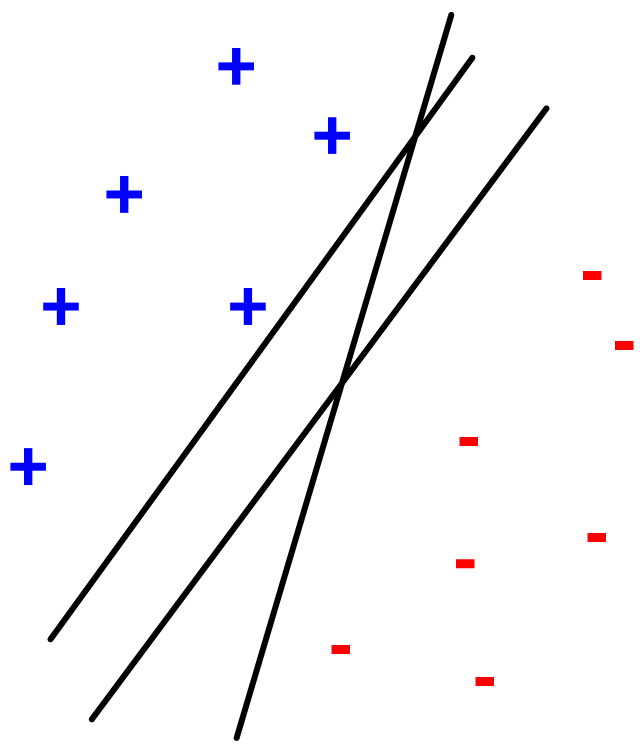
Support Vector Machines

- Want to separate "+" from "-" using a line



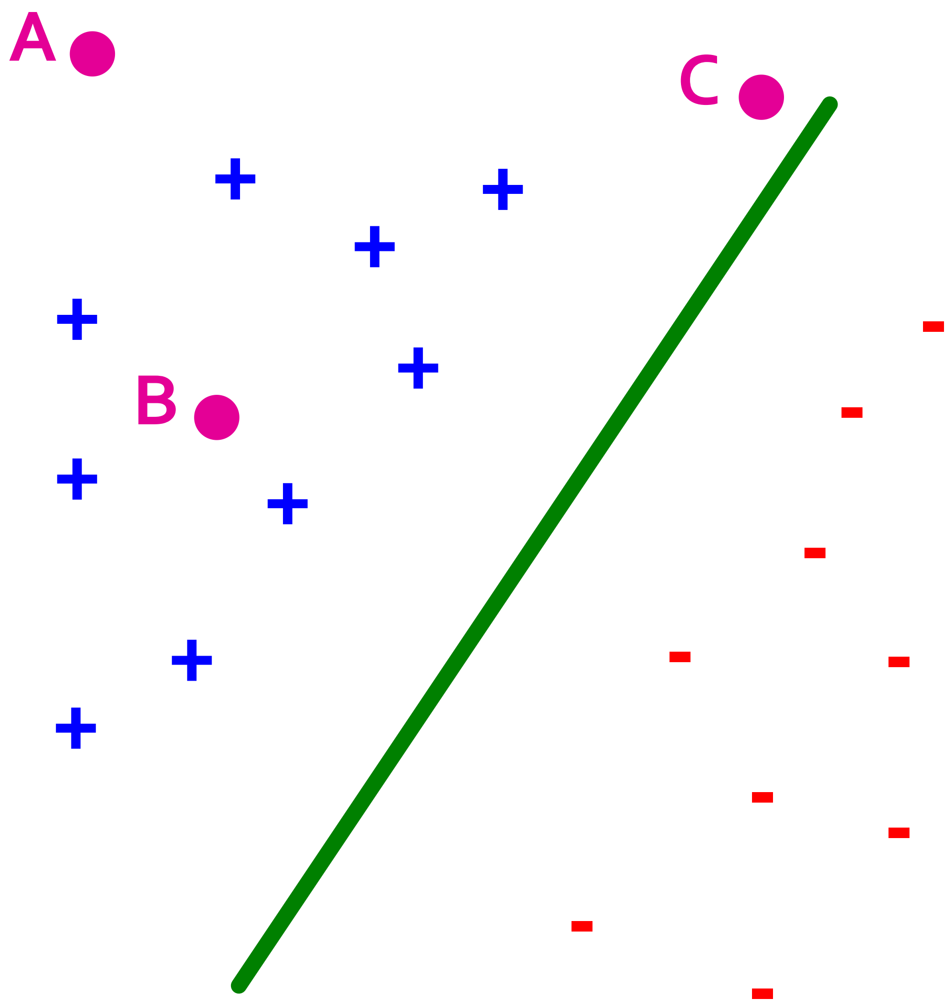
Data:

- **Training examples:**
  - $(x_1, y_1) \ldots (x_n, y_n)$
- **Each example *i*:**
  - $x_i = (x_i^{(1)}, \ldots , x_i^{(d)})$
    - $x_i^{(j)}$ is real valued
  - $y_i \in \{ -1, +1 \}$
- **Inner product:**

$$\boldsymbol{w} \cdot \boldsymbol{x} = \sum_{j=1}^{d} w^{(j)} \cdot x^{(j)}$$

Which is best linear separator (defined by *w*)?

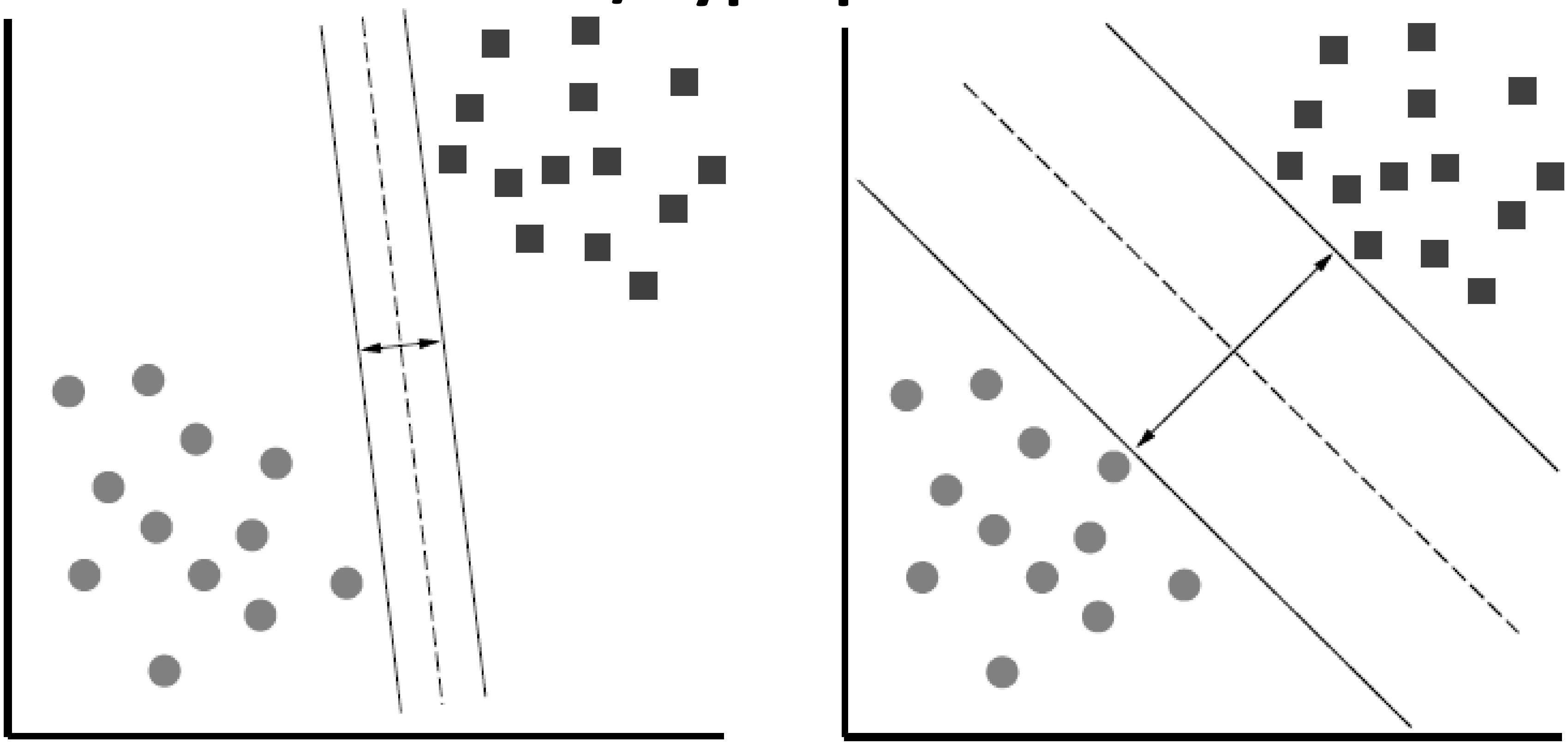
Largest Margin



- **Distance from the separating hyperplane corresponds to the "confidence" of prediction**
- Example:
  - We are more sure about the class of **A** and **B** than of **C**

Largest Margin

- **Margin γ: Distance of closest example from the decision line/hyperplane**

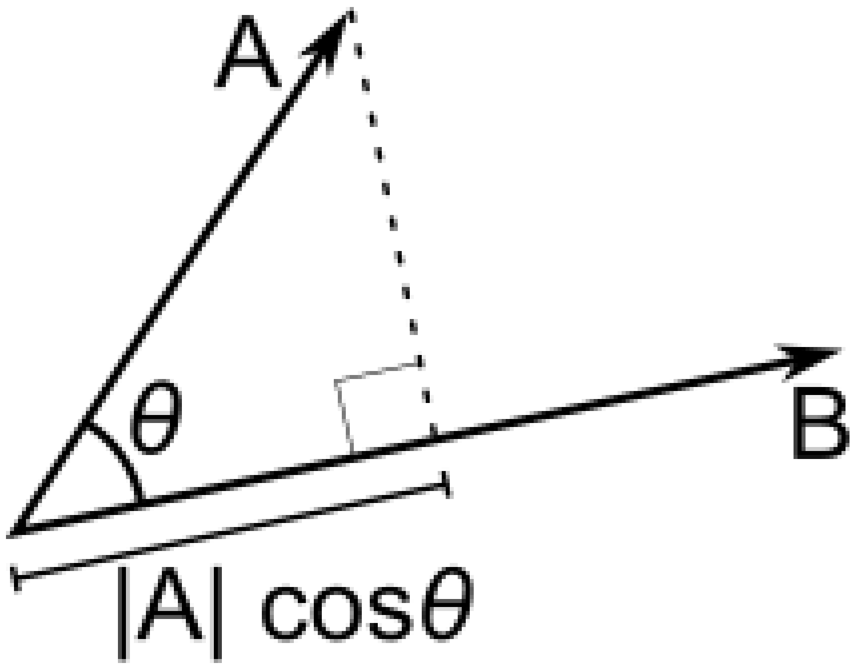


The reason we define margin this way is due to theoretical convenience and existence of generalization error bounds that depend on the value of margin.

Why maximizing γ a good idea?

- **Remember: Dot product**

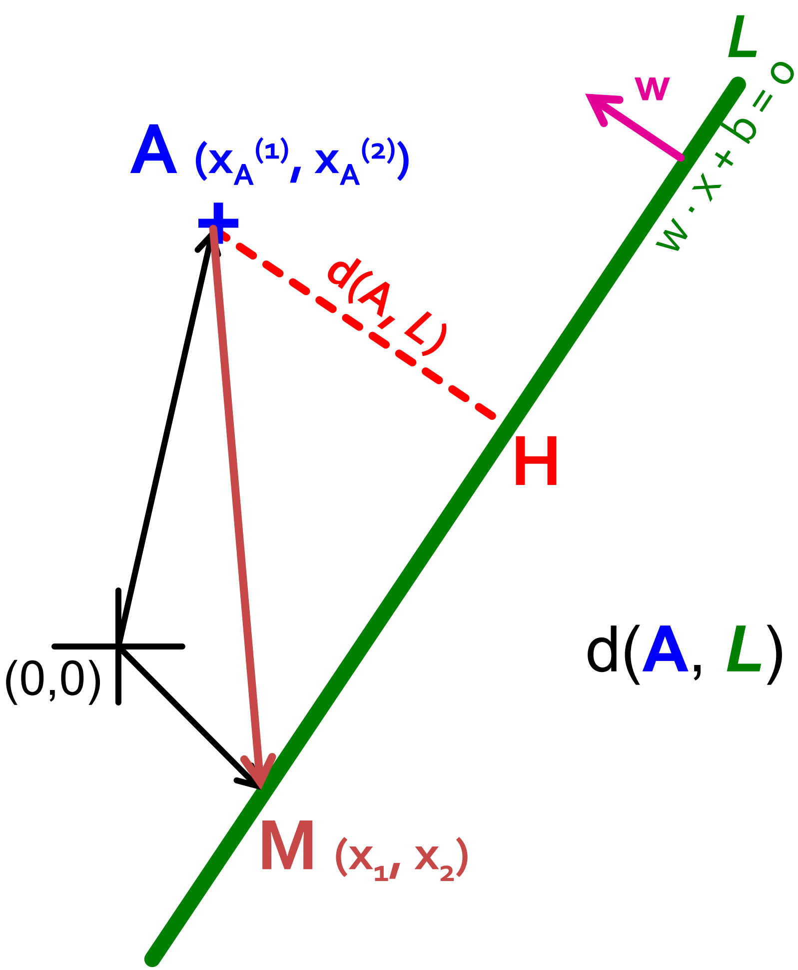
$$A \cdot B = \|A\| \|B\| \cos\theta$$



$$||A|| = \sqrt{\sum_{j=1}^{d} (A^{(j)})^2}$$

What is the margin?

Note we assume $\|w\|_2 = 1$

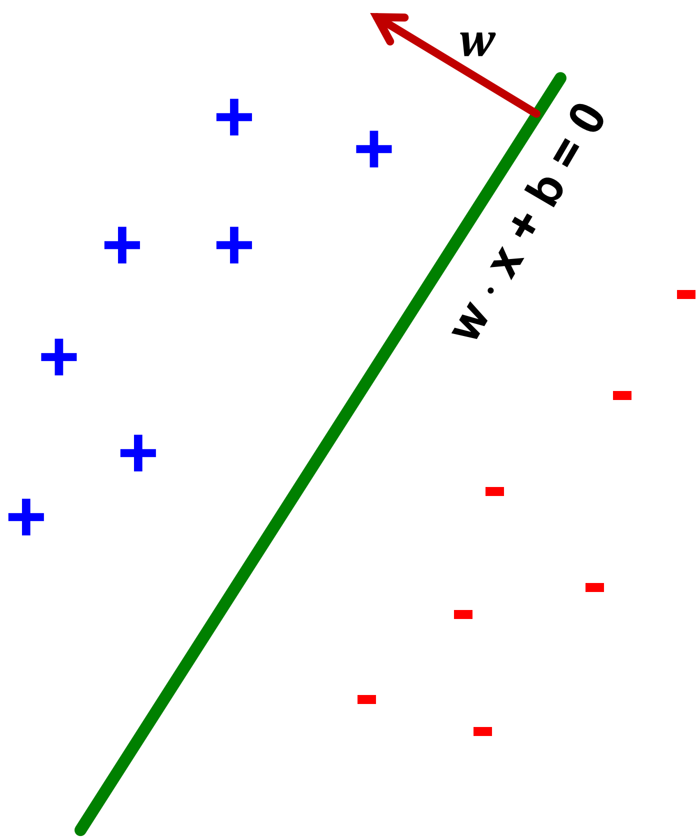


- **Let:**
  - **Line L: w·x+b =** $w^{(1)}x^{(1)}+w^{(2)}x^{(2)}+b=0$
  - **w** = $(w^{(1)}, w^{(2)})$
  - **Point A** = $(x_A^{(1)}, x_A^{(2)})$
  - **Point M** on a line = $(x_M^{(1)}, x_M^{(2)})$

$$
\begin{aligned}
d(A, L) &= |AH| \\
&= |(A-M) \cdot w| \\
&= |(x_A^{(1)} - x_M^{(1)})\, w^{(1)} + (x_A^{(2)} - x_M^{(2)})\, w^{(2)}| \\
&= x_A^{(1)} w^{(1)} + x_A^{(2)} w^{(2)} + b \\
&= w \cdot A + b
\end{aligned}
$$

Remember $\mathbf{x_M^{(1)}w^{(1)} + x_M^{(2)}w^{(2)} = - b}$ since **M** belongs to line **L**

Largest Margin



- Prediction = $\mathbf{sign}(w \cdot x + b)$
- "**Confidence**" = $(w \cdot x + b)\, y$
- **For i-th datapoint:**

$$\gamma_i = (w \cdot x_i + b) y_i$$

- **Want to solve:**

$$\max_{w} \min_{i} \gamma_i$$

- **Can rewrite as**

$$\max_{w,\gamma} \gamma$$

$$s.t. \forall i, y_i (w \cdot x_i + b) \geq \gamma$$

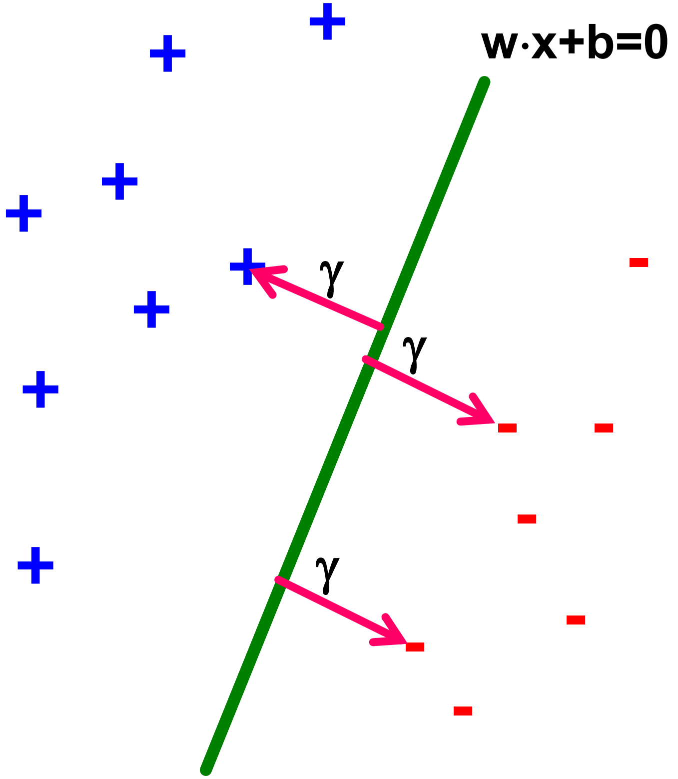
Support Vector Machine

- **Maximize the margin:**
  - **Good according to intuition, theory (VC dimension) & practice**

$$\max_{w,\gamma} \gamma$$

$$s.t. \forall i, y_i (w \cdot x_i + b) \geq \gamma$$

  - **γ is margin … distance from the separating hyperplane**



Maximizing the margin