- In general, Distance of a point from a plane (d) is

$$d = \frac{W^T P}{\|W\|}$$

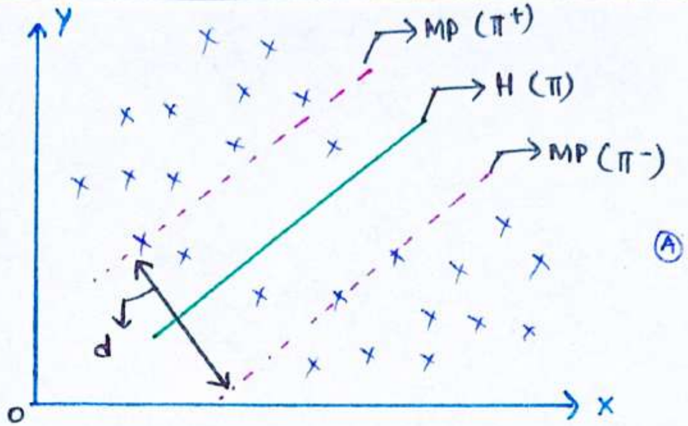
or

$$\boxed{d = \|W\| \cdot \|P\| \cdot \cos\theta}$$

- **Observation:**
  1. Points above the plane make θ_1 angle with $\hat{w}$ and from formula and fact that θ_1 is always less than 90° as $\hat{w} \perp \pi$
     $\therefore$ Value of $\cos\theta$ for $\theta < 90$ is always +ve.
     ⇒ Points above plane always have +ve 'd'.
  2. Points below plane make θ_2 angle with $\hat{w}$ and from formula and fact θ_2 always more than 90°
     $\therefore$ value of $\cos\theta$ for $\theta > 90$ is always -ve.
     ⇒ Points below plane always have -ve 'd'

C Support Vector Classifier

- **Support Vectors (SV)**: Data points that are closer to hyperplane and influence the position and orientation of hyperplane.
- **Marginal Plane (MP)**: Planes closer to the SV are marginal plane and help in choosing hyperplane.
- **Margin (d)**: Distance between two marginal plane.
- **Hyperplane (H)**: Best plane which clearly seperates data points with highest margin.



- Our aim is to find best plane which can clearly sepperate data points.
- This plane has maximum margin and called hyperplane (π).
- Our goal is to maximize the margin (d), classifier using such methodology is called **maximal margin classifier** and that maximum margin plane is called **margin maximizing plane**.
- Maximizing margin (M) given,

$$c + x_1 + x_2 + \dots + x_n \quad \text{— (i)}$$

$$y_i(c + w_1x_{1i} + w_2x_{2i} + \dots + w_nx_{ni}) \geq M \;,\; i = 1 \dots n \quad \text{— (ii)}$$

- This equation means define margin M by tunning coefficents of all variables such that margin is maximized(ii) and product of predicted (observed) value with equation of respective input features should be greater than margin(ii)

- If hyperplane able to clearly sepperate data points like fig A then it is called **Hard margin SVC**.
  In such case,

$$\pi^+ = W^T x_1 + c = +1$$

$$\pi^- = W^T x_2 + c = -1$$

  If we add these two we get,

$$d = \frac{W^T(x_1 - x_2)}{\|W\|}$$

$$\therefore \boxed{d = \frac{2}{\|W\|}}$$

  as $(\pi^+ + \pi^-) = W^T(x_1 - x_2) = 2$

- So, $d = 2/\|W\|$ is margin in case of hard margin classifier.
- We need to maximize 'd' by changin coefficents of features in matrix X which are present in matrix W^T.
- This margin is basically observed values distances under constraint,

$$y_i \begin{cases} 1 \;, & W^T x + c \geq 1 \\ -1 \;, & W^T x + c \leq 1 \end{cases}$$

  for all points which is basically error,
- Combining constraints we get,

$$\boxed{y_i(W^T x + c) \geq 1}$$