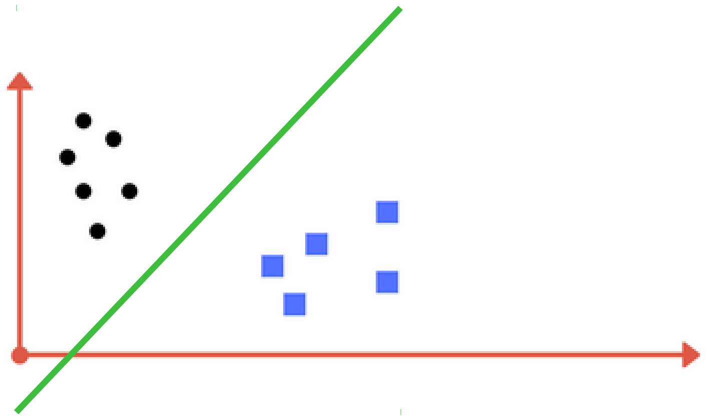
SVM

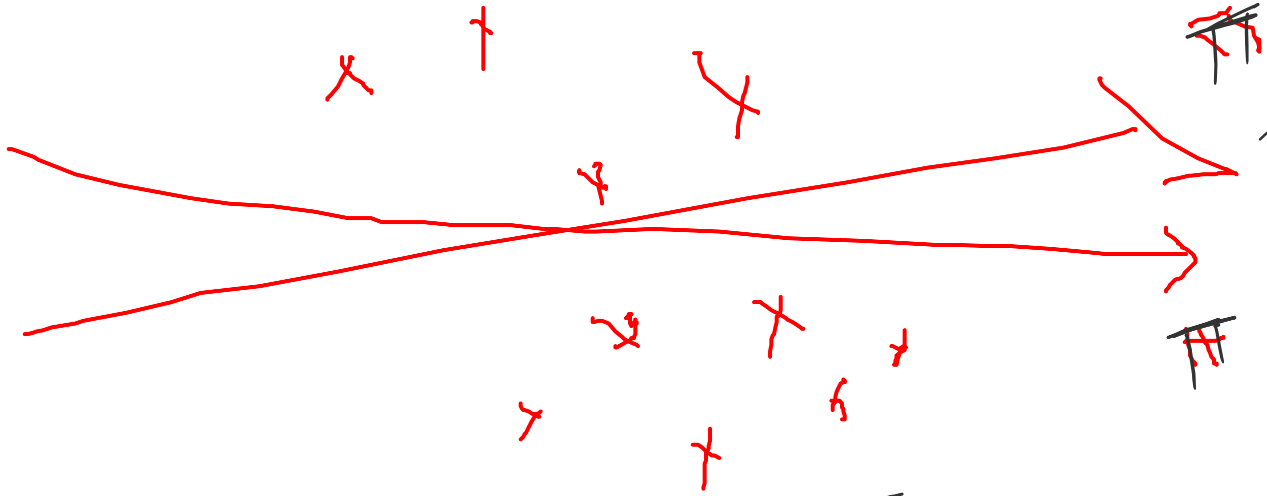
What is SVM?

- Support Vector Machine or SVM is Supervised Learning algorithms, which is used for Classification as well as Regression problems.
- However, primarily, it is used for Classification problems in Machine Learning.
- The goal of the SVM algorithm is to create the decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.
- In two dimensional space this hyperplane is a line dividing a plane in two parts where in each class lay in either side.



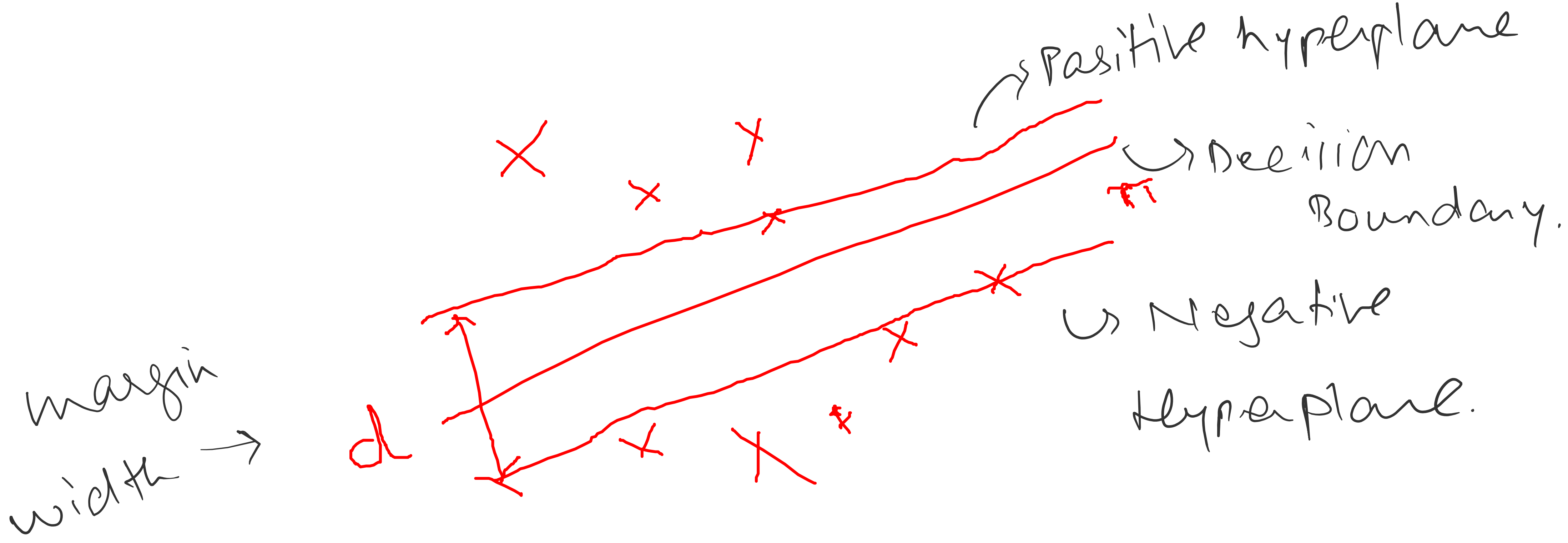
→ The decision Boundary which has maximum margin is choosen

(positive examples)



→ decision Boundary which of the following decision Boundary gives the best result?

(Negative Examples)



Positive hyperplane
Decision Boundary.
Negative Hyperplane.
margin width
d

Types of SVM

- SVM can be of two types:
- **Linear SVM:** Linear SVM is used for linearly separable data, which means if a dataset can be classified into two classes by using a hyperplane, then such data is termed as linearly separable data, and classifier used, is called as Linear SVM classifier.
- **Non-linear SVM:** Non-Linear SVM is used for non-linearly separated data, which means if a dataset cannot be classified by using a hyperplane, then such data is termed as non-linear data and classifier used, is called as Non-linear SVM classifier.

Hyperplane

- The decision boundary which separates data points in two classes is known as hyperplane

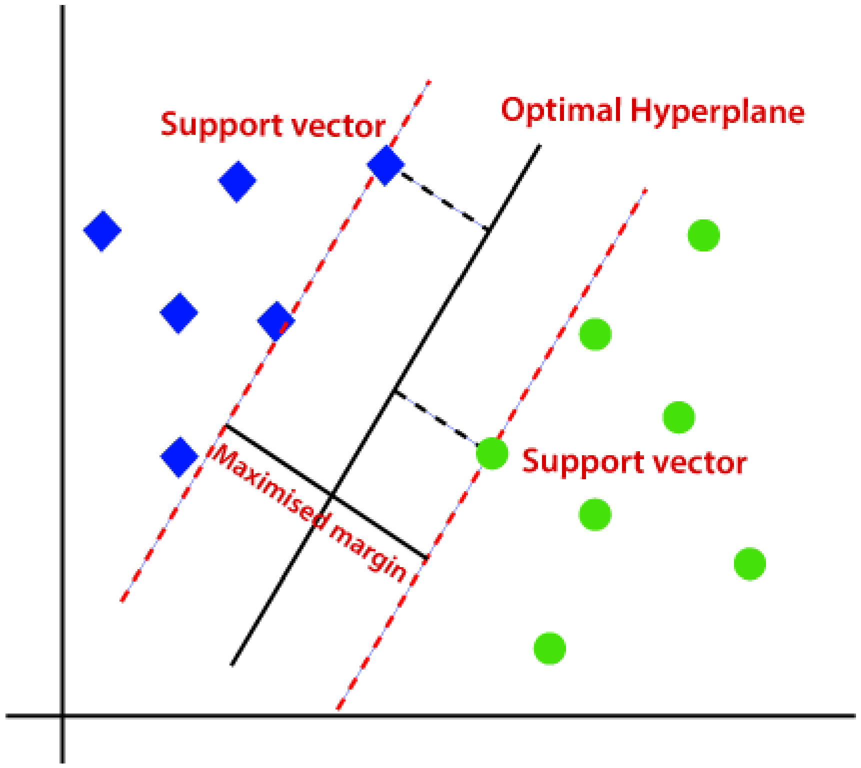
  There can be multiple decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the optimal hyperplane of SVM.

  The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features, then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane.

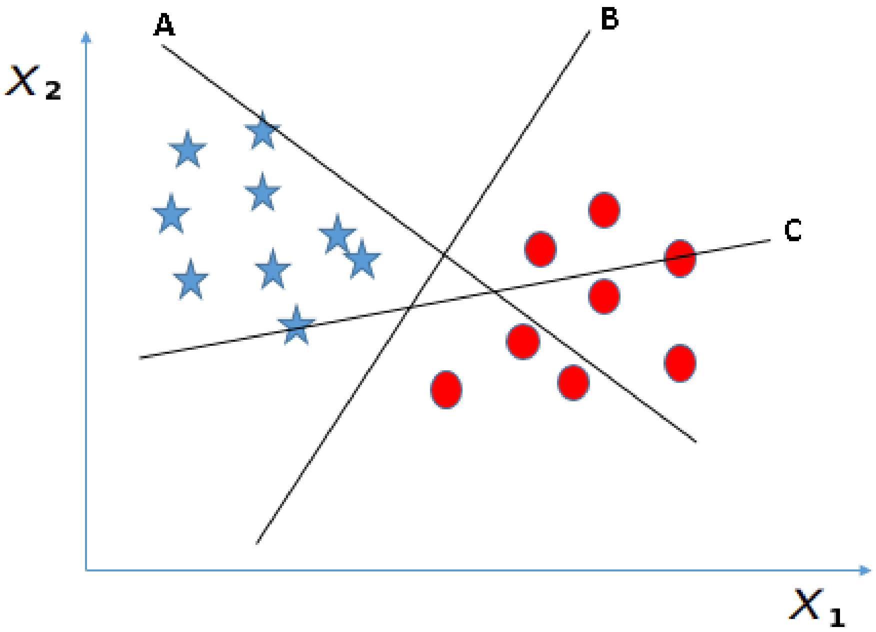
  Always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Support Vectors

- The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector.
- Since these vectors support the hyperplane, hence called a Support vector.

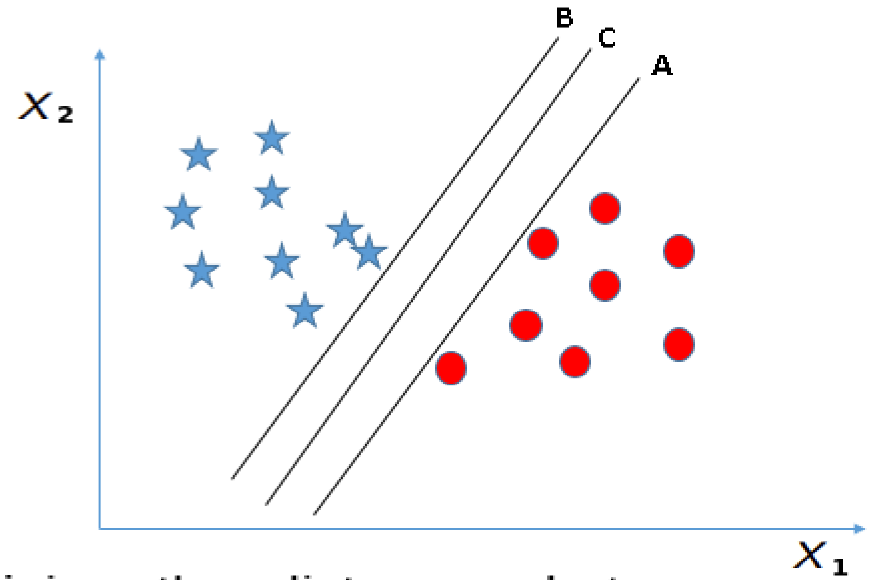


Identify the right hyper-plane - (1)



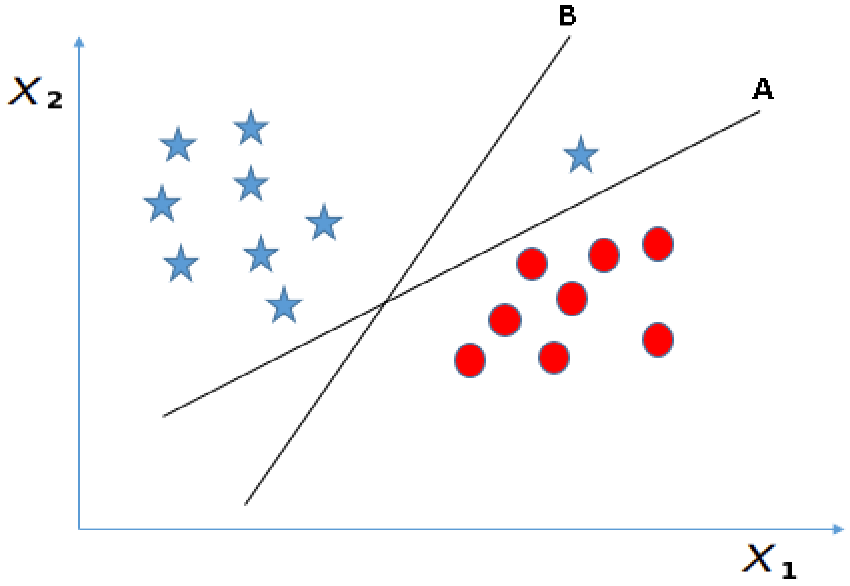
- Select the hyper-plane which segregates the two classes better.
- In this scenario, hyper-plane “B” has excellently performed this job.

Identify the right hyper-plane - (2)



- Here, maximizing the distances between nearest data point (either class) and hyper-plane will help us to decide the right hyper-plane. This distance is called as **Margin**.
- Above, you can see that the margin for hyper-plane C is high as compared to both A and B. Hence, C is the right hyper-plane.
- Another lightning reason for selecting the hyper-plane with higher margin is robustness.
- If we select a hyper-plane having low margin then there is high chance of miss-classification.

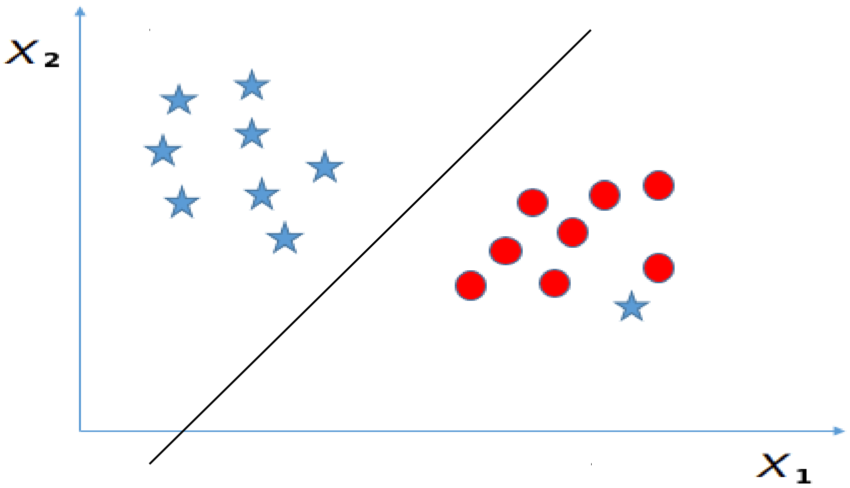
Identify the right hyper-plane - (3)



Soft margin case

- Though the hyper-plane B has higher margin compared to A, SVM selects the hyper-plane which classifies the classes accurately prior to maximizing margin.
- Here, hyper-plane B has a classification error and A has classified all correctly.
- Therefore, the right hyper-plane is A.

Identify the right hyper-plane - (4)

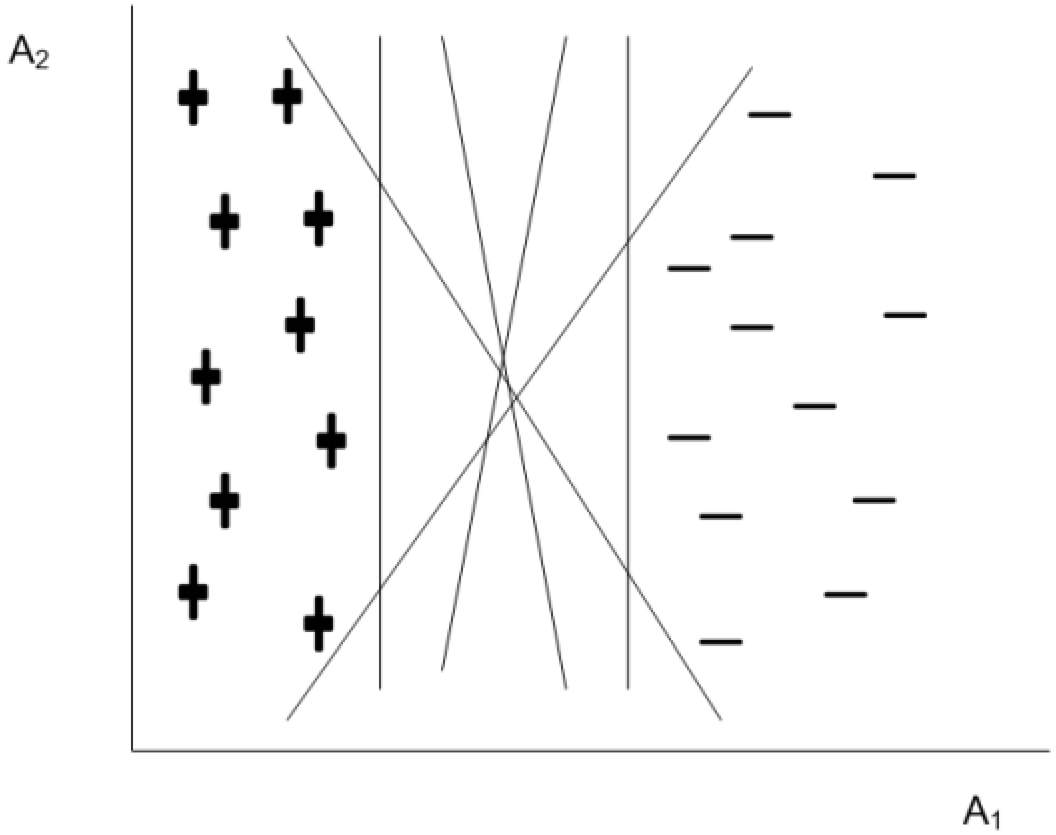


Soft margin

- It is not possible to segregate the two classes using a straight line, as one of the stars lies in the territory of other(circle) class as an outlier.
- The SVM algorithm has a feature to ignore outliers and find the hyper-plane that has the maximum margin.
- Hence, we can say, SVM classification is robust to outliers.

Maximum Margin Hyperplane

Figure 2: A 2D data linearly seperable by hyperplanes

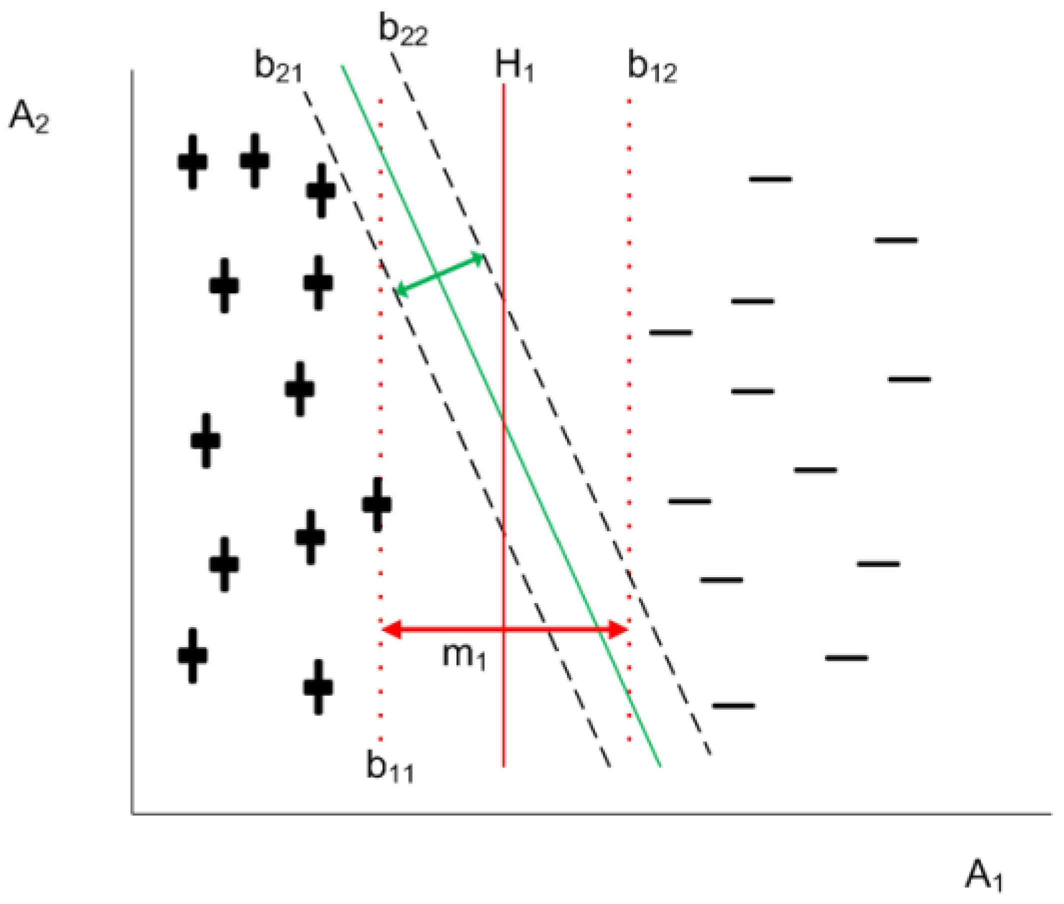


Maximum Margin Hyperplane contd...

- Figure 2 shows a plot of data in 2-D. Another simplistic assumption here is that the data is linearly separable, that is, we can find a hyperplane (in this case, it is a straight line) such that all +'s reside on one side whereas all -'s reside on other side of the hyperplane.
- From Fig. 2, it can be seen that there are an infinite number of separating lines that can be drawn. Therefore, the following two questions arise:
  1. Whether all hyperplanes are equivalent so far the classification of data is concerned?
  2. If not, which hyperplane is the best?

Maximum Margin Hyperplane

Figure 3: Hyperplanes with decision boundaries and their margins.



Algebraic form of margin equation for SVM,

→ Find the values of w, b such that margin is maximized

$w^T x + b = 1$

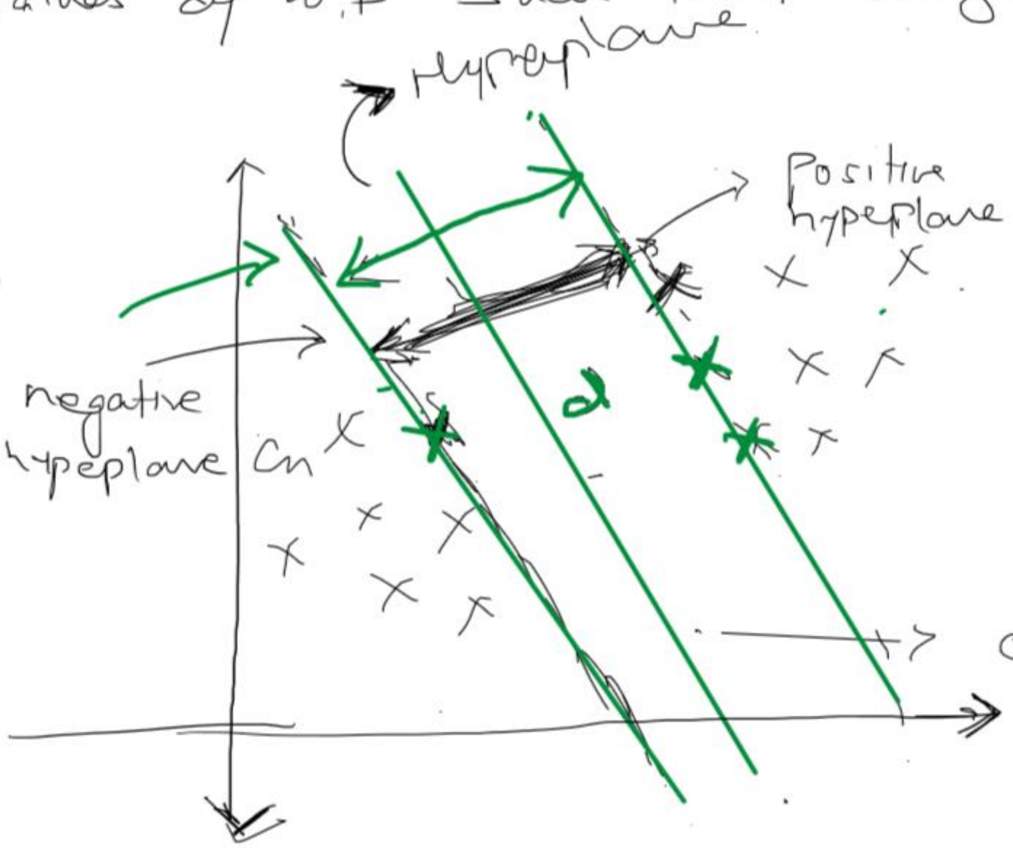
$w^T x + b = -1$

→ set of equation on hyperplane

→ $w^T x + b = 0$

$w^T x + b = 1$

$w^T x + b = -1$



Assumptions

↓

equation of hyperplanes

equation of hyperplane

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

Hyperplane → $w_1 x_1 + w_2 x_2 + w_n x_n + w_0 = 0$

n → dimensional

$\Rightarrow$ Decision Rules ✓ $\rightarrow$ $w^T x + b > 0 \rightarrow 1$

$w^T x + b < 0 \rightarrow -1$

$1, -1 \rightarrow$ classes

$\Rightarrow$ $\boxed{w^T x + b} = 1$

$w^T x + b = -1$

$\Rightarrow$ $-1, +1$ y_i $\rightarrow$ class

$\Rightarrow$ $y_i (w^T x + b) \geq 1 \rightarrow$ correct classification

y_i target variable

Assume

equation of hyperplane

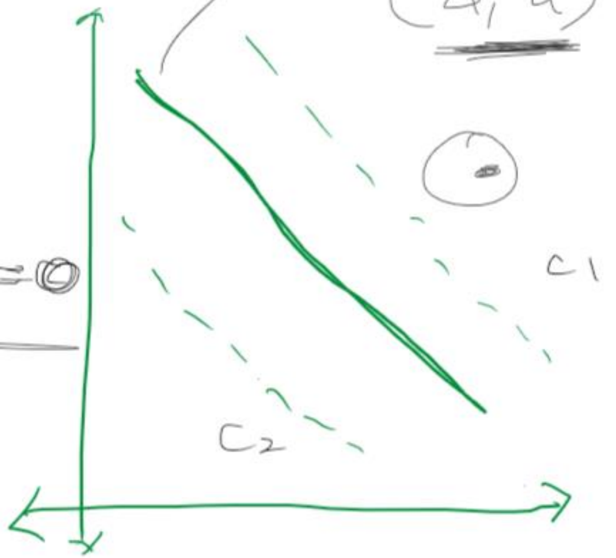
$\rightarrow$ $2x + 3y + 3 = 0$

which is of the form

$\rightarrow$ $w_1 x_1 + w_2 x_2 + w_0 = 0$

Hyperplane

$(4, 4)$ $\rightarrow$ New test point



Decision Rule,

$w^T x + b \geq 0 \rightarrow c_1$

$w^T x + b < 0 \rightarrow c_2$

So, For $(4, 4)$

$2(4) + 3(4) + 3 > 0 \rightarrow$ so class c1

For, $(-3, 1)$

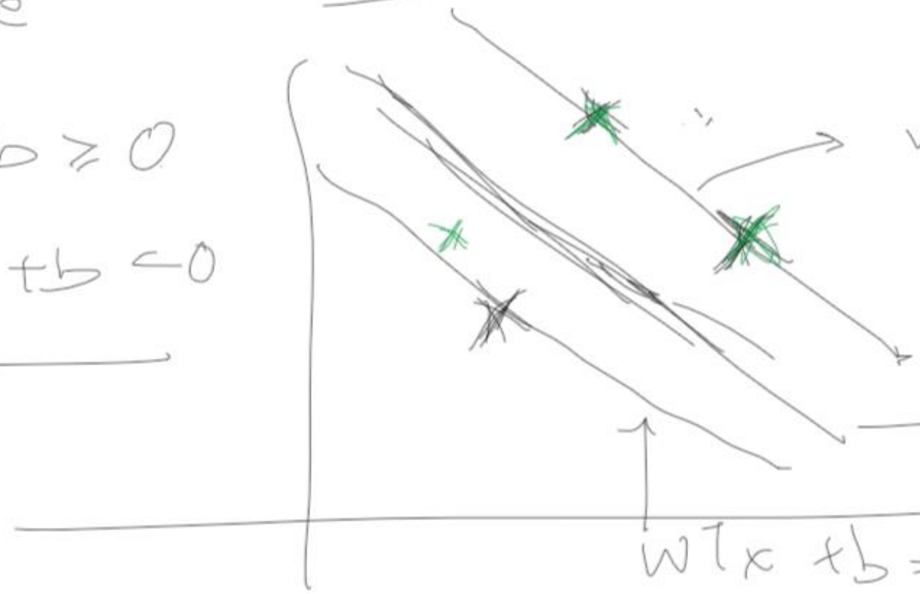
$2(-3) + 3(1) + 3 \leq 0 \rightarrow C_2$

Decision Rule ⇒ Defined → (No misclassification error)

Hard Margin

$c_1 = 1 \to w^Tx + b \geq 0$

$c_2 = -1 \to w^T + b < 0$



$w^Tx + b = 1$

equation $\boxed{w^Tx + b = 0}$

$w^Tx + b = -1$

⇒ $\boxed{y_i(w^Tx + b) \geq 1}$ → correctly classified

For all the correctly classified points

$$y_i(w^T x_i + b) \geq 1$$

For support vectors $y_i(w^T x + b) = 1$

since, we have equation of

positive & negative hyperplane $w^T x + b = +1$

$$w^T x + b = -1$$

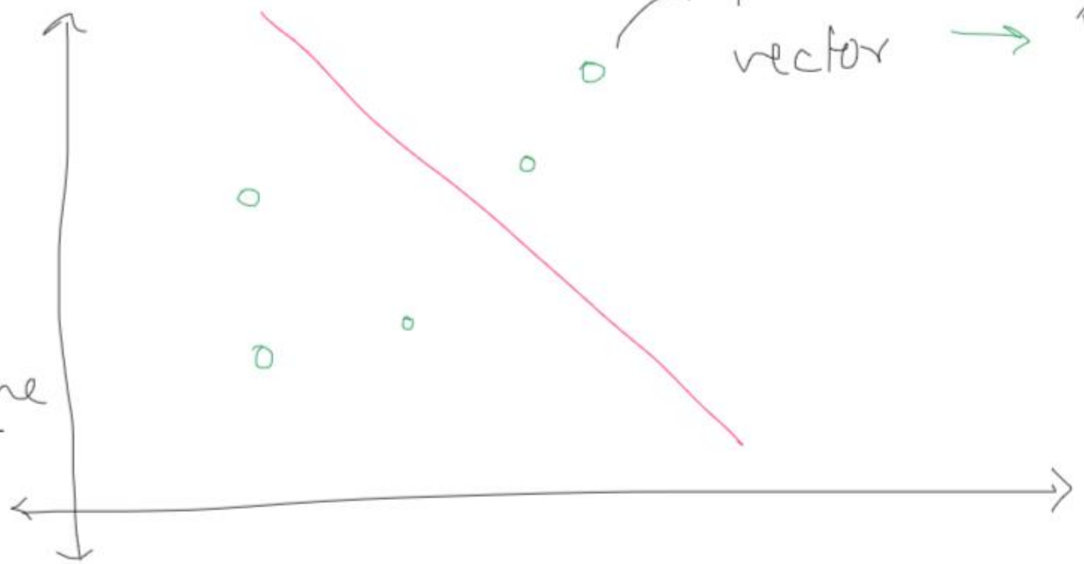
Projection of vector $\vec{a}$ on vector $\vec{b}$

$$\rightarrow \frac{\vec{a}\,\vec{b}}{\|\vec{b}\|}$$

→ Here along with family of equations, consider $\vec{w}$ orthogonal to all three hyperplane.

$\boxed{w^T x + b}$

distance of x from the seperating plane



Feature vector → For different value of w & b

→ where the equation of the hyperplane is

$$w^T x + b = 0$$

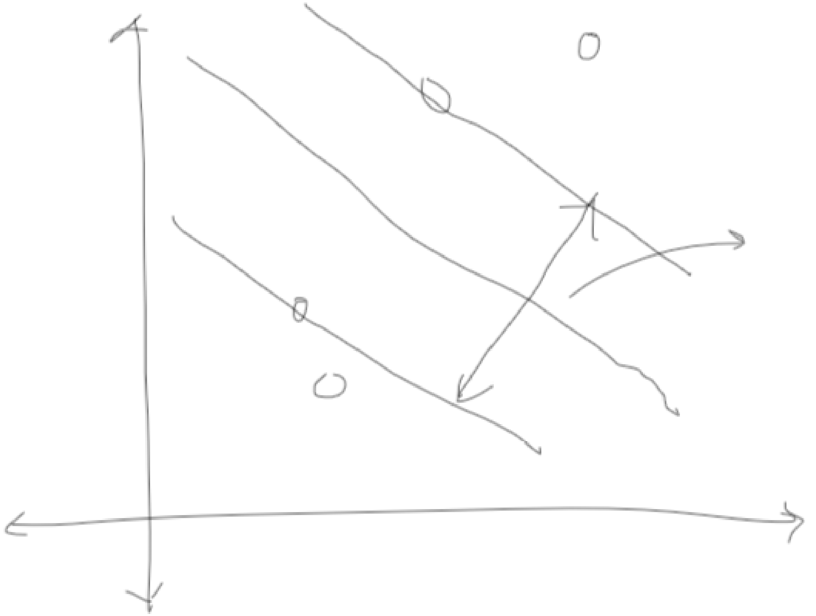
$\vec{w}$ & $\vec{b}$ are also vector in higher dimension.

→ For different value of w & b → different hyperplane can be obtained.

→ Each of these hyperplanes will have different margin or different levels classification:

'decision Boundary'

→



This hyperplane gives me a margin like this.

→ confidence level given by this classifica.

we have to find a seperating plane which has maximum margin.

Our aim is to choose seperating plane, $w^T x + b = 0$ which satisfies the condition,

$$\boxed{y_i (w^T x + b) \geq 1}$$ → After normalization.

→ This particular seperating plane which maximizes the margin.

→ Take vector x^+ on the w_1 side &

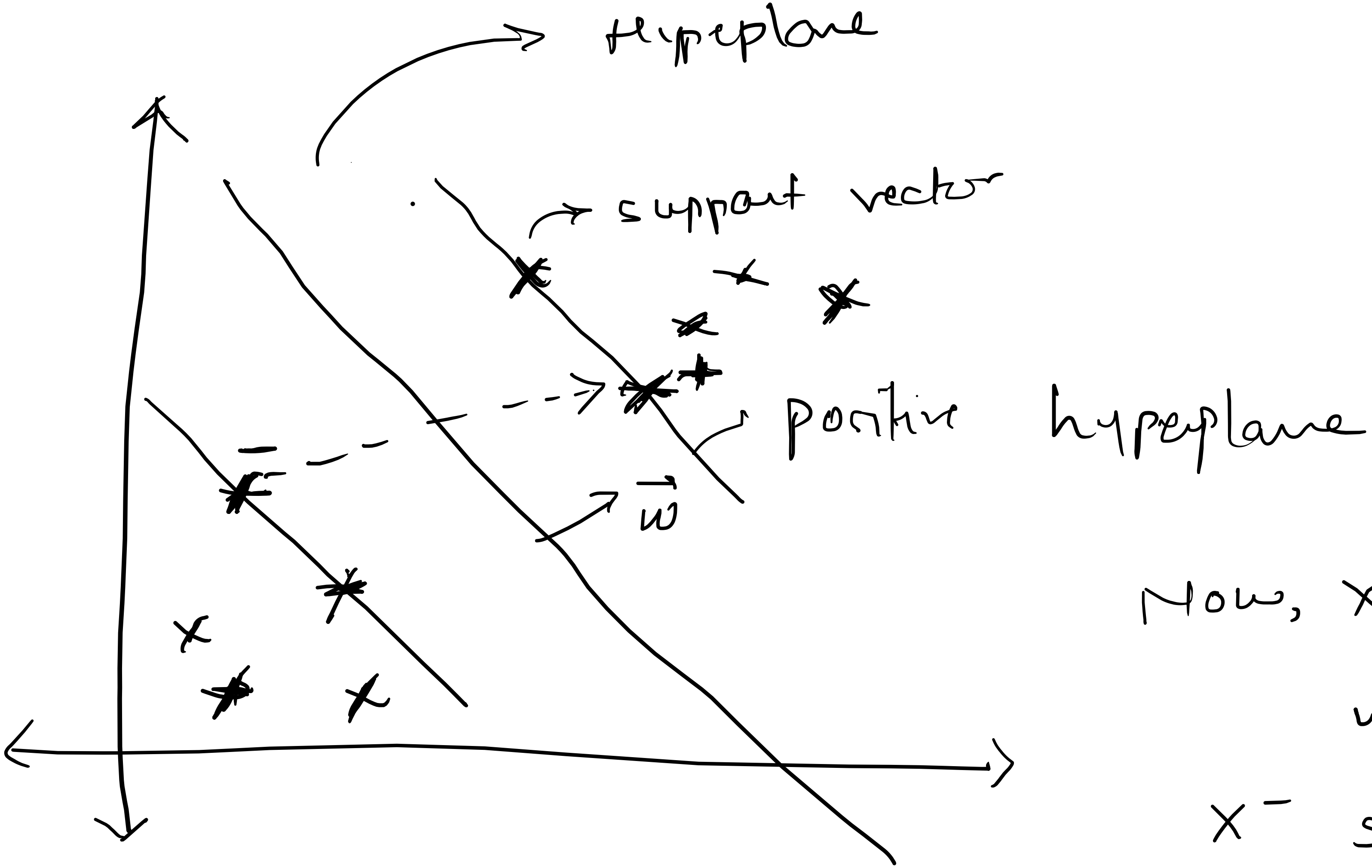
x^- on the other side w_2 → region

So, a vector $x^+ - x^-$ → is a vector drawn from x^- to x^+.

→ Once I have this vector, I can obtain this margin by,

$\rightarrow$ distance $= (x^+ - x^-) \cdot \dfrac{w}{\|w\|}$ Dot product with unit vector in the same direction

Projection of $(x^+ - x^-)$ $\dfrac{w}{\|w\|}$ orthogonal to seperating plane.



Now, x^+ satisfies

$$w^T x + b = +1$$

x^- satisfies

$$w^T x + b = -1$$

So, $w^T x^+ + b = 1,$

$$w^T x^- + b = -1.$$

So, $d = (x^+ - x^-) \dfrac{\vec{w}}{||w||}$

Now, $d = \dfrac{(x^+ w - x^- w)}{||w||}$

$d = \dfrac{|1 - b + b + 1|}{||w||}$

From the equation in previous slide

$$\boxed{d = \frac{2}{||w||}}$$

⇒ And we have to maximize this margin value.

$$d = \frac{2}{\|w\|} \quad \text{given} \quad \boxed{y_i(w^T x_i + b) \geq 1}$$

$\boxed{\text{Hard margin}}$

⇒ Data → The formulation that we have formed → will not mark even if one of the point comes above or below the designated line.

All the positive points should be above the positive hyperplane and all the negative points should be below the negative hyperplane.

↓

For Hard margin

⇒ This → not possible practically.

→ no dataset can be obtained in real time which is perfectly linearly seperable.

→ so modify the equation in such a way that outliers are allowed in the system.

Soft margin SVM → handle outliers.

→ Solving this optimization problem.

→ so far the optimization problem that we have come across are unconstrained.

→ For example, consider the Linear Regression, Logistic Regression, where we are trying to minimize the cost function. But there is no constraint associated with it.

→ Here, we have to maximize the margin, that is we can choose for different value of w & b such that the margin is maximized, but we have to keep in mind that $y_i(x_i w^T + b) \geq 1$.

→ This is a hard margin problem.

→ So, to solve this, we have quadratic optimization problem with linear constraints, to Lagrangian multiplier.

→ Consider the following examples for solving Lagrangian multiplier, where we will be making use of a Lagrangian multiplier α, to solve for this,

Lagrange Multiplier Method

Equality constraint optimization problem solving

- The following steps are involved in this case:

1. Define the Lagrangian as follows:

$$(X, \lambda) = f(X) + \sum_{i=1}^{p} \lambda_i . g_i(x) \tag{18}$$

where $\lambda_i' s$ are dummy variables called Lagrangian multipliers.

2. Set the first order derivatives of the Lagrangian with respect to x and the Lagrangian multipliers $\lambda_i' s$ to zero's. That is

$$\frac{\delta L}{\delta x_i} = 0, i = 1, 2, ..., d$$
$$\frac{\delta L}{\delta \lambda_i} = 0, i = 1, 2, ..., p$$

3. Solve the $(d + p)$ equations to find the optimal value of $X = [x_1, x_2......, x_d]$ and $\lambda_i' s$.

Lagrange Multiplier Method

Example: Equality constraint optimization problem

Suppose, minimize $f(x, y) = x + 2y$
subject to $x^2 + y^2 - 4 = 0$

1. Lagrangian $L(x, y, \lambda) = x + 2y + \lambda(x^2 + y^2 - 4)$
2. $\frac{\delta L}{\delta x} = 1 + 2\lambda x = 0$
   $\frac{\delta L}{\delta y} = 1 + 2\lambda y = 0$
   $\frac{\delta L}{\delta \lambda} = x^2 + y^2 - 4 = 0$
3. Solving the above three equations for x, y and λ, we get $x = \mp\frac{2}{\sqrt{5}}$, $y = \mp\frac{4}{\sqrt{5}}$ and $\lambda = \pm\frac{\sqrt{5}}{4}$

Lagrange Multiplier Method

Example : Equality constraint optimization problem

- When $\lambda = \frac{\sqrt{5}}{4}$,
  $x = -\frac{2}{\sqrt{5}}$,
  $y = -\frac{4}{\sqrt{5}}$,
  we get f(x, y, λ)=$-\frac{10}{\sqrt{5}}$
- Similarly, when $\lambda = -\frac{\sqrt{5}}{4}$,
  $x = \frac{2}{\sqrt{5}}$,
  $y = \frac{4}{\sqrt{5}}$,
  we get f(x, y, λ)=$\frac{10}{\sqrt{5}}$
- Thus, the function f(x, y) has its minimum value at
  $x = -\frac{2}{\sqrt{5}}, y = -\frac{4}{\sqrt{5}}$

Lagrange Multiplier Method

Inequality constraint optimization problem solving

- The method for solving this problem is quite similar to the Lagrange multiplier method described above.
- It starts with the Lagrangian

$$L = f(x) + \sum_{i=1}^{p} \lambda_i . h_i(x) \tag{19}$$

- In addition to this, it introduces additional constraints, called **Karush-Kuhn-Tucker (KKT) constraints**, which are stated in the next slide.

Lagrange Multiplier Method

Inequality constraint optimization problem solving

$$\frac{\delta L}{\delta x_i} = 0, i = 1, 2, ..., d$$

$$\lambda_i \geq 0, i = 1, 2......, p$$

$$h_i(x) \leq 0, i = 1, 2......, p$$

$$\lambda_i . h_i(x) = 0, i = 1, 2......, p$$

Solving the above equations, we can find the optimal value of $f(x)$.

Lagrange Multiplier Method

Example for Reference

Example: Inequality constraint optimization problem

Consider the following problem.

Minimize $f(x, y) = (x-1)^2 + (y-3)^2$
subject to $x + y \leq 2$,
$y \geq x$

- The Lagrangian for this problem is
  $L = (x-1)^2 + (y-3)^2 + \lambda_1(x+y-2) + \lambda_2(x-y)$.

  subject to the KKT constraints, which are as follows:

Lagrange Multiplier Method

Example: Inequality constraint optimization problem

$$\frac{\delta L}{\delta x} = 2(x-1) + \lambda_1 + \lambda_2 = 0$$
$$\frac{\delta L}{\delta y} = 2(y-3) + \lambda_1 - \lambda_2 = 0$$
$$\lambda_1(x+y-2) = 0$$
$$\lambda_2(x-y) = 0$$
$$\lambda_1 \geq 0, \lambda_2 \geq 0$$
$$(x+y) \leq 2, y \geq x$$

Lagrange Multiplier Method

Example: Inequality constraint optimization problem

To solve KKT constraints, we have to check the following tests:

- Case 1: $\lambda_1 = 0, \lambda_2 = 0$

  $$2(x-1) = 0 \mid 2(y-3) = 0 \Rightarrow x = 1, y = 3$$

  since, $x + y = 4$, it violates $x + y \leq 2$; is not a feasible solution.

- Case 2: $\lambda_1 = 0, \lambda_2 \neq 0$ $2(x-y) = 0 \mid$
  $2(x-1) + \lambda_2 = 0 |$
  $2(y-3) - \lambda_2 = 0$
  $\Rightarrow x = 2, y = 2$ and $\lambda_2 = -2$
  since, $x + y \leq 4$, it violates $\lambda_2 \geq 0$; is not a feasible solution.

Lagrange Multiplier Method

Example: Inequality constraint optimization problem

- Case 3: $\lambda_1 \neq 0, \lambda_2 = 0$ $2(x + y) = 2$
  $2(x - 1) + \lambda_1 = 0$
  $2(y - 3) + \lambda_1 = 0$

  $\Rightarrow x = 0, y = 2$ and $\lambda_1 = 2$; this is a feasible solution.

- Case 4: $\lambda_1 \neq 0, \lambda_2 \neq 0$ $2(x + y) = 2$
  $2(x - y) = 0$
  $2(x - 1) + \lambda_1 + \lambda_2 = 0$
  $2(y - 3) + \lambda_1 - \lambda_2 = 0$

  $\Rightarrow x = 1, y = 1$ and $\lambda_1 = 2$ $\lambda_2 = -2$
  This is not a feasible solution.

LMM to Solve Linear SVM

- The optimization problem for the linear SVM is inequality constraint optimization problem.
- The Lagrangian multiplier for this optimization problem can be written as

$$L = \frac{||W||^2}{2} - \sum_{i=1}^{n} \lambda_i(y_i(W.x_i + b) - 1) \tag{20}$$

where the parameters $\lambda_i' s$ are the Lagrangian multipliers, and $W = [w_1, w_2 .... w_m]$ and b are the model parameters.