CS 4342: Class 12

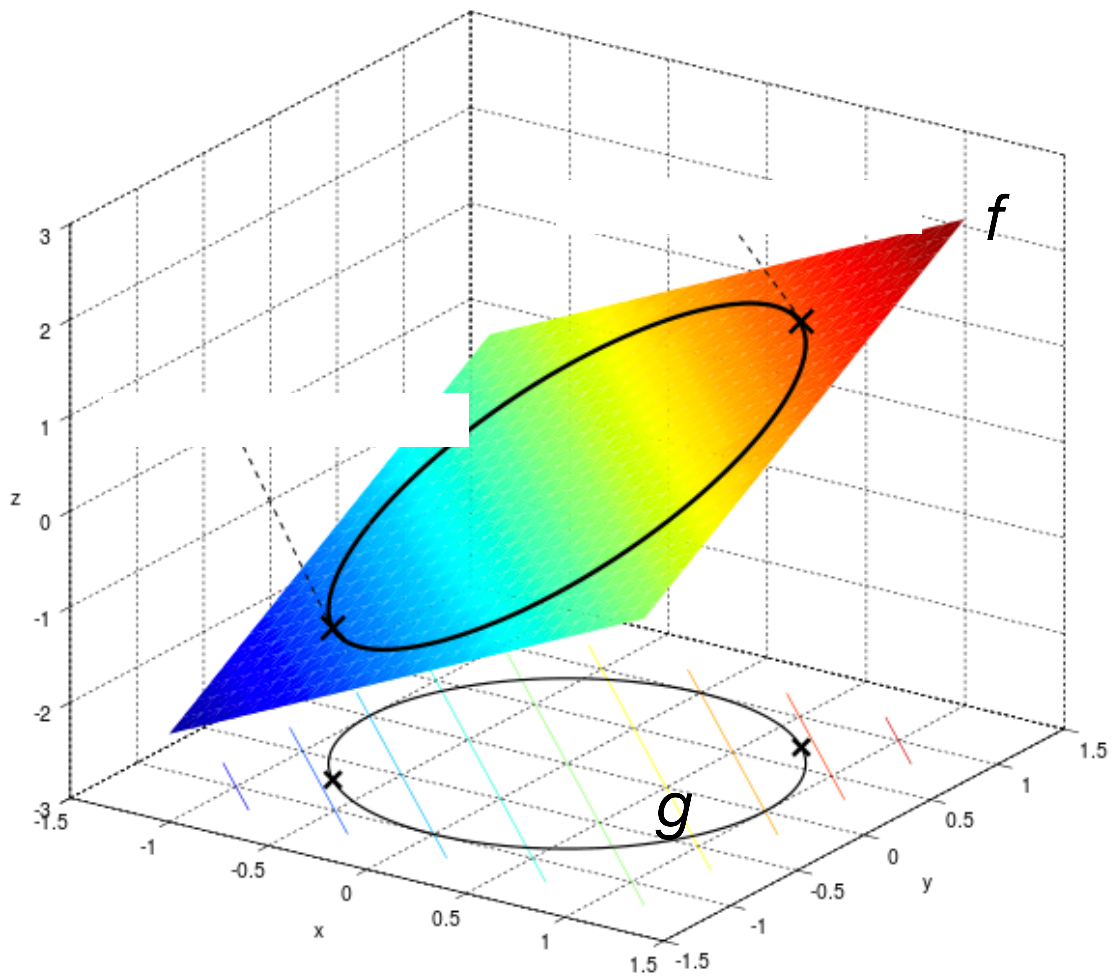
Jacob Whitehill

Lagrange multipliers

Lagrange multipliers

- Minimize:

$$f(x,y) = x + y \quad \text{subject to} \quad x^2 + y^2 = 1$$



Jacob Whitehill, WPI

Lagrange multipliers

- We can express the equality constraint ($x^2+y^2=1$) as a constraint function g.

- We define g so that $g(x,y) = 0$ when the constraint is satisfied:

$$g(x, y) = x^2 + y^2 - 1$$

Lagrange multipliers

- We can express the equality constraint ($x^2+y^2=1$) as a constraint function g.
- We define g so that $g(x,y) = 0$ when the constraint is satisfied:

$$g(x, y) = \tanh(x^2 + y^2 - 1)$$

Example

$$
\begin{aligned}
f(x,y) &= x+y \quad \text{subject to} \quad x^2+y^2=1 \\
L(x,y,\alpha) &= x+y+\alpha(x^2+y^2-1) \\
\frac{\partial L}{\partial x} &= 1+2\alpha x = 0 \\
\frac{\partial L}{\partial y} &= 1+2\alpha y = 0 \\
\frac{\partial L}{\partial \alpha} &= x^2+y^2-1=0 \\
2\alpha x &= -1 \\
x &= -1/(2\alpha) \\
y &= -1/(2\alpha) = x \\
x^2+(x)^2-1 &= 0 \\
2x^2 &= 1 \\
x^2 &= 1/2 \\
x &= y = \pm 1/\sqrt{2}
\end{aligned}
$$

Jacob Whitehill, WPI

Example

$$
\begin{aligned}
f(x,y) &= x+y \quad \text{subject to} \quad x^2+y^2=1 \\
L(x,y,\alpha) &= x+y+\alpha\tanh(x^2+y^2-1) \\
\frac{\partial L}{\partial x} &= 1+2\alpha(1-\tanh^2(x^2+y^2-1))x=0 \\
\frac{\partial L}{\partial y} &= 1+2\alpha(1-\tanh^2(x^2+y^2-1))y=0 \\
\frac{\partial L}{\partial \alpha} &= \tanh(x^2+y^2-1)=0 \\
\implies x &= y \\
x &= -1/(2\alpha) \\
y &= -1/(2\alpha)=x \\
\tanh(x^2+(x)^2-1) &= 0 \quad \implies \quad x^2+(x)^2-1=0 \\
2x^2 &= 1 \\
x^2 &= 1/2 = \pm 1/\sqrt{2} \\
y &= \pm 1/\sqrt{2}
\end{aligned}
$$

Lagrange multipliers

- Both constraint functions *g* yield the same solution.
- In this example, the constrained optimum can be deduced algebraically.
- However, with machine learning we typically need to solve constrained optimization problems numerically.
- In such cases, using simpler (e.g., linear) constraint functions is both faster and easier.

Support vector machines

Support vector machines

- **Support vector machines (SVMs)** are a ML model for binary classification.

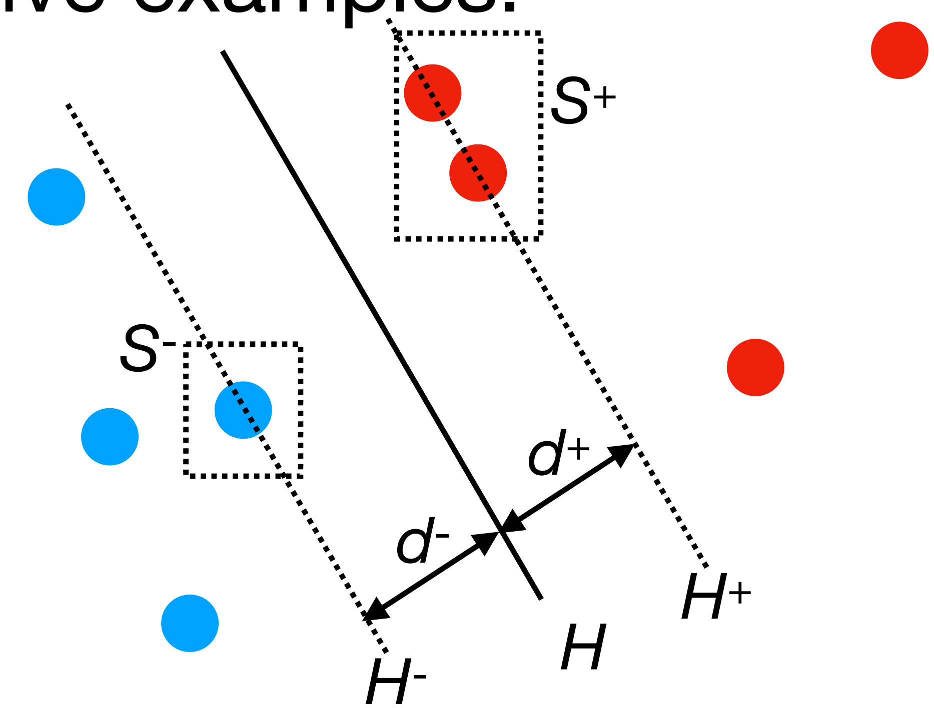
- SVMs are optimized using **constrained optimization** rather than unconstrained optimization (e.g., for logistic regression).

- For notational convenience, if example *i* belongs to the positive class, we write $y^{(i)} = +1$; if example *i* belongs to the negative class, we write $y^{(i)} = -1$.

Jacob Whitehill, WPI

Support vector machines

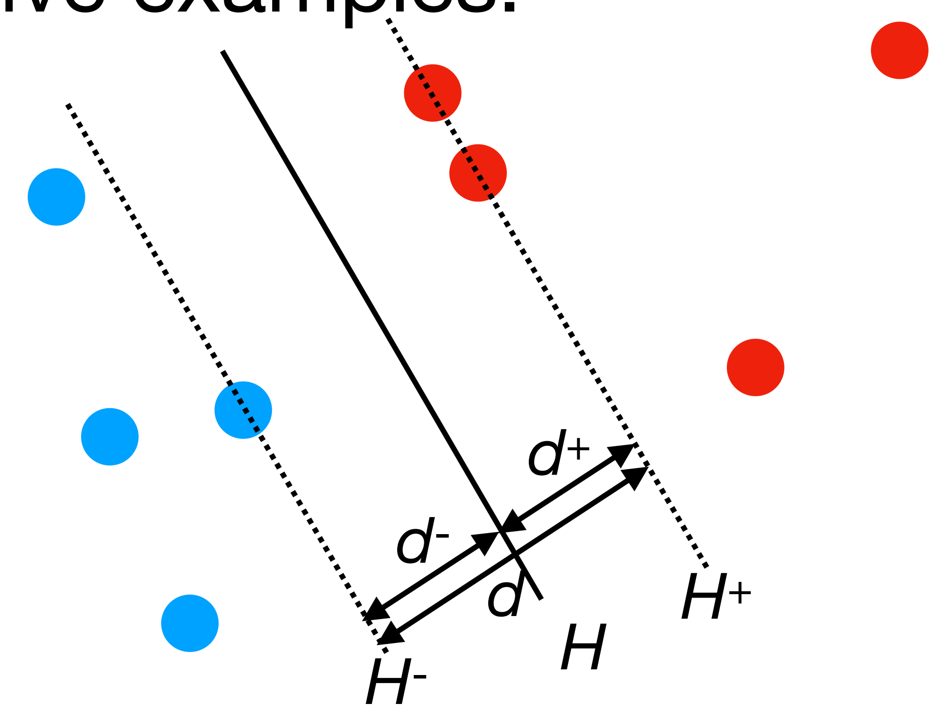
- For any hyperplane H that perfectly separates the positive from the negative examples:



- Find the subset S^+ of + examples that lie closest to H.
- The points in S^+ lie in a hyperplane H^+ parallel to H.
- Denote the shortest distance between H^+ and H as d^+.

Support vector machines

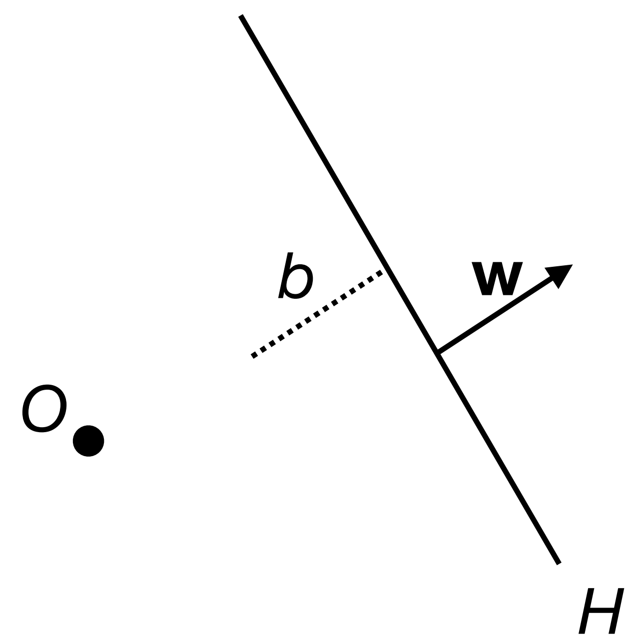
- For any hyperplane *H* that perfectly separates the positive from the negative examples:



- Let *d* denote the **margin** — the sum of d^+ and d^-.
- The optimization objective of SVMs is to find a separating hyperplane *H* that **maximizes** *d*.

Hyperplanes

Defining a hyperplane

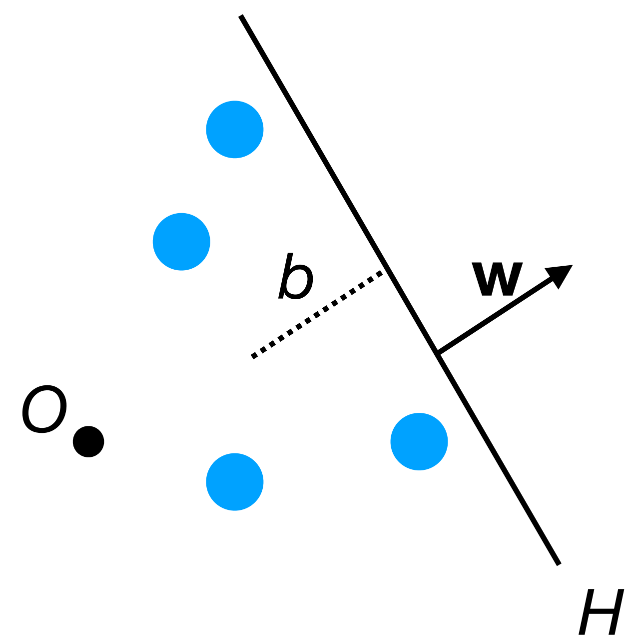


- A **hyperplane** is defined by a normal vector **w** ($\perp$ to *H*) and a bias *b* that is proportional to the distance to the origin.

- The points on hyperplane *H* are those values of **x** that satisfy:

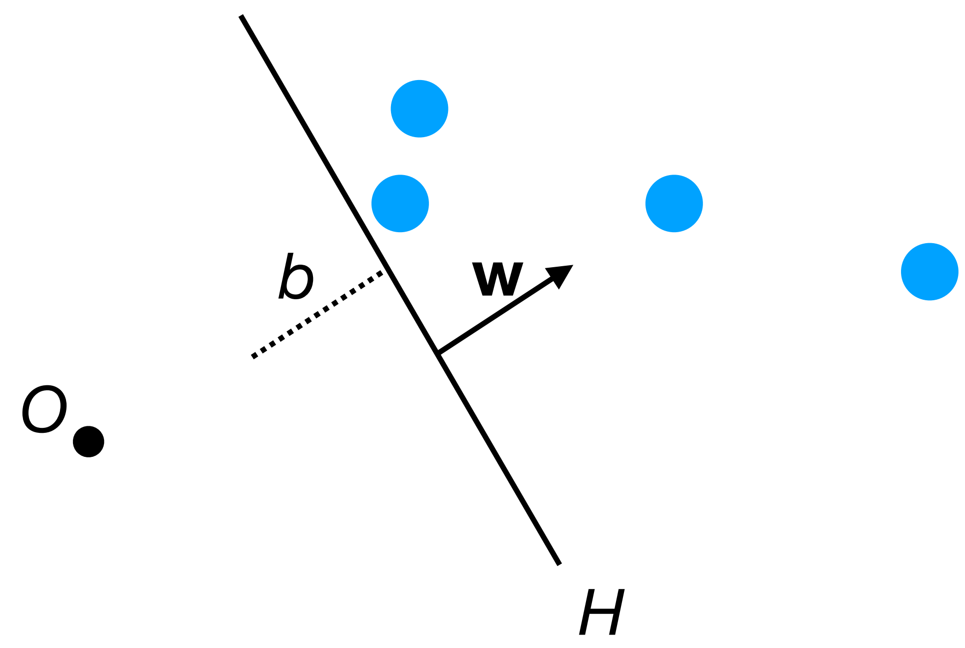
$$\mathbf{x}^\top \mathbf{w} + b = 0$$

Defining a hyperplane



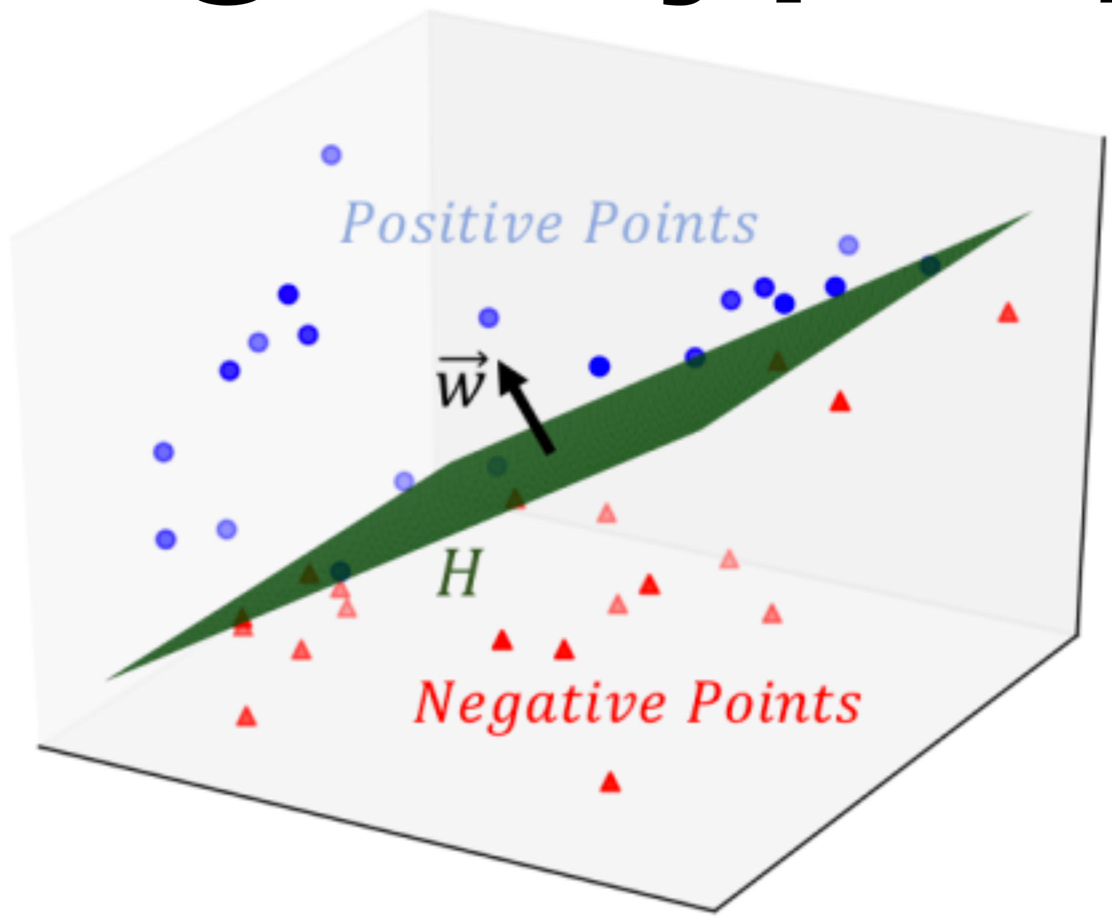
- The hyperplane separates points $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{w} + b > 0$ from points $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{w} + b < 0$.

Defining a hyperplane



- The hyperplane separates points $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{w} + b > 0$ from points $\mathbf{x}$ such that $\mathbf{x}^\top\mathbf{w} + b < 0$.

Defining a hyperplane

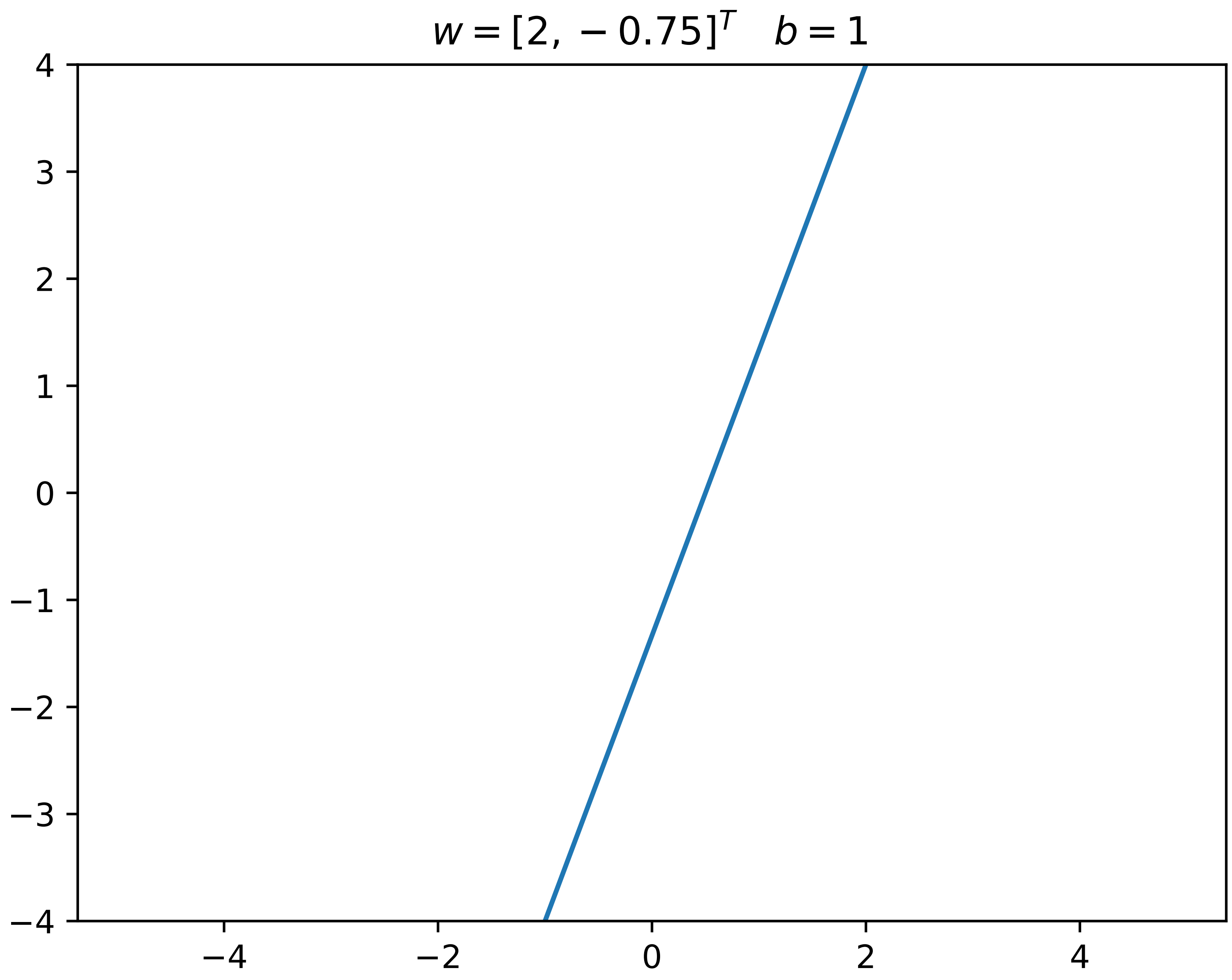


- A **hyperplane** is defined by a normal vector **w** ($\perp$ to *H*) and a bias *b* that is proportional to the distance to the origin.

- The points on hyperplane *H* are those values of **x** that satisfy:
$$\mathbf{x}^\top \mathbf{w} + b = 0$$

Hyperplane examples

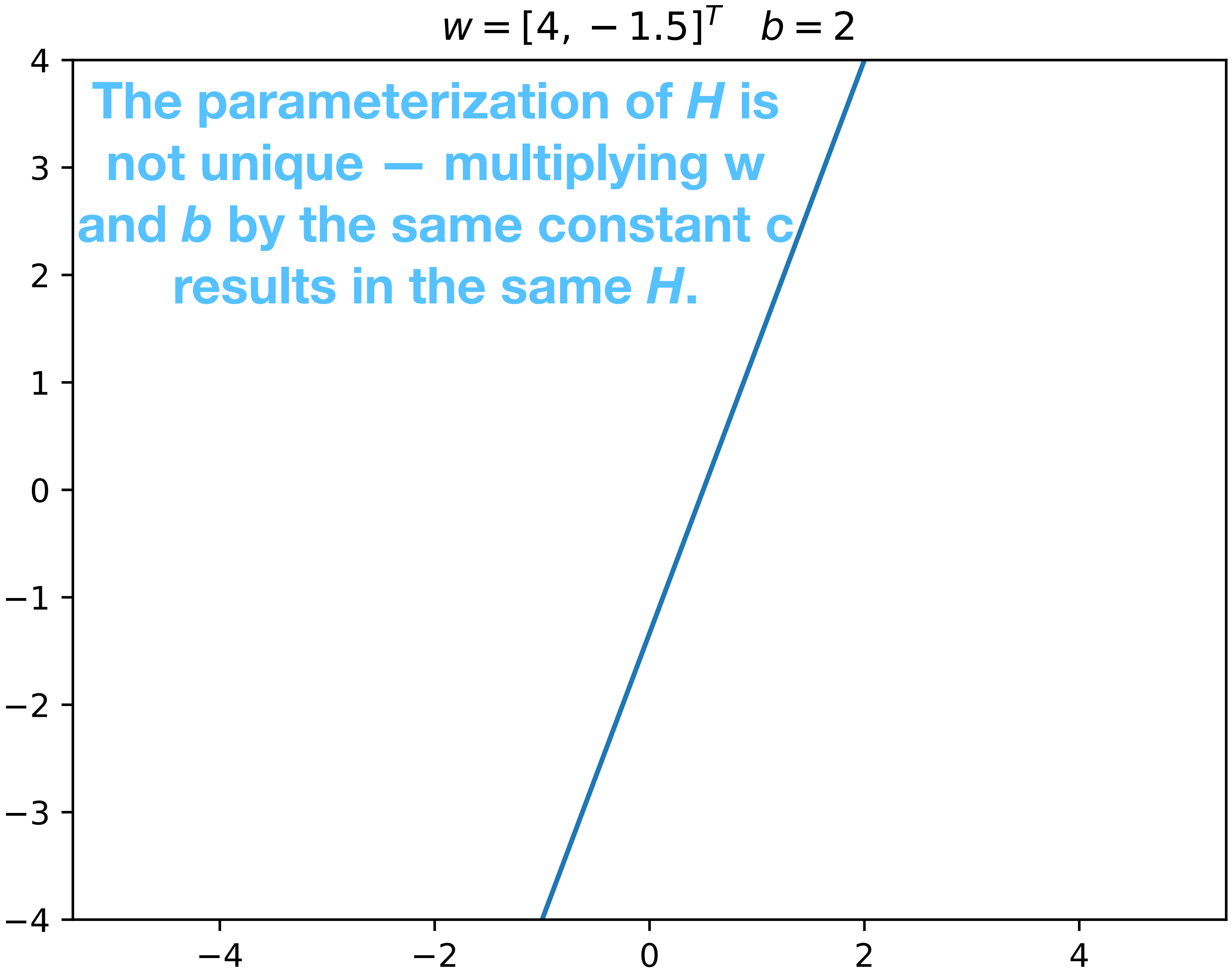
$$H = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{w} + b = 0\}$$



Jacob Whitehill, WPI

Hyperplane examples

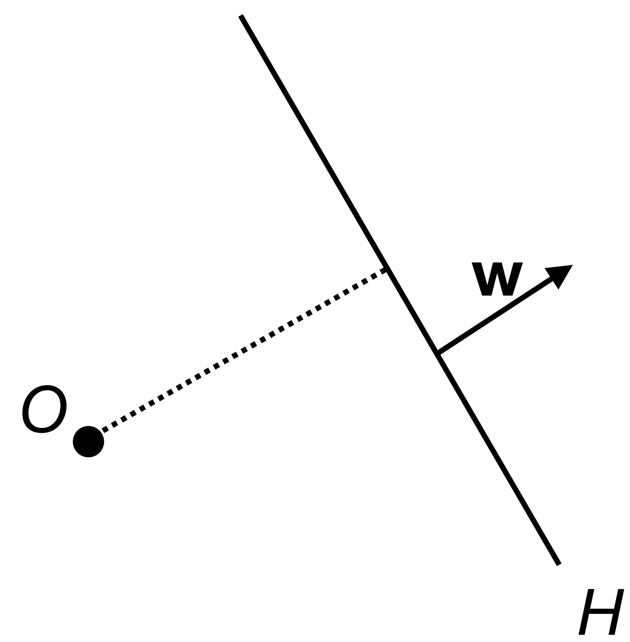
$$H = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{w} + b = 0\}$$



The parameterization of *H* is not unique — multiplying w and *b* by the same constant c results in the same *H*.

Jacob Whitehill, WPI

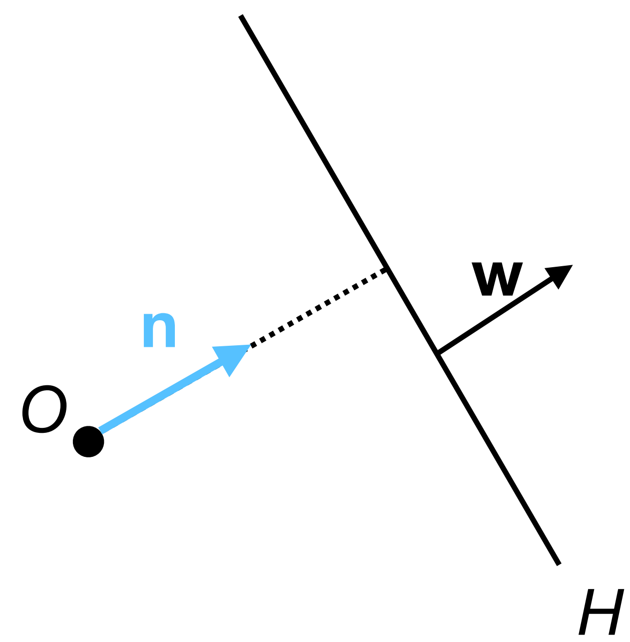
Distance from *O* to *H*



- To find the shortest (perpendicular) distance *c* between the origin *O* and the hyperplane *H*:

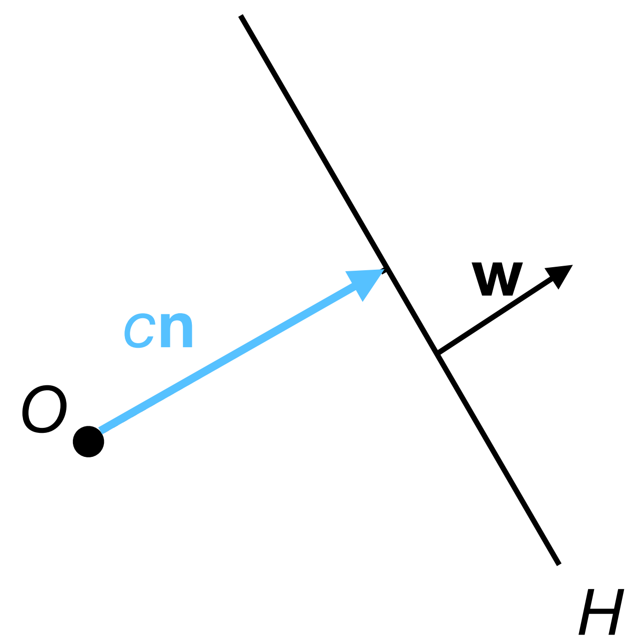
Jacob Whitehill, WPI

Distance from *O* to *H*



- To find the shortest (perpendicular) distance *c* between the origin *O* and the hyperplane *H*:
  - Define a *unit* vector **n** with same direction as **w**: $\mathbf{n} = \frac{\mathbf{w}}{|\mathbf{w}|}$

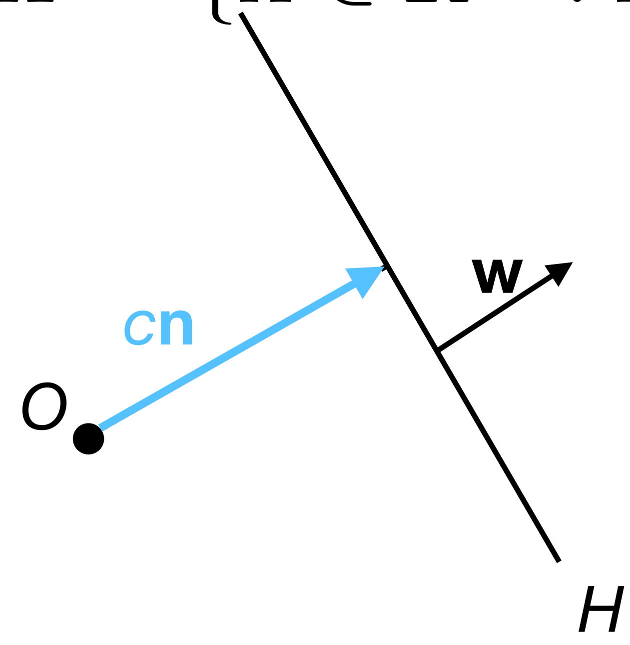
Distance from O to H



- To find the shortest (perpendicular) distance c between the origin O and the hyperplane H:
  - Define a *unit* vector **n** with same direction as **w**: $\mathbf{n} = \frac{\mathbf{w}}{|\mathbf{w}|}$
  - The shortest line from O to H ends at $c\mathbf{n}$.

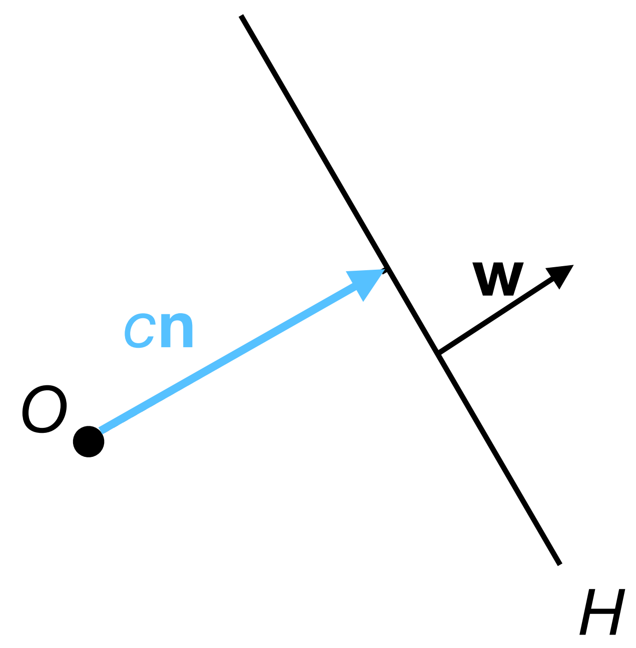
Distance from O to *H*

$$H = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{w} + z = 0\}$$



- Since $c\mathbf{n}$ is within H, we have: $$c\mathbf{n}^\top \mathbf{w} + z \quad = \quad 0$$

Distance from *O* to *H*

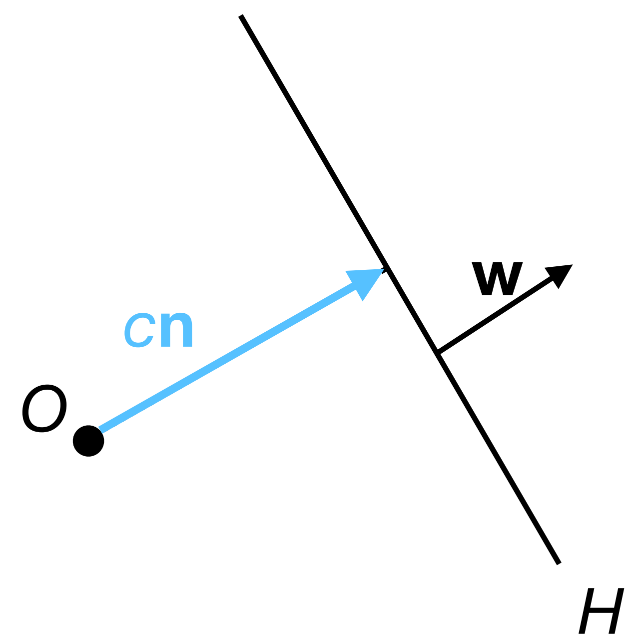


- Since *c***n** is within *H*, we have: $$c\mathbf{n}^\top \mathbf{w} + z \quad = \quad 0$$

- We can then solve for *c* (distance from *O* to *H*):

Jacob Whitehill, WPI

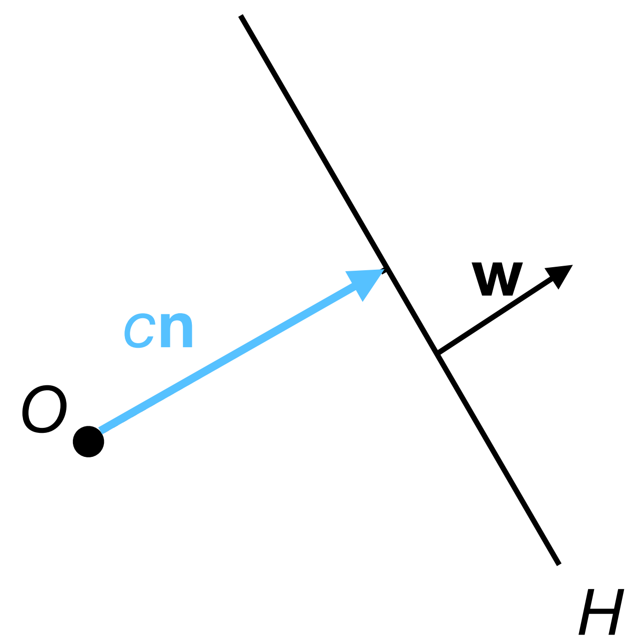
Distance from *O* to *H*



- Since *c***n** is within *H*, we have:
- We can then solve for *c* (distance from *O* to *H*):

$$\begin{aligned} c\mathbf{n}^\top \mathbf{w} + z &= 0 \\ c\left(\frac{\mathbf{w}}{|\mathbf{w}|}\right)^\top \mathbf{w} &= -z \end{aligned}$$

Distance from O to H

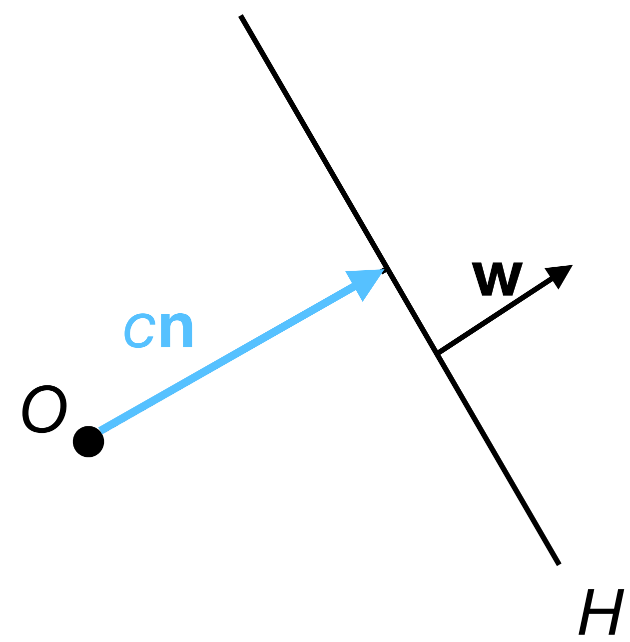


- Since $c\mathbf{n}$ is within H, we have:

- We can then solve for c (distance from O to H):

$$
\begin{aligned}
c\mathbf{n}^\top \mathbf{w} + z &= 0 \\
c\left(\frac{\mathbf{w}}{|\mathbf{w}|}\right)^\top \mathbf{w} &= -z \\
\frac{c}{|\mathbf{w}|}\mathbf{w}^\top \mathbf{w} &= -z
\end{aligned}
$$

Distance from *O* to *H*

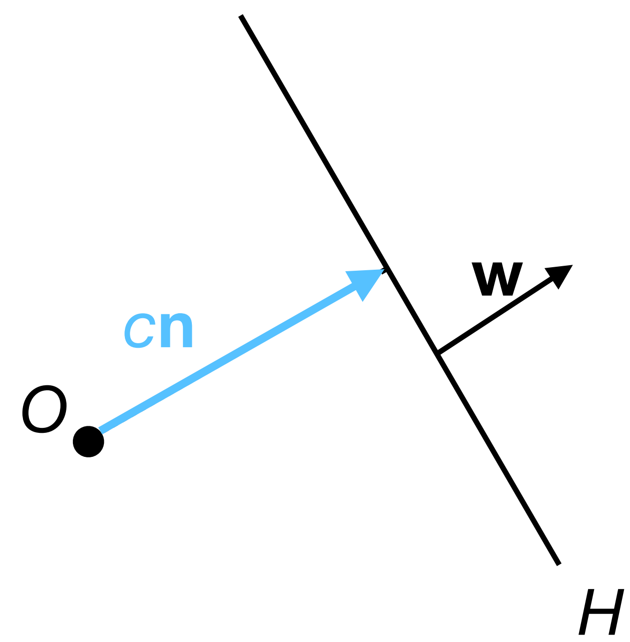


- Since *c***n** is within *H*, we have:

- We can then solve for *c* (distance from *O* to *H*):

$$
\begin{aligned}
c\mathbf{n}^\top \mathbf{w} + z &= 0 \\
c\left(\frac{\mathbf{w}}{|\mathbf{w}|}\right)^\top \mathbf{w} &= -z \\
\frac{c}{|\mathbf{w}|}\mathbf{w}^\top \mathbf{w} &= -z \\
\frac{c}{|\mathbf{w}|}|\mathbf{w}|^2 &= -z
\end{aligned}
$$

Distance from O to H

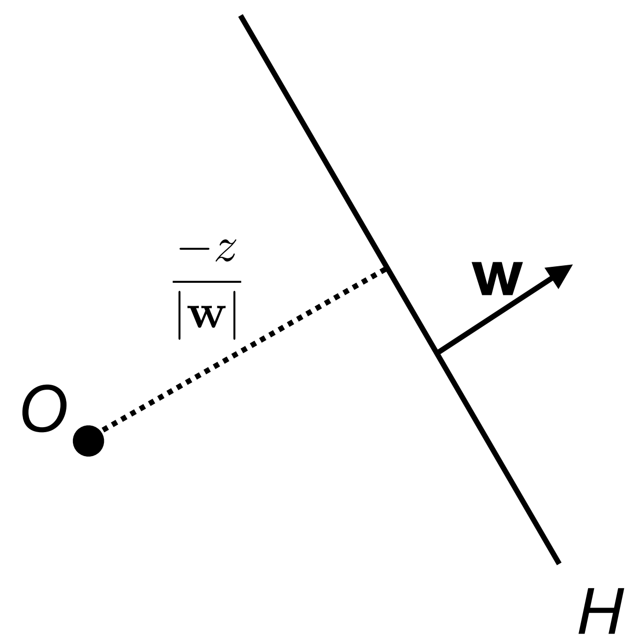


- Since $c\mathbf{n}$ is within H, we have:

- We can then solve for c (distance from O to H):

$$
\begin{aligned}
c\mathbf{n}^\top \mathbf{w} + z &= 0 \\
c\left(\frac{\mathbf{w}}{|\mathbf{w}|}\right)^\top \mathbf{w} &= -z \\
\frac{c}{|\mathbf{w}|}\mathbf{w}^\top \mathbf{w} &= -z \\
\frac{c}{|\mathbf{w}|}|\mathbf{w}|^2 &= -z \\
c|\mathbf{w}| &= -z \\
c &= \frac{-z}{|\mathbf{w}|}
\end{aligned}
$$

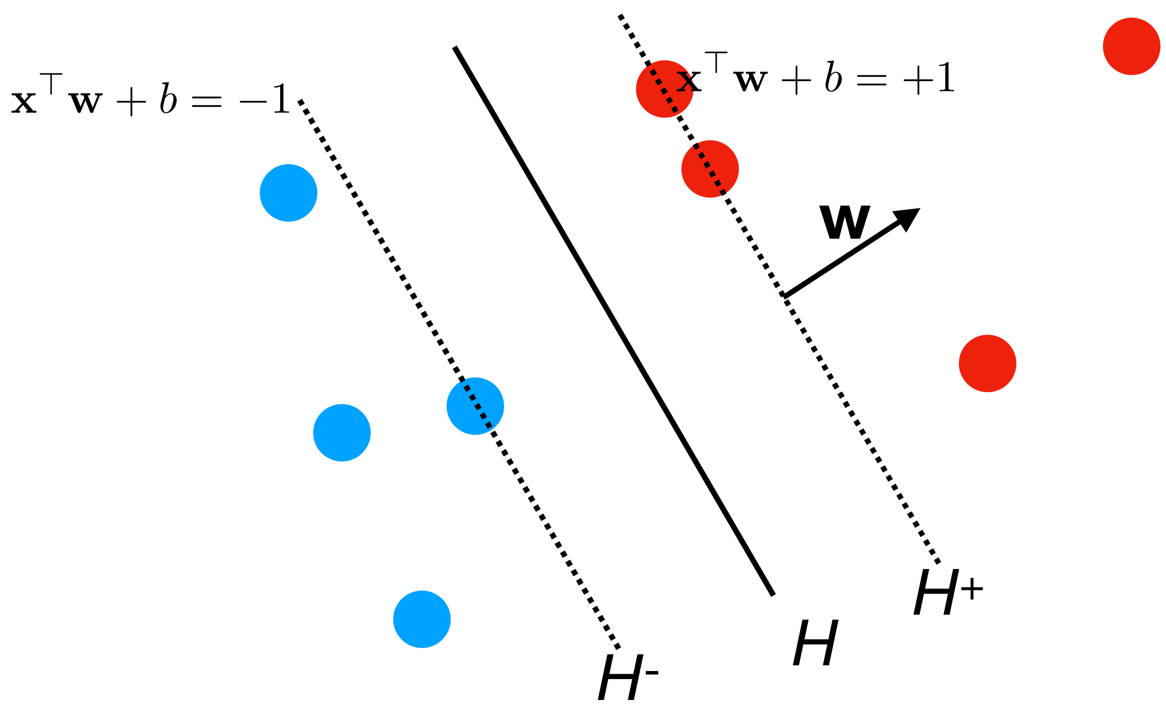
Distance from *O* to *H*



- Therefore, the shortest distance between the origin *O* and the hyperplane *H* is: $\frac{-z}{|\mathbf{w}|}$

Support vector machines

- Recall that $H \parallel H^+ \parallel H^-$. Then they can share the same **w**.

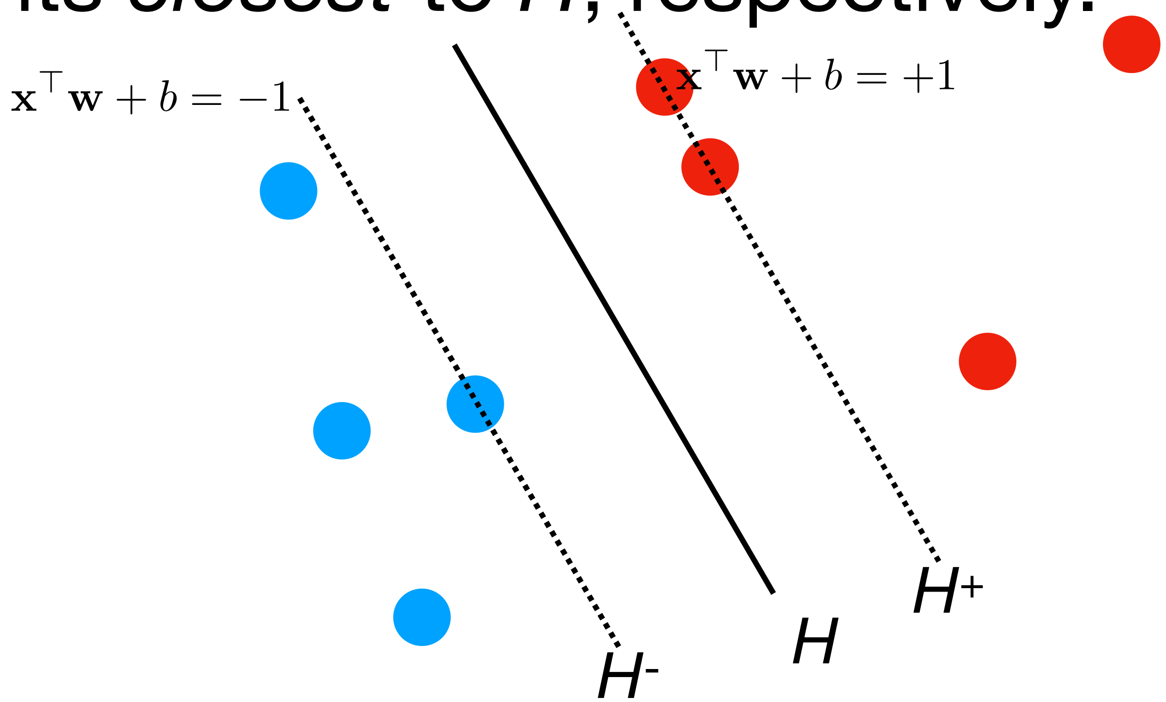


- We can scale **w** and *b* such that:

$$\begin{aligned} H^- &: \quad \mathbf{x}^\top \mathbf{w} + b = -1 \\ H &: \quad \mathbf{x}^\top \mathbf{w} + b = 0 \\ H^+ &: \quad \mathbf{x}^\top \mathbf{w} + b = +1 \end{aligned}$$

Support vector machines

- H^- and H^+ intersect the negatively and positively labeled data points *closest* to H, respectively.



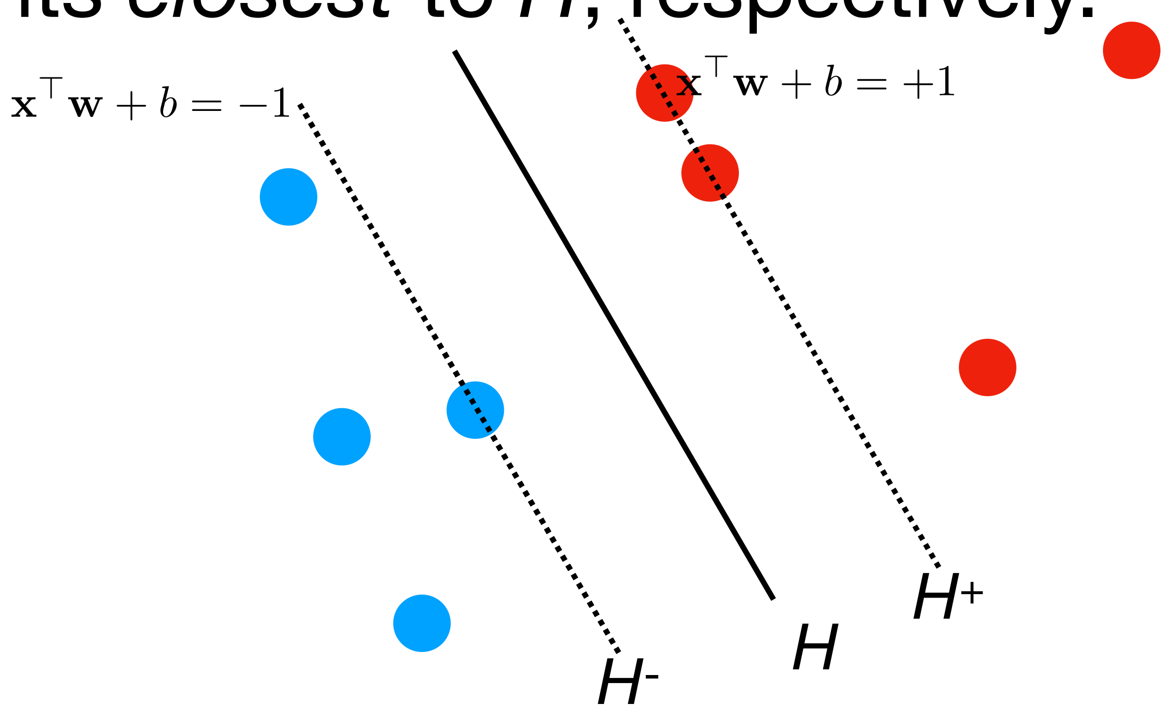
- Since all data points not in H^+ or H^- must lie even farther from H, we require that:

$$y^{(i)} = +1 \quad \Longrightarrow \quad {\mathbf{x}^{(i)}}^\top \mathbf{w} + b \geq +1$$
$$y^{(i)} = -1 \quad \Longrightarrow \quad {\mathbf{x}^{(i)}}^\top \mathbf{w} + b \leq -1$$

Jacob Whitehill, WPI

Support vector machines

- H^- and H^+ intersect the negatively and positively labeled data points *closest* to H, respectively.



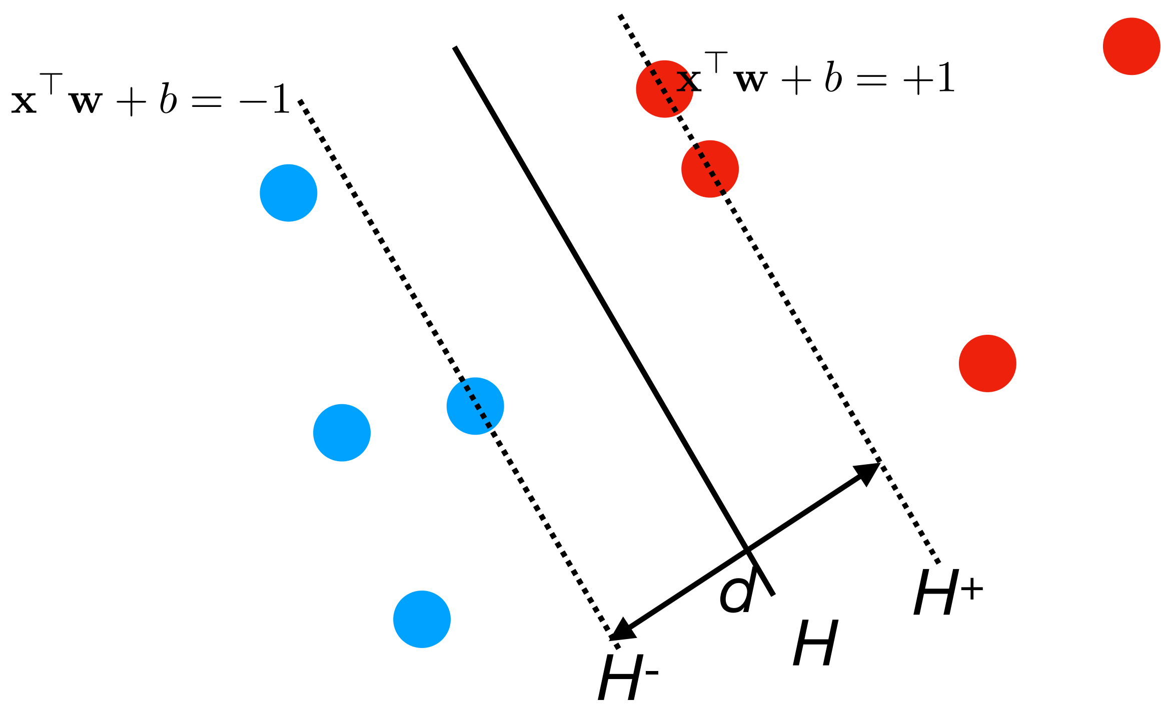
- These two sets of **constraints** can be unified:

$$y^{(i)}(\mathbf{x}^{(i)\top}\mathbf{w} + b) \geq 1 \quad \forall i$$

Inequality constraints

Maximizing the margin

- How do we maximize the margin *d*?

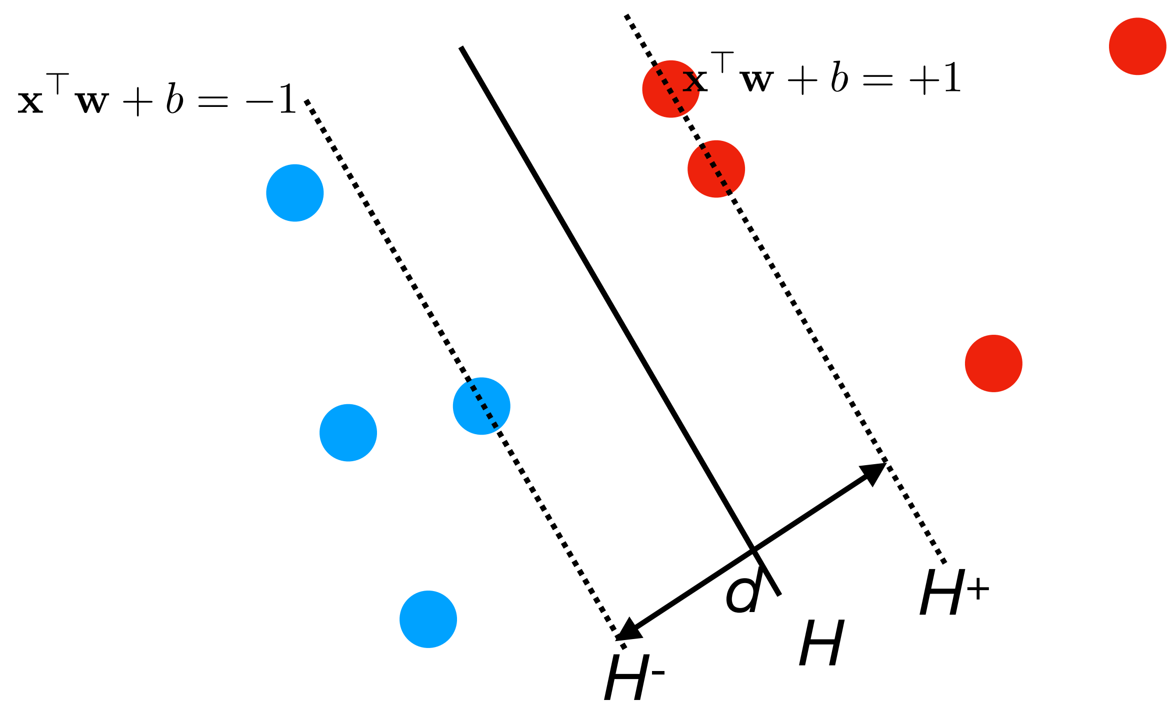


- Distance from origin for a hyperplane *H* ($\mathbf{x}^\top\mathbf{w}+z=0$):

$$c = \frac{-z}{|\mathbf{w}|}$$

Maximizing the margin

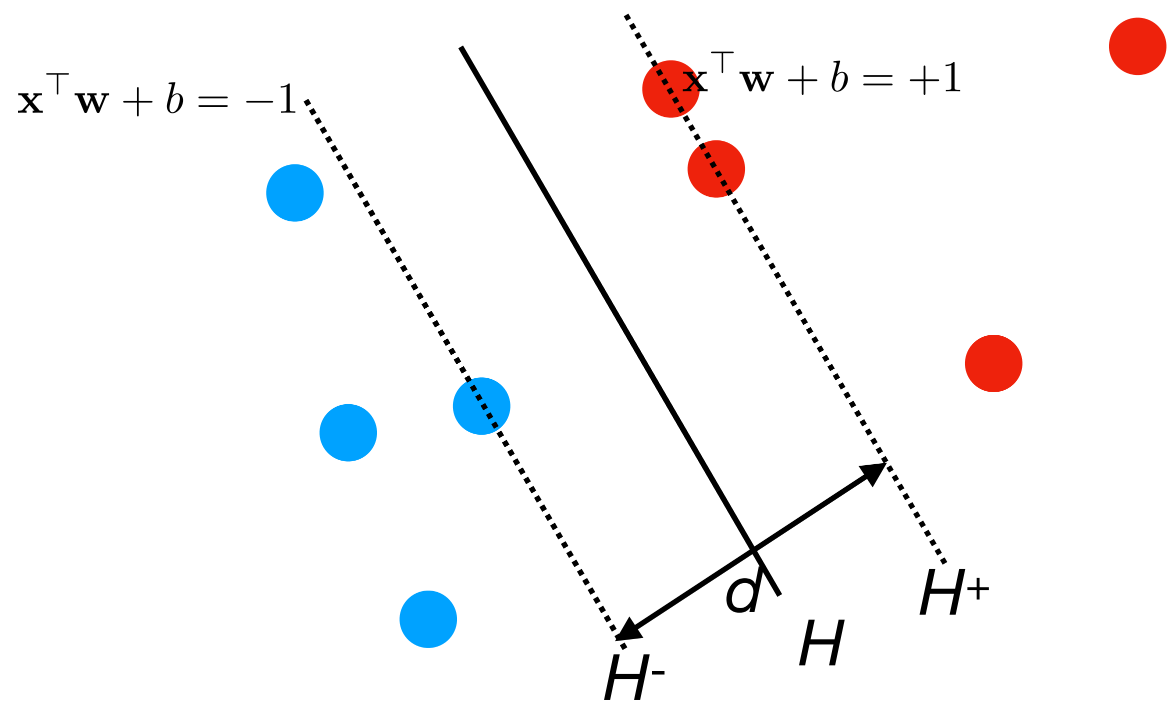
- How do we maximize the margin *d*?



- How far is H^- from H+?

Maximizing the margin

- How do we maximize the margin *d*?

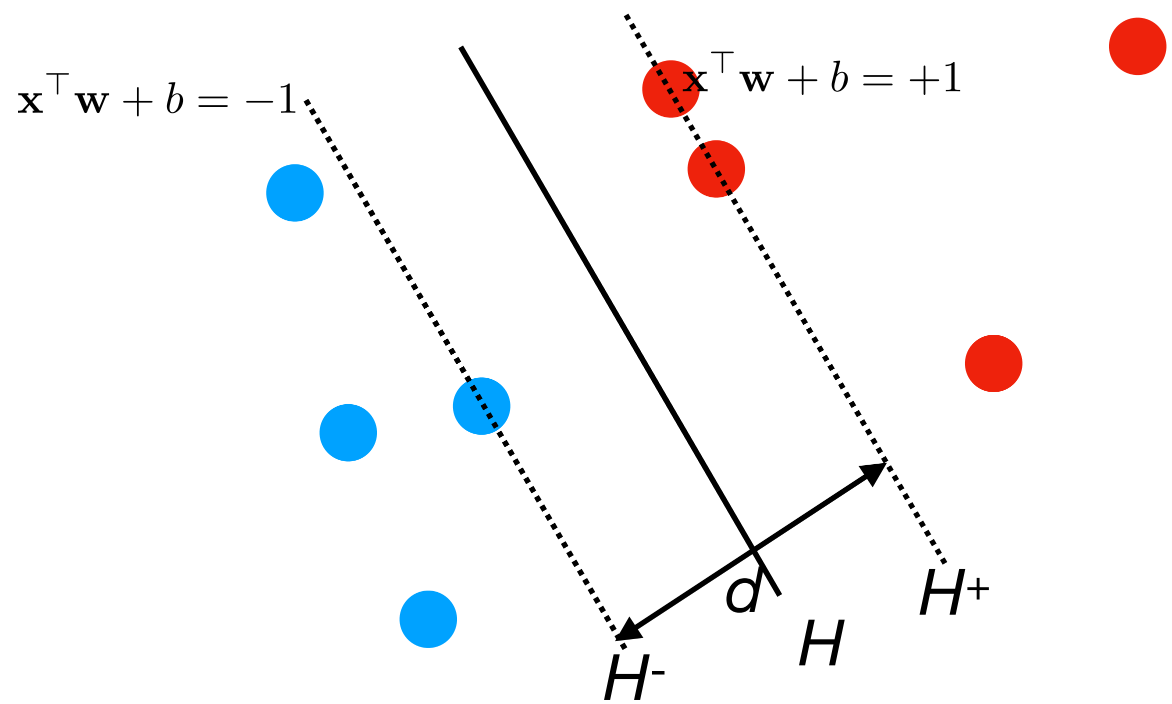


- *H*- is (-1-*b*)/|**w**| from the origin.

$$\frac{-1-b}{|\mathbf{w}|}$$

Maximizing the margin

- How do we maximize the margin *d*?



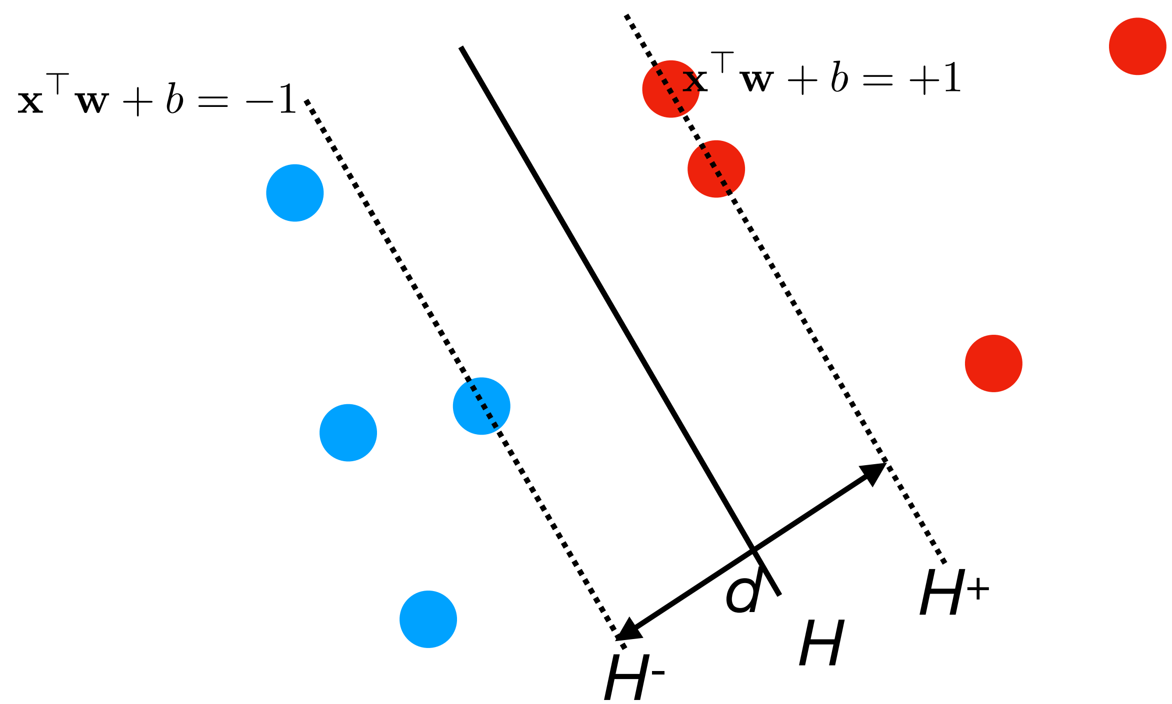
- H^+ is (1-*b*)/|**w**| from the origin.

$$\frac{1-b}{|\mathbf{w}|}$$

Jacob Whitehill, WPI

Maximizing the margin

- How do we maximize the margin *d*?

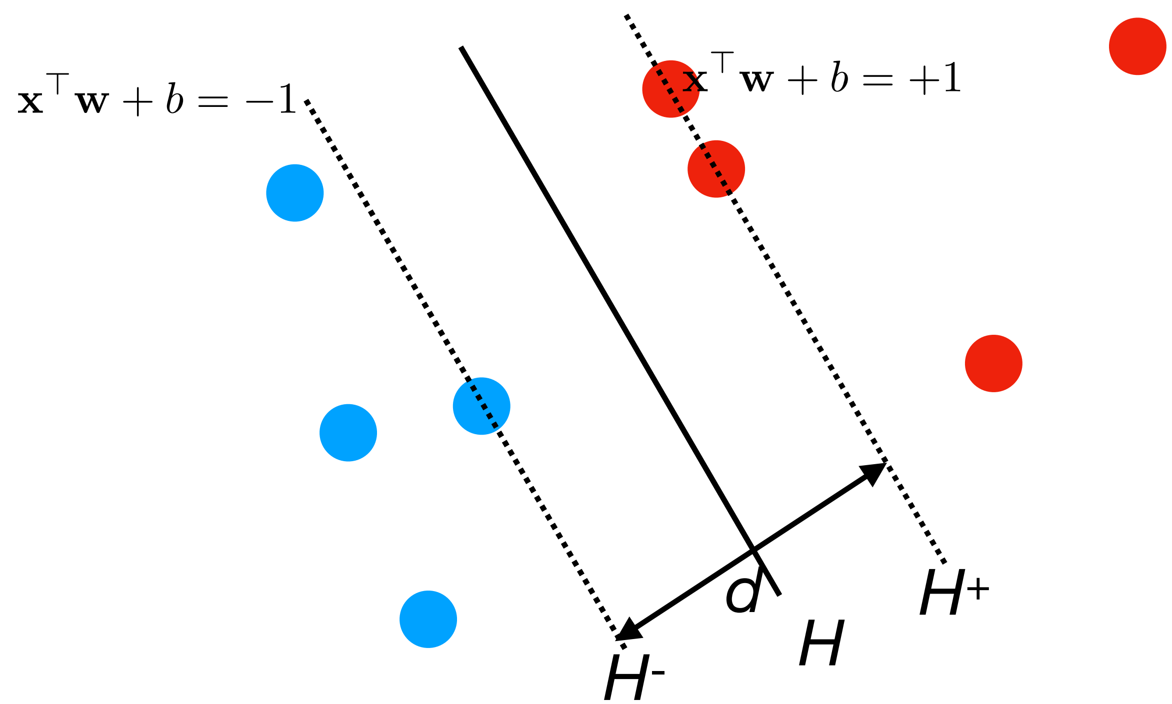


- Therefore, the **margin** (distance between the hyperplanes) must be:

$$d = \frac{1-b}{|\mathbf{w}|} - \frac{-1-b}{|\mathbf{w}|} = \frac{2}{|\mathbf{w}|}$$

Maximizing the margin

- How do we maximize the margin *d*?



- To *maximize d*=2/|**w**|, we can thus *minimize* |**w**|/2 or (equivalently) minimize:

$$\frac{1}{2}\mathbf{w}^\top \mathbf{w}$$

Optimization objective (cost function)

SVM optimization problem

- Putting the parts together, we wish to:
  - Minimize: $\frac{1}{2}\mathbf{w}^\top\mathbf{w}$
  - Subject to: $y^{(i)}(\mathbf{x}^{(i)^\top}\mathbf{w}+b) \geq 1 \quad \forall i$

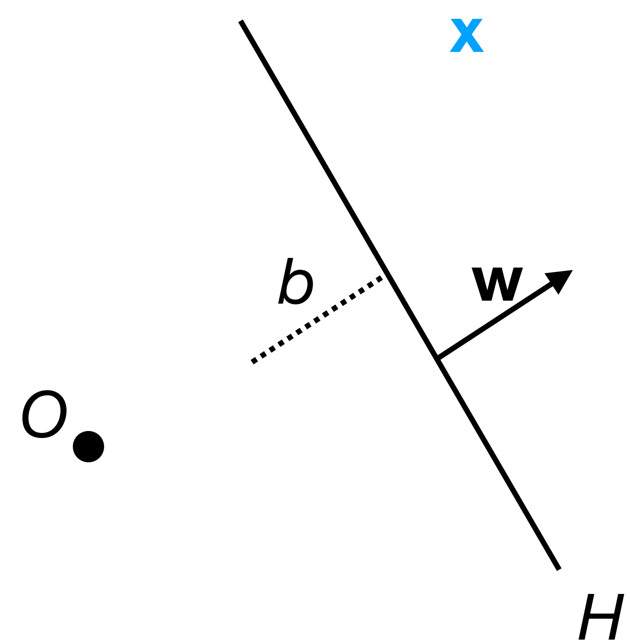
SVM optimization problem

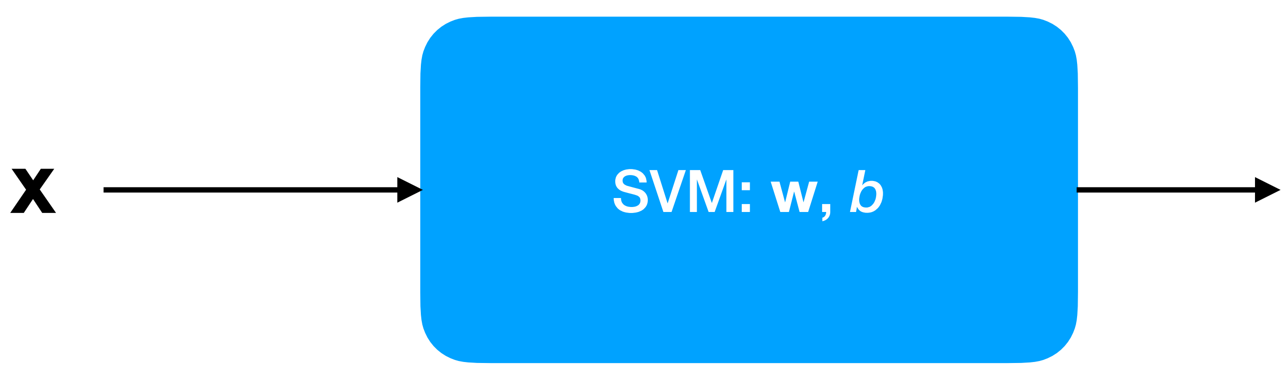
- Putting the parts together, we wish to:
  - Minimize: $\frac{1}{2}\mathbf{w}^\top\mathbf{w}$
  - Subject to: $y^{(i)}(\mathbf{x}^{(i)}{}^\top\mathbf{w} + b) \geq 1 \quad \forall i$
- This is a **quadratic programming** problem: quadratic objective with linear inequality (and/or equality) constraints. There are many efficient solvers for quadratic programs.
- The optimization variables are both **w** and *b*.

SVM: classification

SVM: classification

- Here’s how an SVM classifies a new example:





$$\hat{y} = \begin{cases} 1 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b > 0 \\ \text{-}1 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b < 0 \end{cases}$$

Can decide class arbitrarily if $\mathbf{x}^\top\mathbf{w} + b = 0$.

Exercise

- Suppose **w** = [1, 3, -2]^T and *b* = -2.
- What is the class (+ or -) of the following **x**?
  - **x** = [-2, 4, 2]^T
  - **x** = [1, 3, -2]^T
  - **x** = [6, 0.5, 5]^T

$$\hat{y} = \begin{cases} 1 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b > 0 \\ 0 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b < 0 \end{cases}$$

Exercise

- Suppose $\mathbf{w}$ = [1, 3, -2]$^\mathsf{T}$ and b = -2.
- What is the class (+ or -) of the following $\mathbf{x}$?
  - $\mathbf{x}$ = [-2, 4, 2]$^\mathsf{T}$ => $\mathbf{x}^\mathsf{T}\mathbf{w}+b$ = -2+12-4-2=4 => +
  - $\mathbf{x}$ = [1, 3, -2]$^\mathsf{T}$ => $\mathbf{x}^\mathsf{T}\mathbf{w}+b$ = 1+9+4-2=12 => +
  - $\mathbf{x}$ = [6, 0.5, 5]$^\mathsf{T}$ => $\mathbf{x}^\mathsf{T}\mathbf{w}+b$ = 6+1.5-10-2=-4.5 => -

$$\hat{y} = \begin{cases} 1 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b > 0 \\ 0 & \text{if} \quad \mathbf{x}^\top \mathbf{w} + b < 0 \end{cases}$$

Jacob Whitehill, WPI