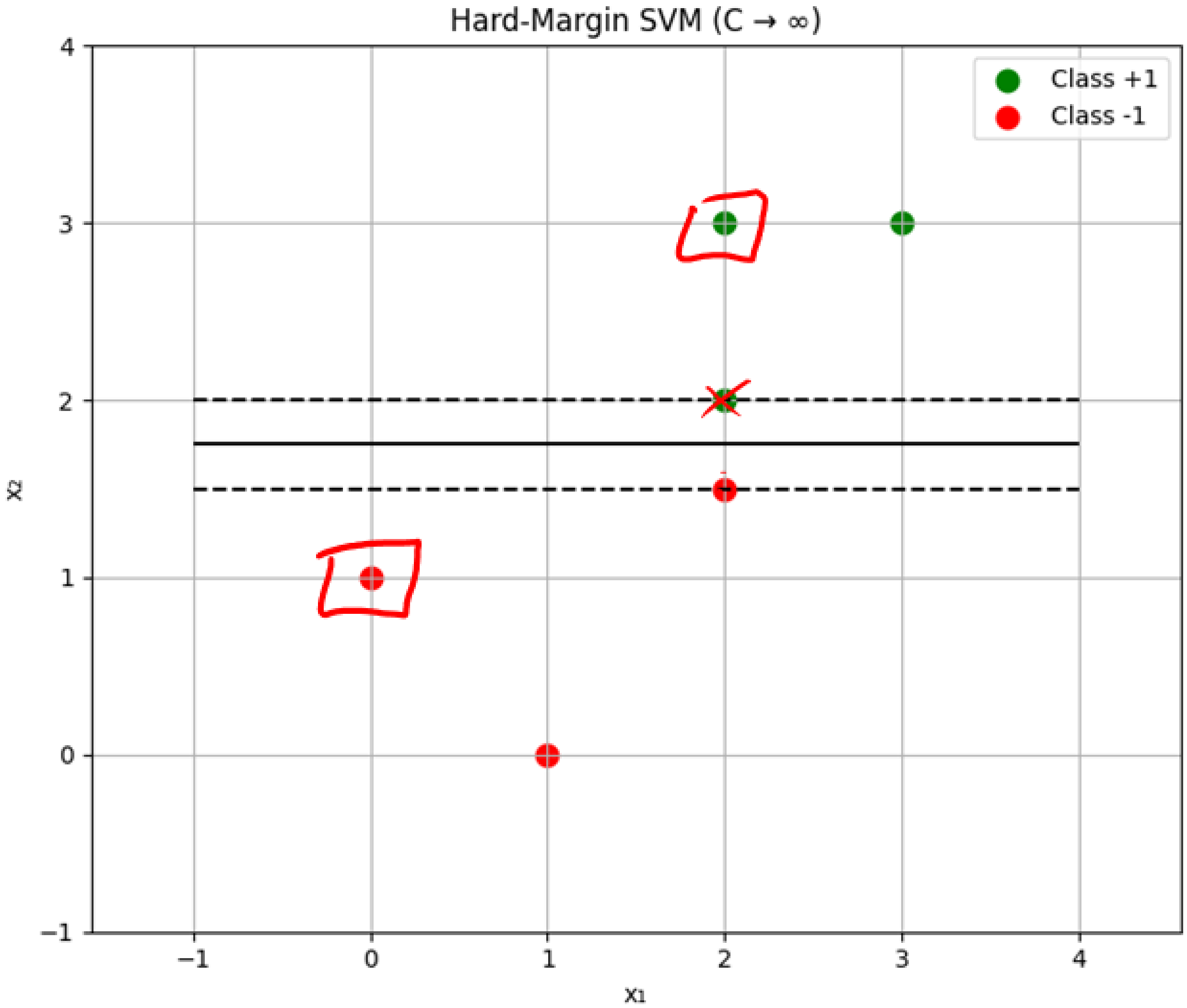
Case I $\quad y_i(w^T x^{(i)} + b) \geq y_i * 1$ $\quad \boxed{y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1}$

$$\min_{\mathbf{w},b} \quad \frac{1}{2}\mathbf{w}^T\mathbf{w}$$
$$\text{subject to:} \quad y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1 \quad \forall i$$

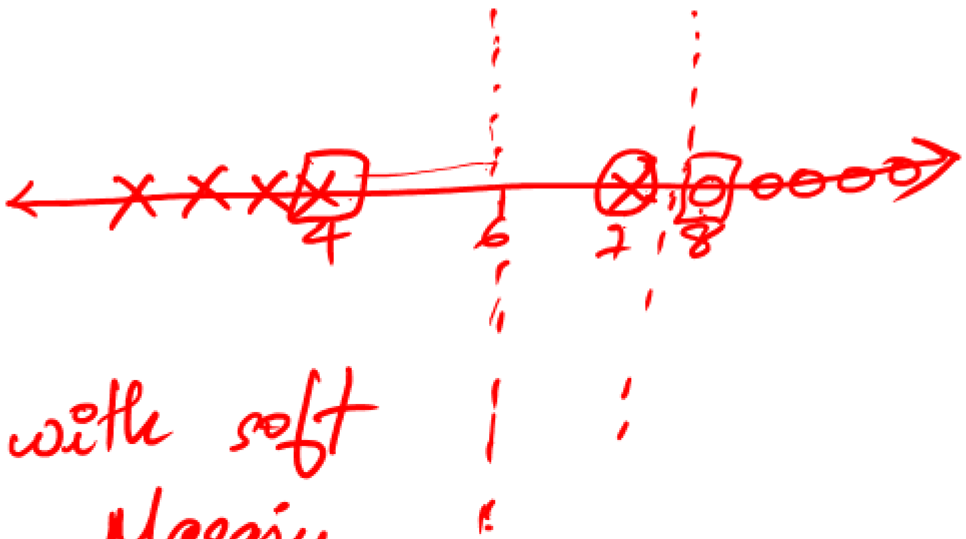
Case II $\quad y^{(i)}(w^T x^{(i)} + b) \geq \underbrace{y_i (-1)}_{1}$



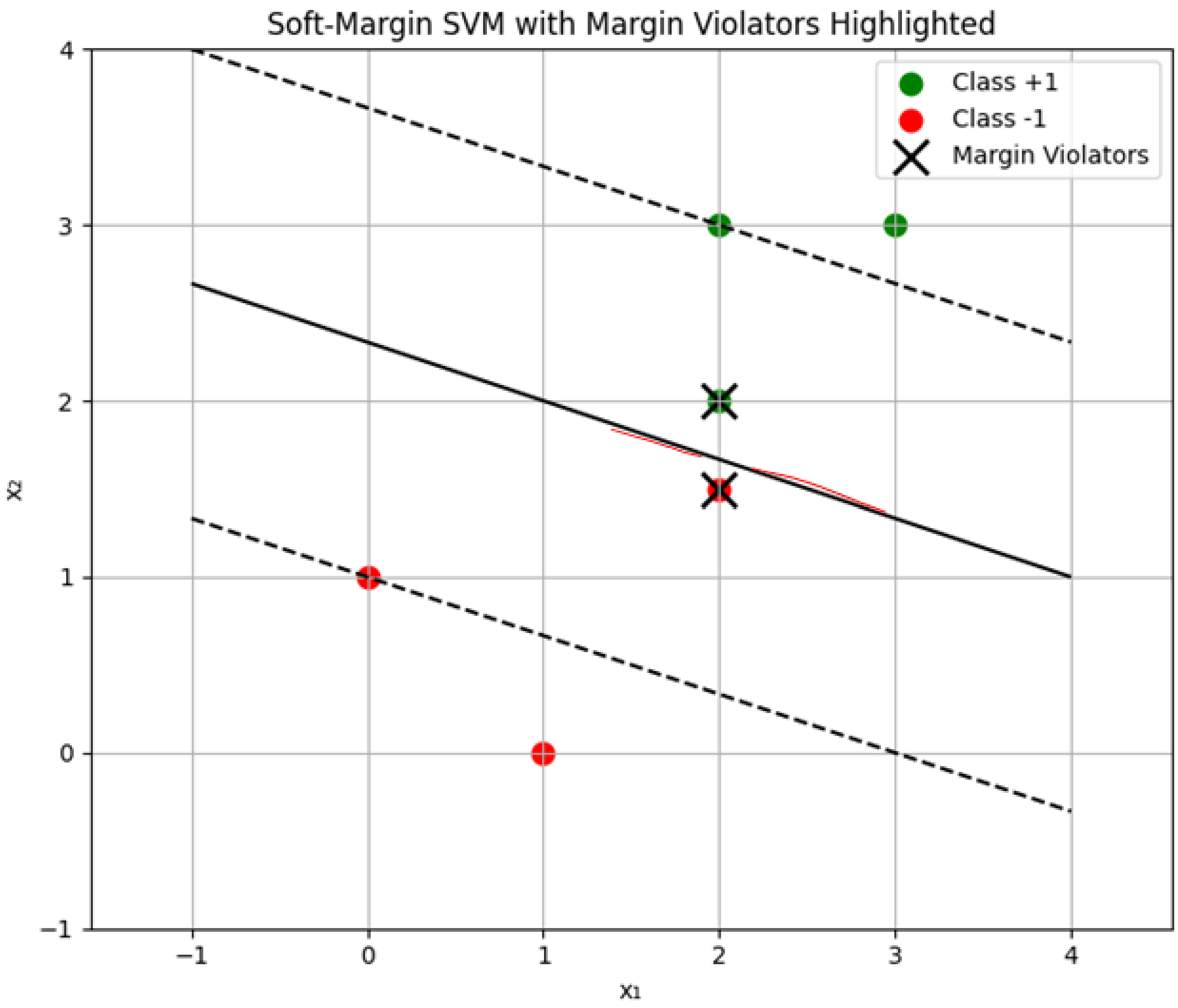

Class +1
Class -1
3.0
2.5
2.0
1.5
1.0
0.5
0.0
x2
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
x1



Hard-Margin SVM (C → ∞)
Class +1
Class -1
x1
x2
4
3
2
1
0
−1
−1
0
1
2
3
4




SVM with soft Margin



Soft-Margin SVM with Margin Violators Highlighted
Class +1
Class -1
Margin Violators
x₁
x₂
−1
0
1
2
3
4

# SVM with soft margin

$$y^{(i)}(w^T x_i + b) \geq 1 - \xi_i$$

### ✅ Soft-Margin SVM Formulation (Non-linearly separable)

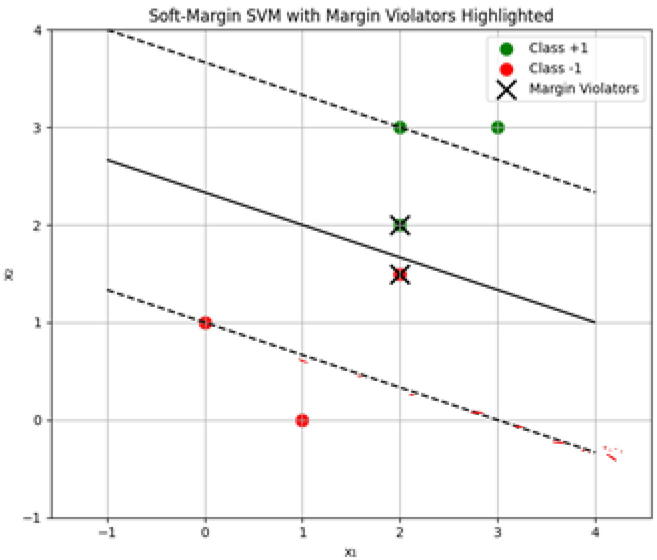
We now allow slack $\xi^i \geq 0$ to **soften** the constraints:

$$y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1 - \xi^i$$

and penalize the slack in the objective.

### ✅ Final Soft-Margin SVM (Primal Form):

$$\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi^i \\ \text{subject to:} \quad & y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1 - \xi^i, \quad \forall i \\ & \xi^i \geq 0, \quad \forall i \end{aligned}$$




### 🧠 Interpretation

- $\xi^i$ measures how much a point violates the margin.
- $C > 0$ controls the **trade-off** between maximizing margin and allowing violations:
  - **Large** $C$ → penalizes violations heavily → acts like hard margin
  - **Small** $C$ → more tolerant of misclassification

# SVM with soft margin

### ✅ Soft-Margin SVM Formulation (Non-linearly separable)

We now allow slack $\xi^i \geq 0$ to **soften** the constraints:

$$y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1 - \xi^i$$

and penalize the slack in the objective.

---

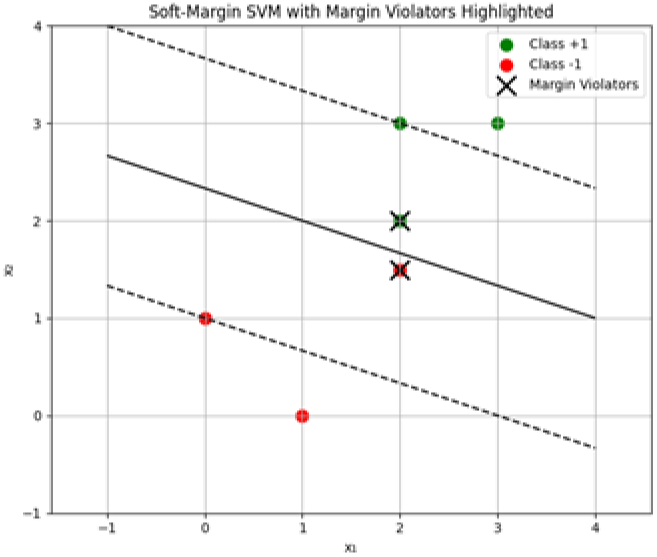
### ✅ Final Soft-Margin SVM (Primal Form):

$$\boxed{\begin{aligned} \min_{\mathbf{w},b,\xi} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi^i \\ \text{subject to:} \quad & y^i(\mathbf{w}^T\mathbf{x}^i + b) \geq 1 - \xi^i, \quad \forall i \\ & \xi^i \geq 0, \quad \forall i \end{aligned}}$$

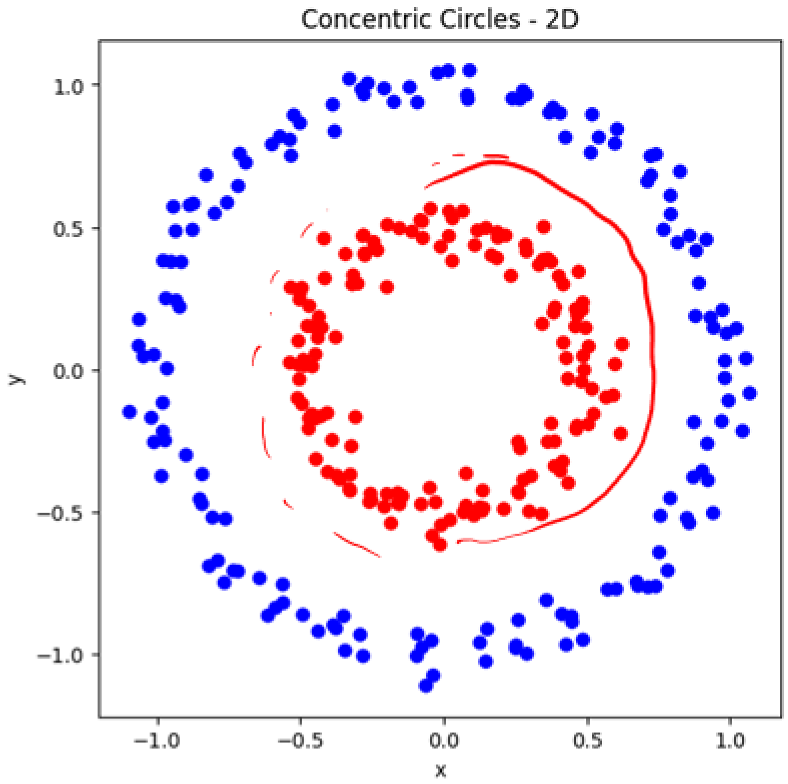
---

### 🧠 Interpretation

- $\xi^i$ measures how much a point violates the margin.
- $C > 0$ controls the **trade-off** between maximizing margin and allowing violations:
  - **Large** $C$ → penalizes violations heavily → acts like hard margin
  - **Small** $C$ → more tolerant of misclassification

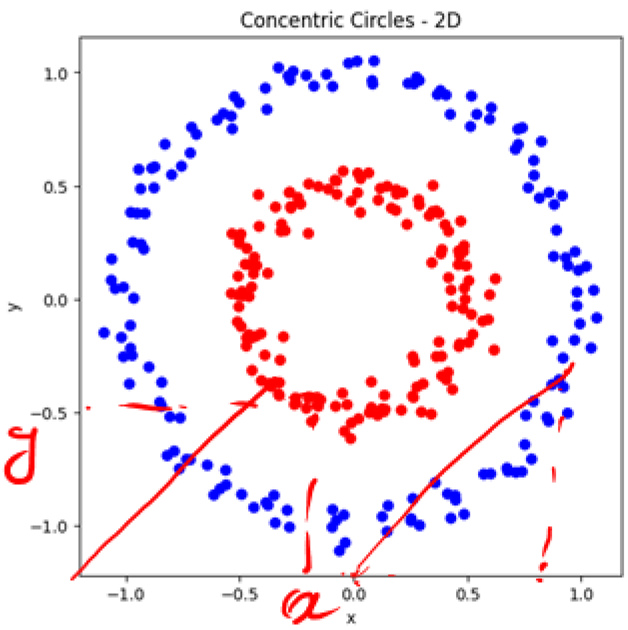


# What if the data is not linearly separable?

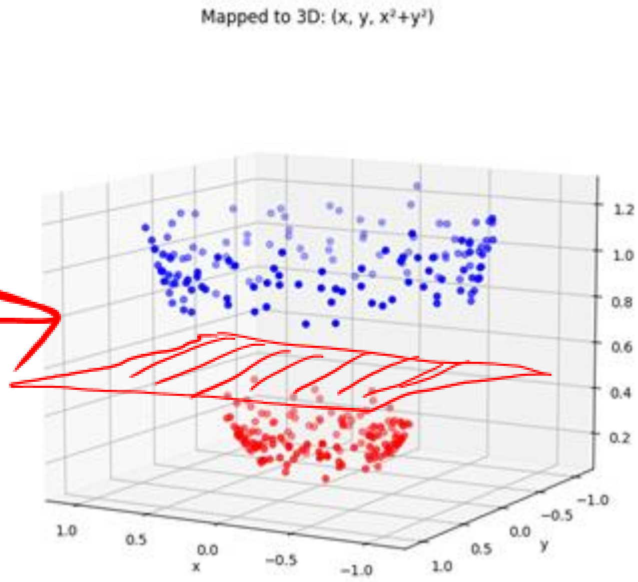


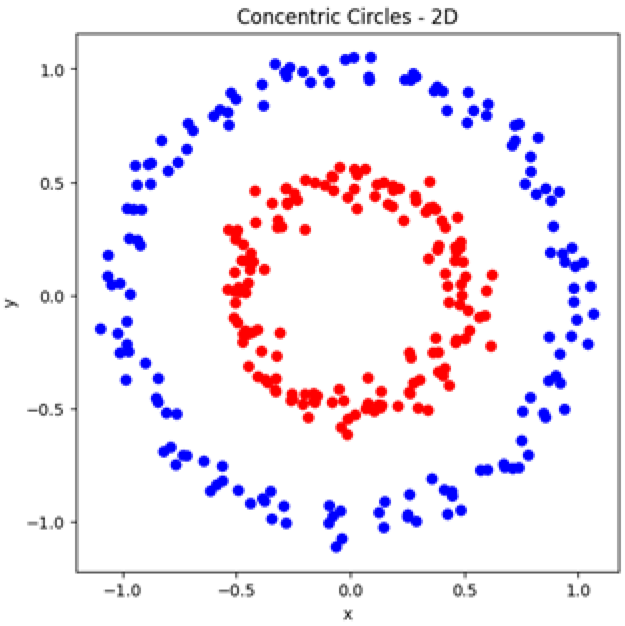
# Kernel Trick

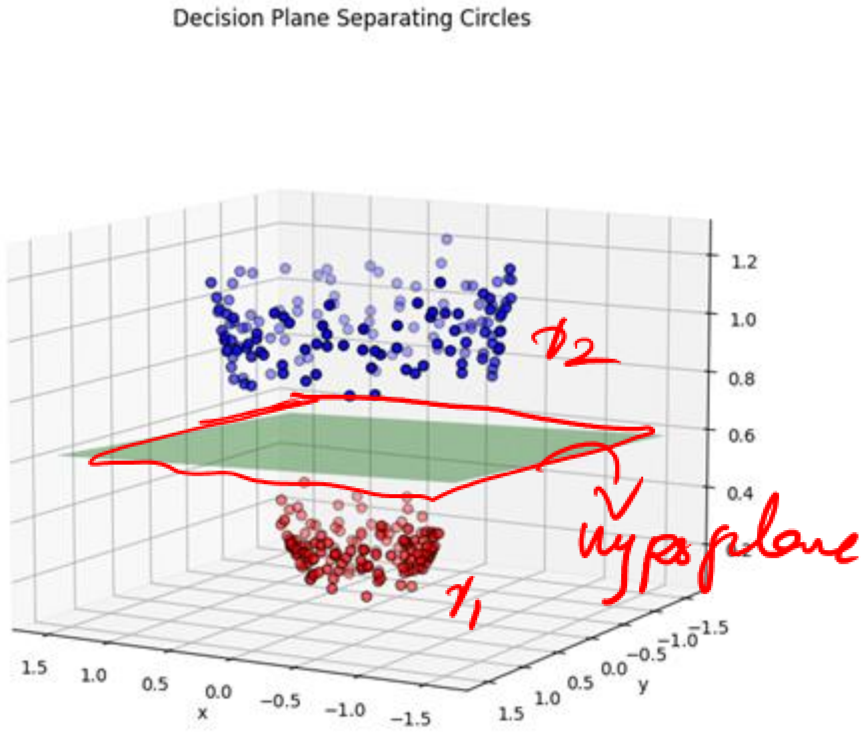
$(x,y) \rightarrow (x, y, x^2+y^2)$




$\varphi(x, y) \rightarrow (x, y, x^2 + y^2)$



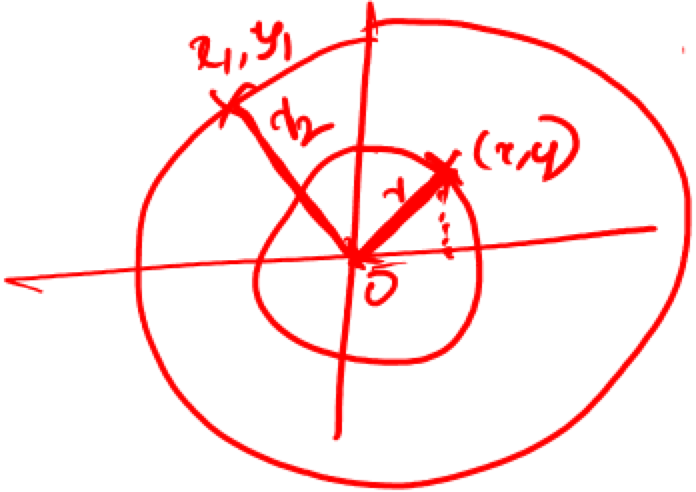

Concentric Circles - 2D
1.0
0.5
0.0
−0.5
−1.0
y
−1.0
−0.5
0.0
0.5
1.0
x


Decision Plane Separating Circles
1.2
1.0
0.8
0.6
0.4
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
x
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
y
$y_2$
$y_1$
hyperplane

# Kernels

Hyperparameter $\boxed{\text{Kernel, } C}$

| Kernel Name | Equation | Notes |
|---|---|---|
| Linear Kernel | $K(x,z) = x^\top z$ | No mapping — simple dot product. |
| Polynomial Kernel | $K(x,z) = (x^\top z + c)^d$ | Adds non-linearity with degree $d$. $c$ is a constant (often 1). |
| Radial Basis Function (RBF) Kernel | $K(x,z) = \exp\left(-\gamma \lVert x - z \rVert^2\right)$ | Infinite-dimensional mapping! Most popular. |
| Sigmoid Kernel | $K(x,z) = \tanh(\kappa x^\top z + c)$ | Inspired by neural networks (like MLPs). |
| Laplacian Kernel | $K(x,z) = \exp\left(-\gamma \lVert x - z \rVert\right)$ | Similar to RBF but uses $\lVert x - z \rVert$ instead of $\lVert x - z \rVert^2$. |
| Chi-Squared Kernel | $K(x,z) = \exp\left(-\gamma \sum \frac{(x_i - z_i)^2}{x_i + z_i}\right)$ | Used in histogram-based data (e.g., images). |
| ANOVA Kernel | $K(x,z) = \sum \exp\left(-\gamma (x_i - z_i)^2\right)^d$ | Used in regression and bioinformatics. |




On squaring both sides

$r = \sqrt{x^2+y^2} \Rightarrow x^2+y^2$

$r_2 = \sqrt{x_1^2+y_1^2} \quad x_1^2+y_1^2$

$r^2 < r_2^2$

$r < r_2$

$(x, y, r^2), (x, y, r_2^2)$