Support Vector Machines (SVM)

Data Intelligence and Learning (DIAL) Lab
Prof. Jongwuk Lee

Support Vector Machines Basics

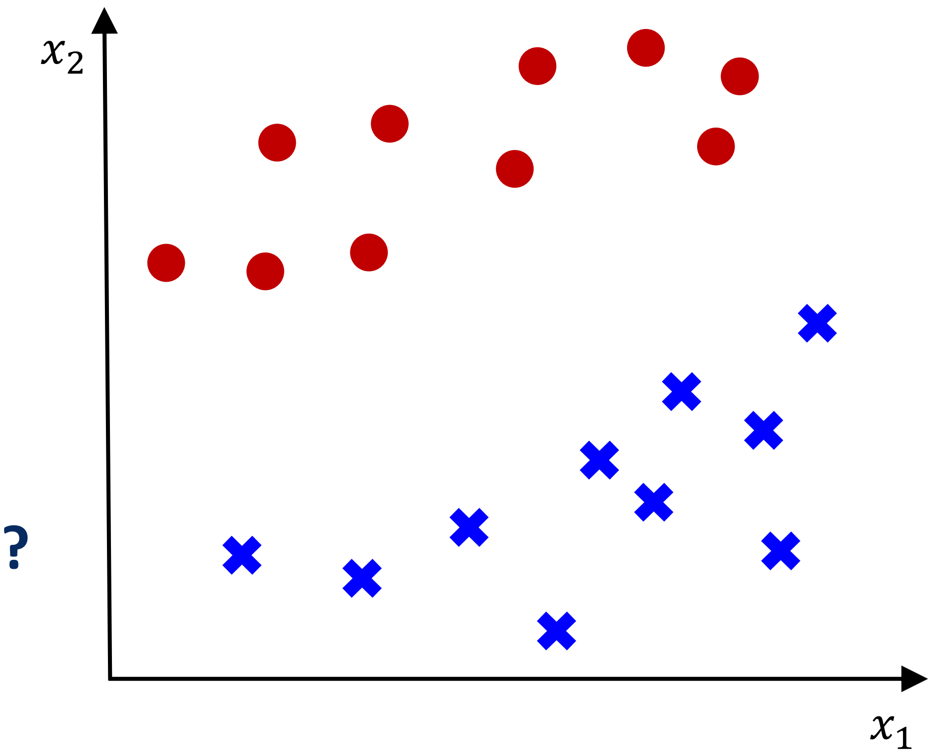
Classification using Linear Models

- **Binary classification: output** $y \in \{-1, +1\}$

- **Formulating a linear model**

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

$$\hat{y} = \mathrm{sign}(\mathbf{z})$$



- **How do we choose w and** b**?**

0-1 Loss Function

- **We can use the 0-1 loss function.**

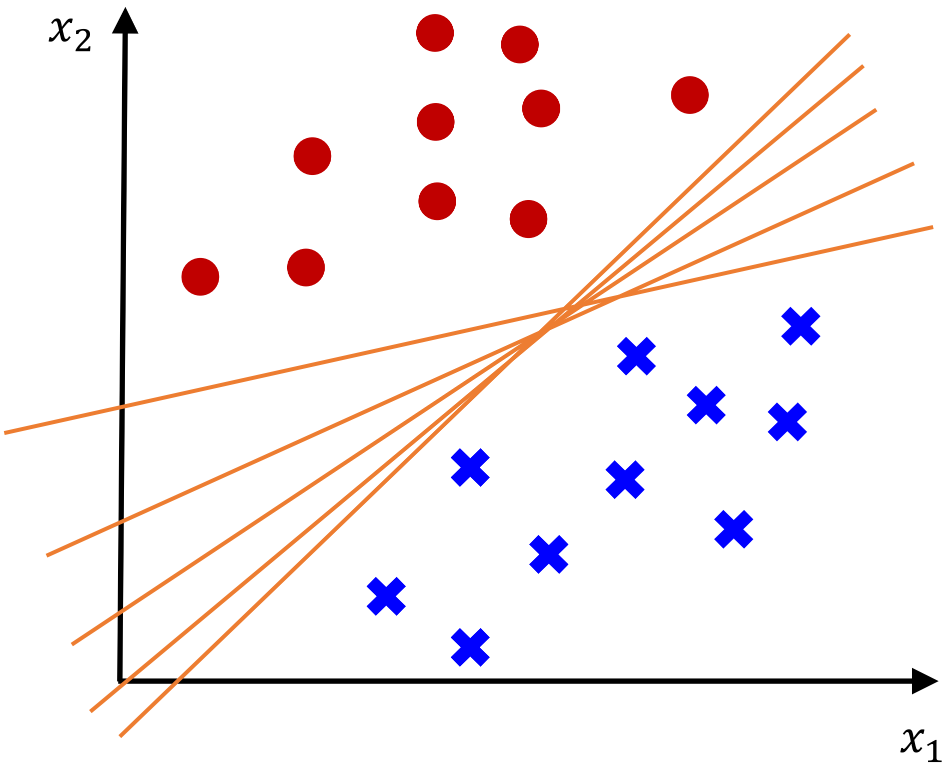
$$\mathcal{L}_{0-1}(h(\mathbf{x}), y) = \mathbb{I}[h(\mathbf{x}) \neq y] = \begin{cases} 0 & \text{if } h(\mathbf{x}) = y \\ 1 & \text{otherwise} \end{cases}$$

- **However, it does not distinguish different hypotheses that achieve the same accuracy.**
  - The cross-entropy loss $\mathcal{L}_{\text{CE}}$ can address this problem.

- **Can we use a different approach using the geometry of binary classifiers?**

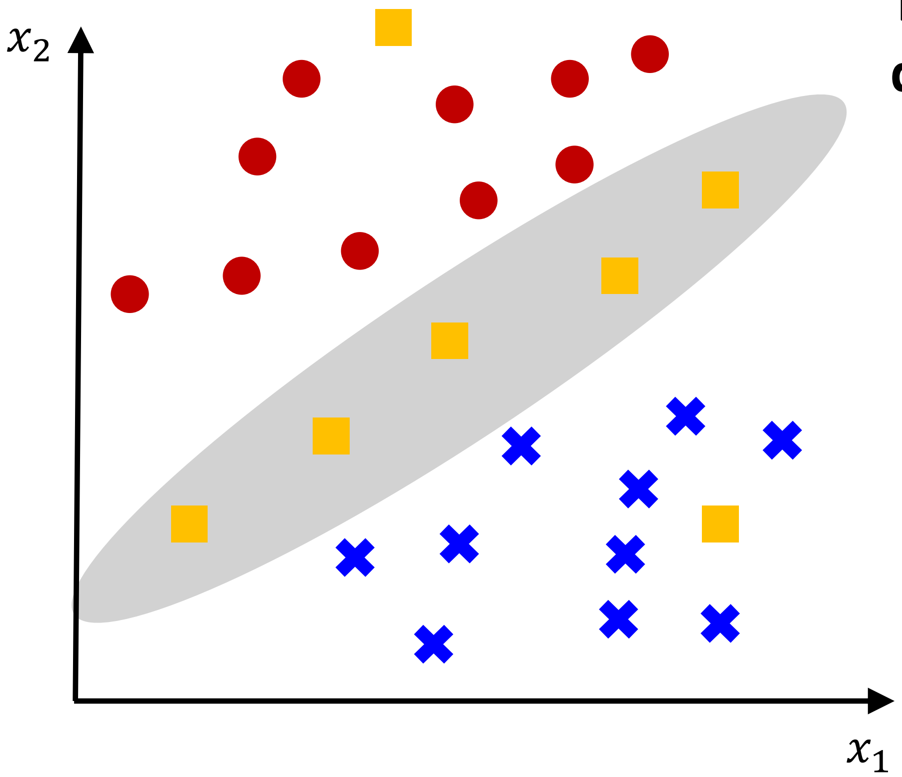
Motivation

➢ **What is the best decision boundary without using the probability distribution?**



Motivation

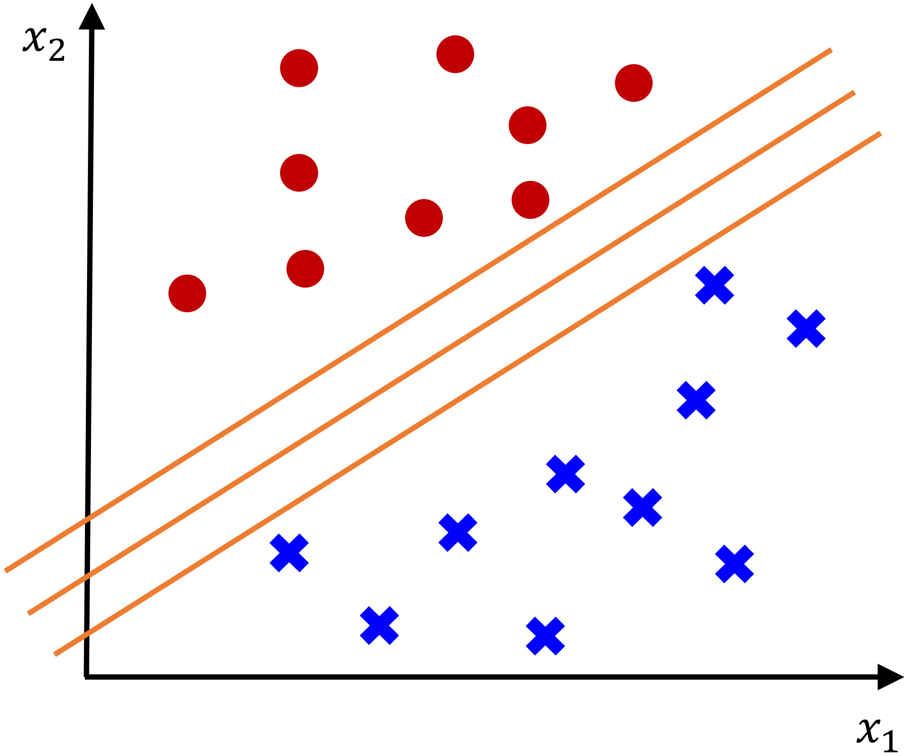
➢ Given an unknown data sample, which does it belong to?



The samples in the gray area depend on a decision boundary.

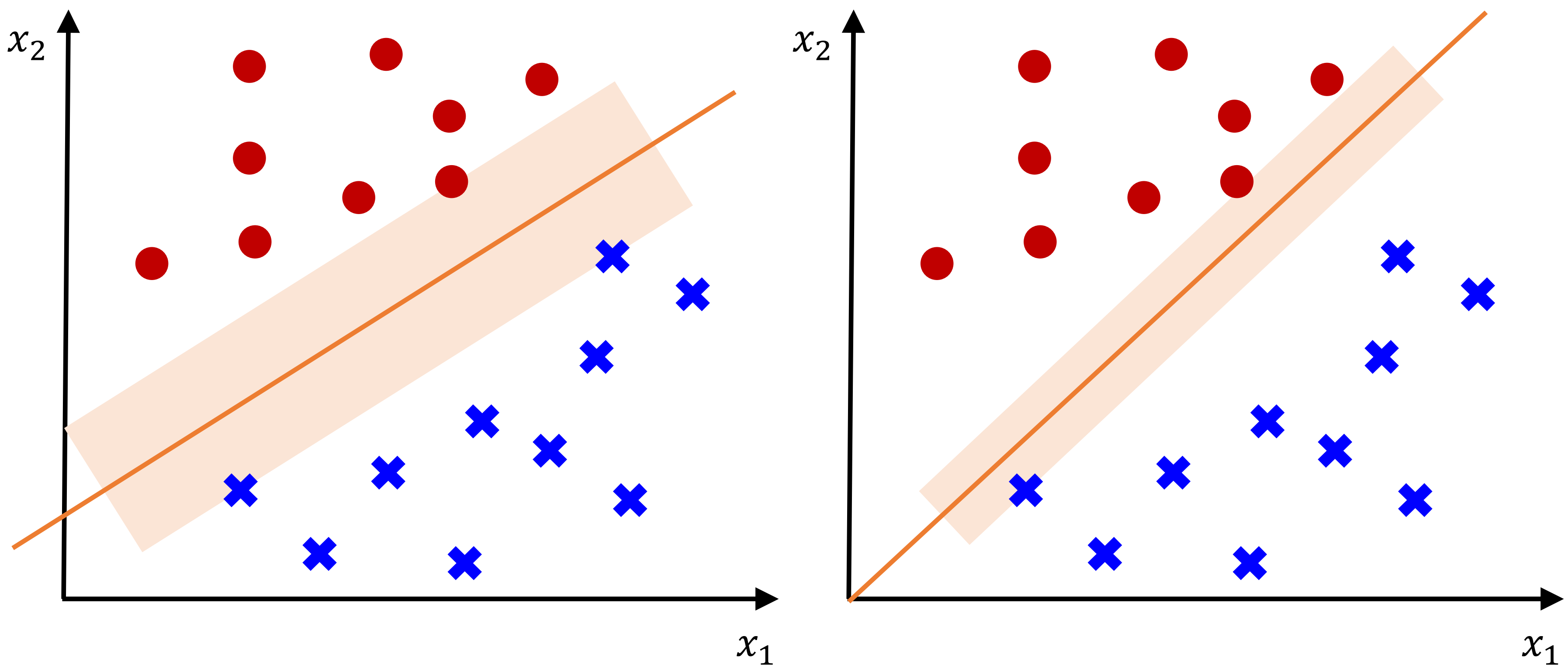
Motivation

- Q: Which is the best?
- A: The central position line is the best.



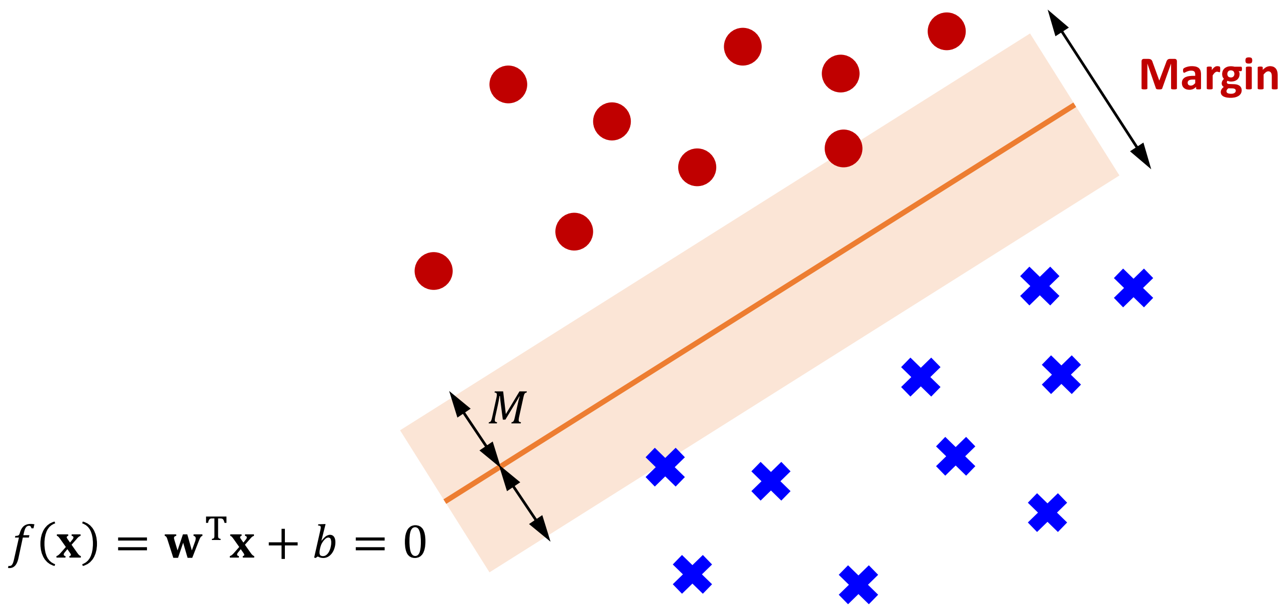
Motivation

- Q: What is better?
- A: Maximizing the margin is better.



Optimal Separating Hyperplane

- **A hyperplane that separates two classes and maximizes the distance from the closest point to either class.**
- **It maximizes the margin of the classifier.**
  - It helps achieve **better generalization** of test data.



How to Calculate the Margin?

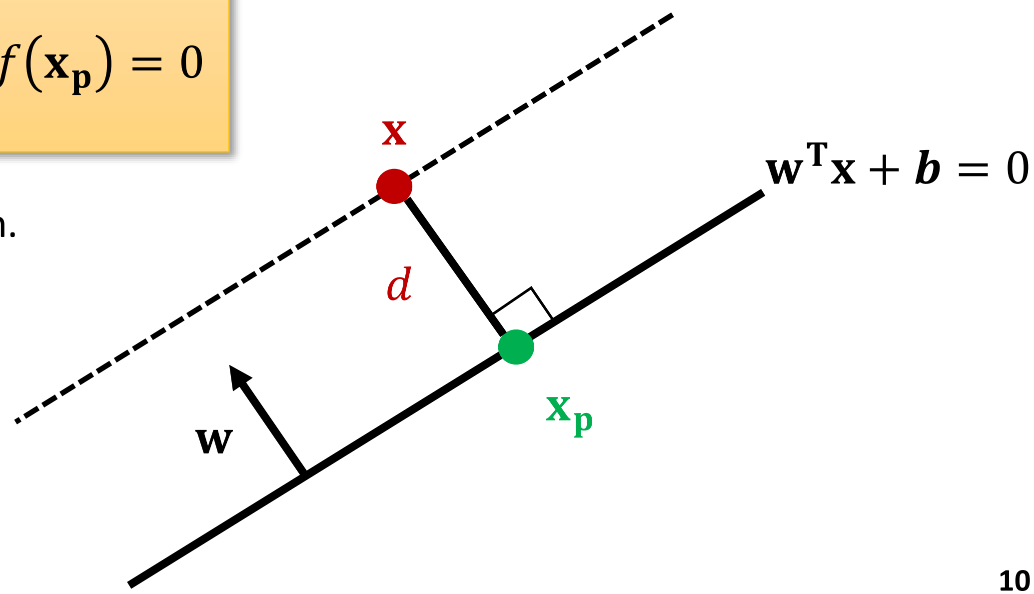
- **Given the decision boundary $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$,**
  - A point $\mathbf{x}$ on the boundary has $f(\mathbf{x}) = \mathbf{w}^{T}\mathbf{x} + \boldsymbol{b} = 0$.
  - A positive point $\mathbf{x}$ has $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = a$ where $a > 0$.

- **The perpendicular line from $\mathbf{x}$ to the $f(\mathbf{x})$ is**

$$\mathbf{x} = \mathbf{x}_{\mathbf{p}} + d\,\frac{\mathbf{w}}{\|\mathbf{w}\|_2}, \qquad \text{where } f(\mathbf{x}_{\mathbf{p}}) = 0$$

Let $\|\mathbf{w}\|_2 = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$ be the L2-norm.

$\frac{\mathbf{w}}{\|\mathbf{w}\|_2}$ is the **unit vector**.



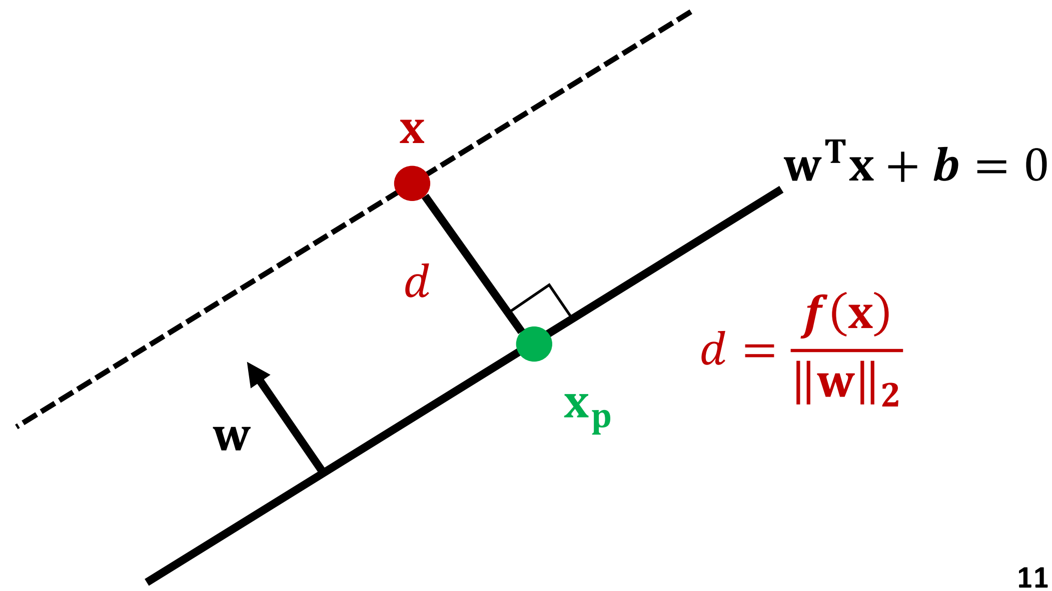
How to Calculate the Margin?

- Given a point $\mathbf{x}$ and $f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b$,

$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = \mathbf{w}^{\mathrm{T}}\left(\mathbf{x}_{\mathrm{p}} + d\frac{\mathbf{w}}{\|\mathbf{w}\|_2}\right) + b = \mathbf{w}^{\mathrm{T}}\mathbf{x}_{\mathrm{p}} + b + d\frac{\mathbf{w}^{\mathrm{T}}\mathbf{w}}{\|\mathbf{w}\|_2} = d \cdot \|\mathbf{w}\|_2$$

$$\|\mathbf{w}\|_2 = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$$

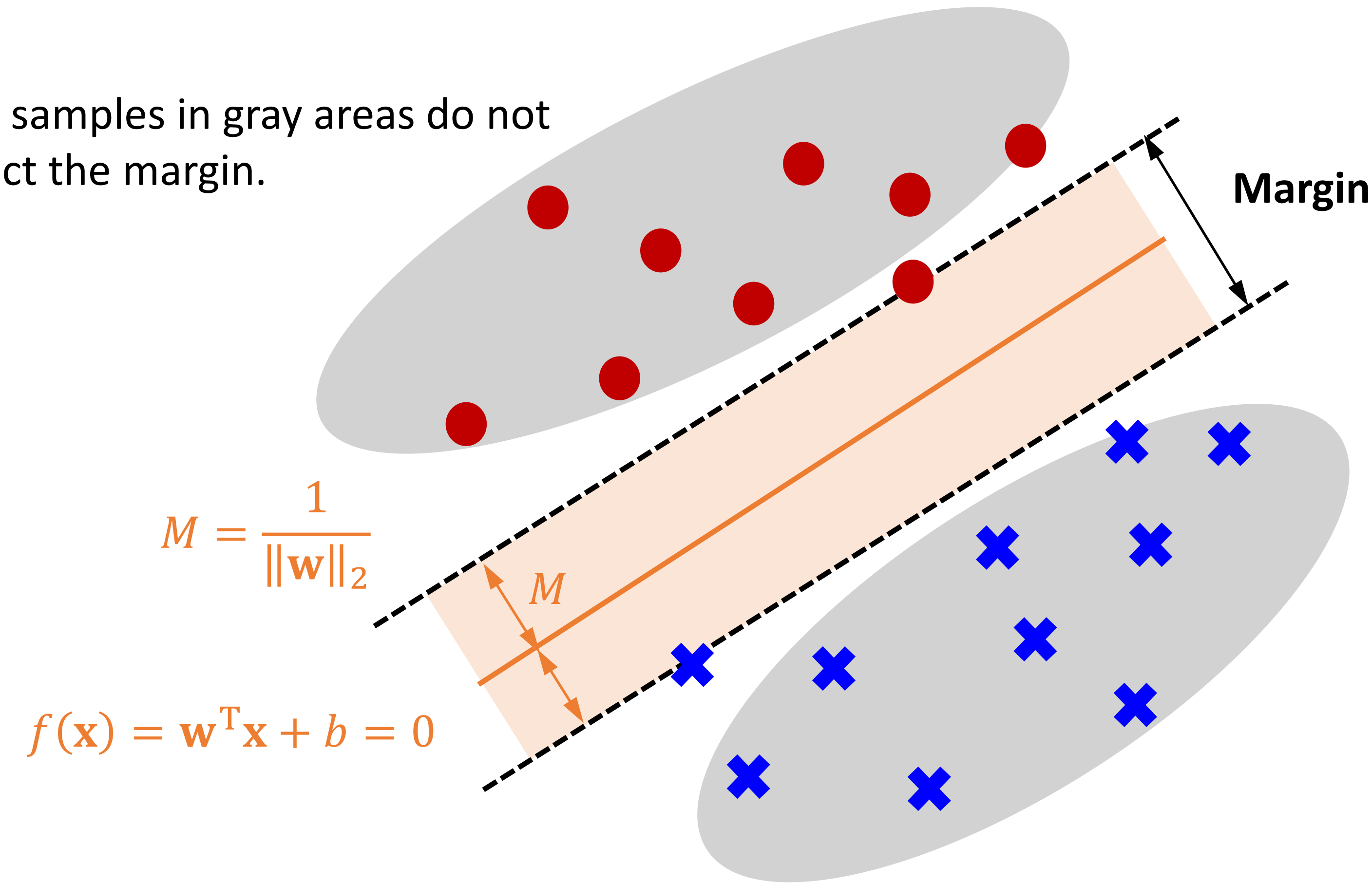
- The distance d from $\mathbf{x}$ to $\mathbf{x}_{\mathrm{p}}$ is $\frac{f(\mathbf{x})}{\|\mathbf{w}\|_2}$.



How to Maximize the Margin?

- If **the margin is not tight** for $\mathbf{x}^{(i)}$, **we can remove it from a training set**, and an optimal $\mathbf{w}$ would be the same.

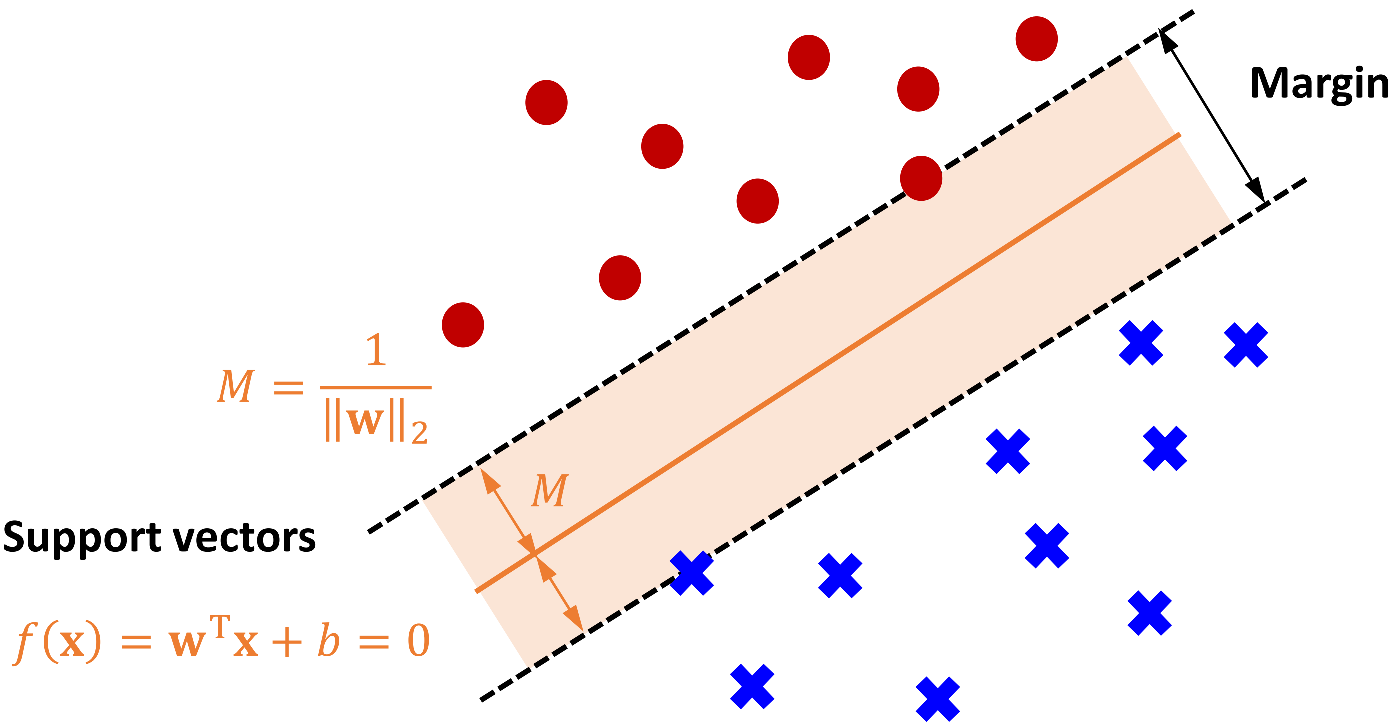
The samples in gray areas do not affect the margin.



Linear Support Vector Machines

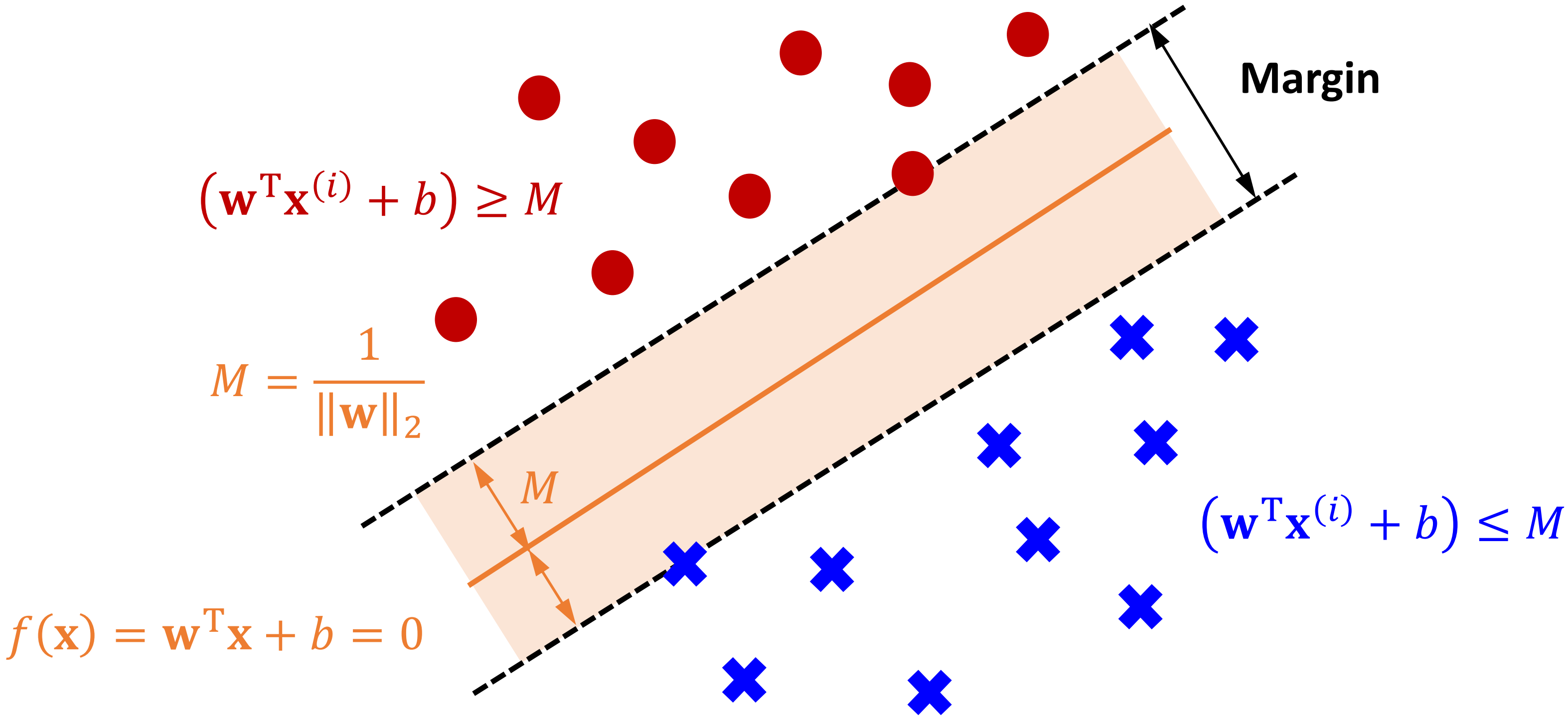
- **Support vectors** are the sample **closest** to a decision surface (or hyperplane).

- They are the samples **most difficult** to classify.



Learning Linear SVM

- **We want to maximize the margin for all the samples.**
  - If $y^{(i)} = +1$, then $y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \geq M$.
  - If $y^{(i)} = -1$, then $y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \geq M$.



Formulating Linear SVM

- The margin for $\mathcal{D}$ is

$$\min_{\mathbf{x}\in\mathcal{D}} \frac{|f(\mathbf{x})|}{\|\mathbf{w}\|_2} = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x} + b|}{\|\mathbf{w}\|_2}$$

- The classification for the i-th sample is correct

$$\operatorname{sign}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) = y^{(i)} \quad \Leftrightarrow \quad y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) > 0$$

- **Enforcing a margin of M**

$$y^{(i)} \cdot \frac{\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)}{\|\mathbf{w}\|_2} \geq M$$

Formulating Linear SVM

➢ Objective function

$$\max_{\mathbf{w},b} 2M \quad \text{such that} \quad \frac{y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)}{\|\mathbf{w}\|_2} \geq M, \qquad \forall i \in [1, n]$$

Let $2M$ be the margin size.

- We can scale $\mathbf{w}$ and b by any positive value and represent the same decision boundary.
- It is also possible to enforce $\|\mathbf{w}\|_2 = d$ for any $d > 0$ without changing the original solution.

➢ We can add a constraint $\|\mathbf{w}\|_2 = 1/M$.

Formulating Linear SVM

- By plugging $M = 1/\|\mathbf{w}\|_2$,

Because $\|\mathbf{w}\|_2 > 0$

$$\frac{y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)}{\|\mathbf{w}\|_2} \geq \frac{1}{\|\mathbf{w}\|_2} \quad \Leftrightarrow \quad y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \geq 1$$

- It is equivalent to the following objective function.
  - This is the **quadratic optimization** problem.

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|_2} \text{ subject to } \frac{y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right)}{\|\mathbf{w}\|_2} \geq M \text{ for } i = 1, 2, \cdots, n$$

$$\Downarrow$$

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \text{ subject to } y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \geq 1 \text{ for } i = 1, 2, \cdots, n$$

Example: Training Linear SVM

- Let $\mathcal{D} = \{(1, 1, -1), (2, 2, +1)\}$.

Objective function

$$\min_{\mathbf{w},b} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \text{ subject to } y^{(i)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(i)} + b\right) \geq 1 \text{ for } i = 1, 2, \cdots, n$$

- Apply the dataset for the objective function.

$$\min_{\mathbf{w},b} \frac{1}{2}(w_1^2 + w_2^2)\ s.t.\ (w_1 + w_2 + b) + 1 \leq 0 \text{ and } (-2w_1 - 2w_2 - b) + 1 \leq 0$$

Example: Training Linear SVM

- **Introduce Lagrange multipliers.**

$$\min_{\mathbf{w},b} \frac{1}{2}(w_1^2 + w_2^2)\ s.t.\ (w_1 + w_2 + b) + 1 \leq 0 \text{ and } (-2w_1 - 2w_2 - b) + 1 \leq 0$$

⬇

$$\mathrm{L}(\mathrm{w}, \mathrm{b}, \lambda) = \frac{1}{2}(w_1^2 + w_2^2) + \mu_1(w_1 + w_2 + b + 1) + \mu_2(-2w_1 - 2w_2 - b + 1)$$

⬇

$$\frac{\partial L}{\partial w_1} = w_1 + \mu_1 - 2\mu_2 = 0$$

$$\frac{\partial L}{\partial w_2} = w_2 + \mu_1 - 2\mu_2 = 0$$

$$\mu_1(w_1 + w_2 + b + 1) = 0$$

$$\mu_2(-2w_1 - 2w_2 - b + 1) = 0$$

$$w_1 + w_2 + b + 1 \leq 0$$

$$-2w_1 - 2w_2 - b + 1 \leq 0$$

Example: Training Linear SVM

$$\frac{\partial L}{\partial w_1} = w_1 + \mu_1 - 2\mu_2 = 0$$

$$\frac{\partial L}{\partial w_2} = w_2 + \mu_1 - 2\mu_2 = 0$$

$$\mu_1(w_1 + w_2 + b + 1) = 0$$
$$\mu_2(-2w_1 - 2w_2 - b + 1) = 0$$
$$w_1 + w_2 + b + 1 \leq 0$$
$$-2w_1 - 2w_2 - b + 1 \leq 0$$

⬇

C1: $\boldsymbol{\mu_1 = 0, \mu_2 = 0}$

$$w_1 = 0$$
$$w_2 = 0$$
$$w_1 + w_2 + b + 1 \leq 0$$
$$-2w_1 - 2w_2 - b + 1 \leq 0$$

C2: $\boldsymbol{\mu_1 = 0, \mu_2 \neq 0}$

$$w_1 - 2\mu_2 = 0$$
$$w_2 - 2\mu_2 = 0$$
$$w_1 + w_2 + b + 1 \leq 0$$
$$-2w_1 - 2w_2 - b + 1 = 0$$

C3: $\boldsymbol{\mu_1 \neq 0, \mu_2 = 0}$

$$w_1 + \mu_1 = 0$$
$$w_2 + \mu_1 = 0$$
$$w_1 + w_2 + b + 1 = 0$$
$$-2w_1 - 2w_2 - b + 1 \leq 0$$

C4: $\boldsymbol{\mu_1 \neq 0, \mu_2 \neq 0}$

$$w_1 + \mu_1 - 2\mu_2 = 0$$
$$w_2 + \mu_1 - 2\mu_2 = 0$$
$$w_1 + w_2 + b + 1 = 0$$
$$-2w_1 - 2w_2 - b + 1 = 0$$

Example: Training Linear SVM

- Solve each subproblem and check the feasible solution.

C1: $\boldsymbol{\mu_1 = 0, \mu_2 = 0}$	C2: $\boldsymbol{\mu_1 = 0, \mu_2 \neq 0}$	C3: $\boldsymbol{\mu_1 \neq 0, \mu_2 = 0}$	C4: $\boldsymbol{\mu_1 \neq 0, \mu_2 \neq 0}$
$w_1 = 0$	$w_1 - 2\mu_2 = 0$	$w_1 + \mu_1 = 0$	$w_1 + \mu_1 - 2\mu_2 = 0$
$w_2 = 0$	$w_2 - 2\mu_2 = 0$	$w_2 + \mu_1 = 0$	$w_2 + \mu_1 - 2\mu_2 = 0$
$w_1 + w_2 + b + 1 \leq 0$	$w_1 + w_2 + b + 1 \leq 0$	$w_1 + w_2 + b + 1 = 0$	$w_1 + w_2 + b + 1 = 0$
$-2w_1 - 2w_2 - b + 1 \leq 0$	$-2w_1 - 2w_2 - b + 1 = 0$	$-2w_1 - 2w_2 - b + 1 \leq 0$	$-2w_1 - 2w_2 - b + 1 = 0$

C1:

$$b \leq -1$$
$$b \geq 1$$

Infeasible!

C2:

$$w_1 = 2\mu_2$$
$$w_2 = 2\mu_2$$
$$b = -8\mu_2 + 1$$
$$\mu_2 \geq 0.5$$

C3:

$$w_1 = -\mu_1$$
$$w_2 = -\mu_1$$
$$b = 2\mu_1 - 1$$
$$\mu_1 \leq -1$$

C4:

$$w_1 = w_2$$
$$b = 2w_1 - 1$$
$$w_1 = 1$$

$$w_1 = 1$$
$$w_2 = 1$$
$$b = -3$$

Example: Training Linear SVM

- When $\mu_1 \neq 0, \mu_2 \neq 0$, we can a solution.

$$w_1 = 1$$
$$w_2 = 1$$
$$b = -3$$

- It implies that two samples are **support vectors**.

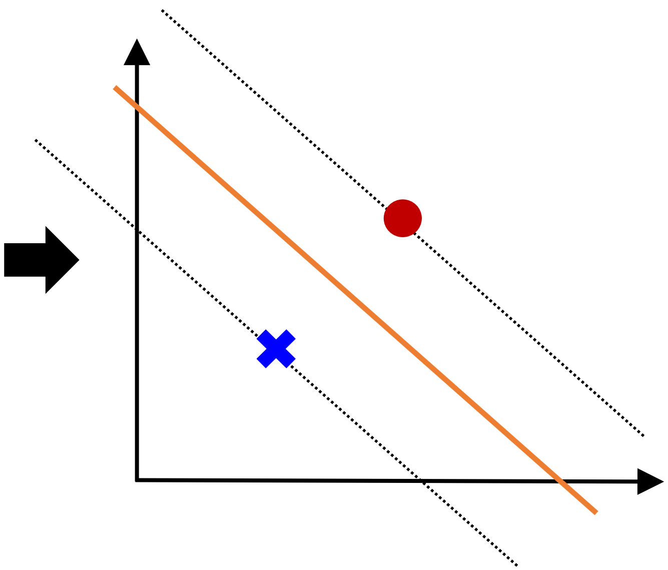
C4: $\boldsymbol{\mu_1 \neq 0, \mu_2 \neq 0}$

$$w_1 + \mu_1 - 2\mu_2 = 0$$
$$w_2 + \mu_1 - 2\mu_2 = 0$$
$$w_1 + w_2 + b + 1 = 0$$
$$-2w_1 - 2w_2 - b + 1 = 0$$

$$\Rightarrow$$

$$y^{(1)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(1)} + b\right) = 1$$
$$y^{(2)}\left(\mathbf{w}^{\mathrm{T}}\mathbf{x}^{(2)} + b\right) = 1$$

$$\Rightarrow$$



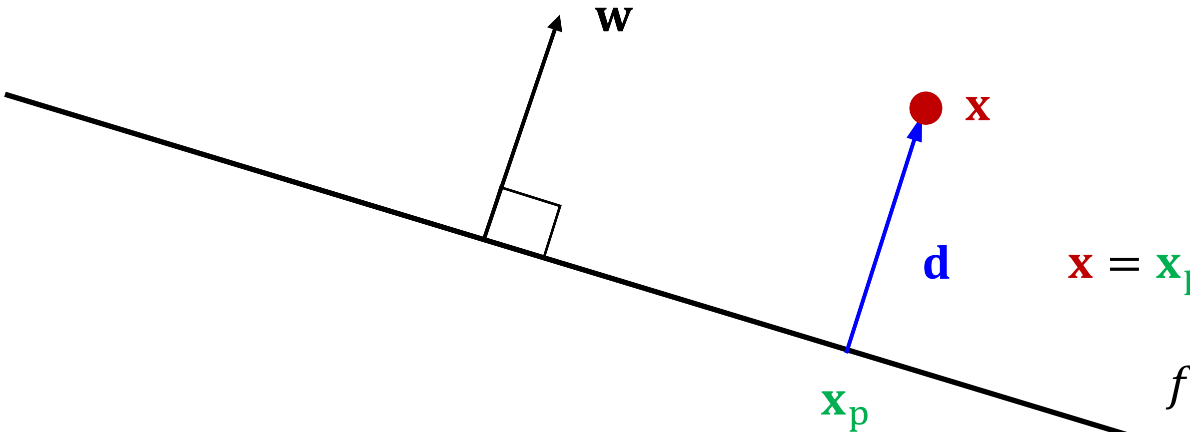
$f(x) = x_1 + x_2 - 3$

Q&A

Geometry of Points and Planes

- Recall that the decision hyperplane is perpendicular to $\mathbf{w}$.

- How to compute the distance $\|\mathbf{d}\|_2 = \sqrt{\mathbf{d}^{\mathrm{T}}\mathbf{d}}$?



$$\mathbf{x} = \mathbf{x}_{\mathrm{p}} + \mathbf{d}$$

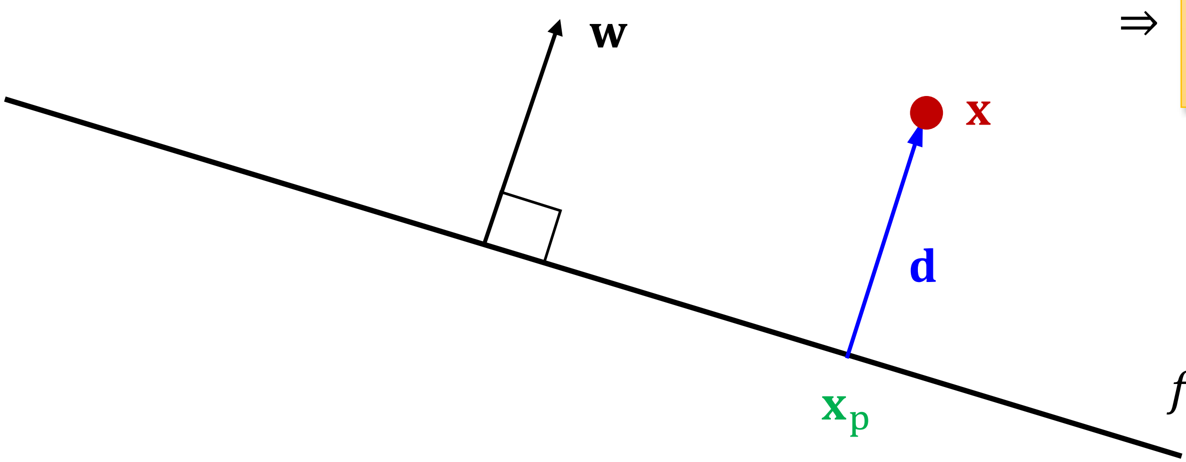
$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$

Geometry of Points and Planes

- **It follows the equation $\mathbf{x} = \mathbf{x_p} + \mathbf{d}$.**
  - Let $\mathbf{x_p}$ be a projected point of $\mathbf{x}$ onto $f(\mathbf{x})$.

- **Since the distance d is parallel to $\mathbf{w}^*$, $r = \alpha\mathbf{w}$.**
  - $\mathbf{x_p} = \mathbf{x} - \mathbf{d} = \mathbf{x} - \alpha\mathbf{w}$
  - Since $\mathbf{x_p} \in \mathcal{H}$, $f(\mathbf{x_p}) = \mathbf{w}^{\mathrm{T}}\mathbf{x_p} + b = \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \alpha\mathbf{w}) + b = 0$

$$\Rightarrow \boxed{\alpha = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{x} + b}{\mathbf{w}^{\mathrm{T}}\mathbf{w}}}$$



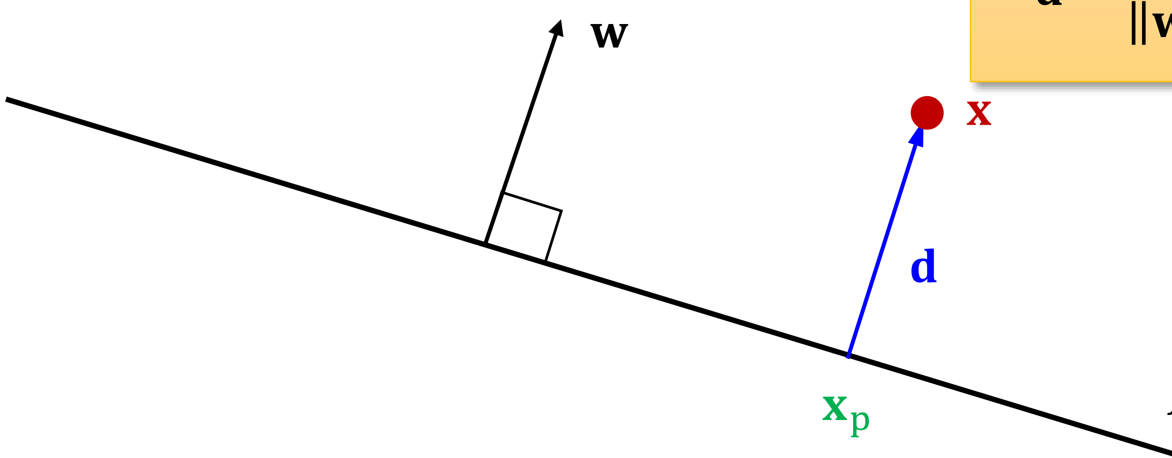
$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$

Geometry of Points and Planes

- **The length of d is**
  - $\|\mathbf{d}\|_2 = \sqrt{\mathbf{d}^{\mathrm{T}}\mathbf{d}} = \sqrt{\alpha^2 \mathbf{w}^{\mathrm{T}}\mathbf{w}} = \sqrt{\left(\frac{\mathbf{w}^{\mathrm{T}}\mathbf{x}+b}{\mathbf{w}^{\mathrm{T}}\mathbf{w}}\right)^2 \mathbf{w}^{\mathrm{T}}\mathbf{w}} = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}+b|}{\sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}} = \frac{|\mathbf{w}^{\mathrm{T}}\mathbf{x}+b|}{\|\mathbf{w}\|_2}$

- **The (signed) distance of a point $\mathbf{x}$ to the hyperplane is**

$$\mathbf{d} = \frac{f(\mathbf{x})}{\|\mathbf{w}\|_2} = \frac{\mathbf{w}^{\mathrm{T}}\mathbf{x} + b}{\|\mathbf{w}\|_2}$$



$$f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x} + b = 0$$