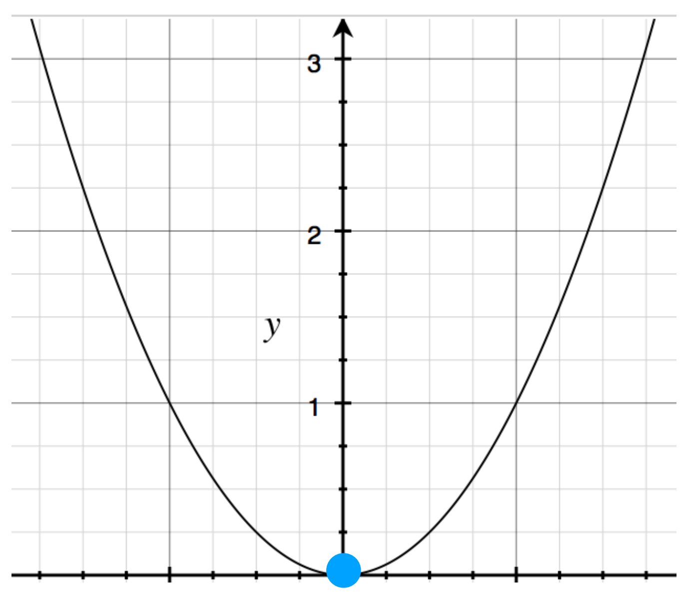
CS 4342: Class 11

Jacob Whitehill

Constrained optimization

Unconstrained optimization

- So far, the ML methods we have examined are based on optimizing some **objective function** (loss or accuracy).
- The optimization variable has been **unconstrained** — it can be any value in $\mathbb{R}^m$.
- Unconstrained optimal solutions exist at critical points of the objective function *f*, i.e., where the gradient of *f* is 0, e.g.:



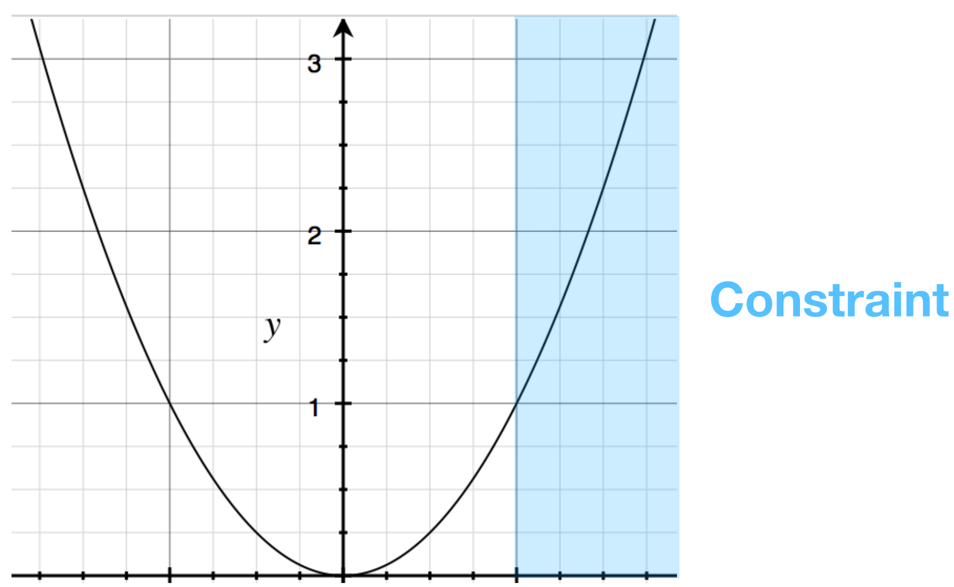
  - The minimum of this function is at x=0.

Jacob Whitehill, WPI

Constrained optimization

- Things become more complicated when we put a constraint on the optimization variables.

- What if we want to minimize *f* subject to the **inequality constraint** that $x \geq 1$?

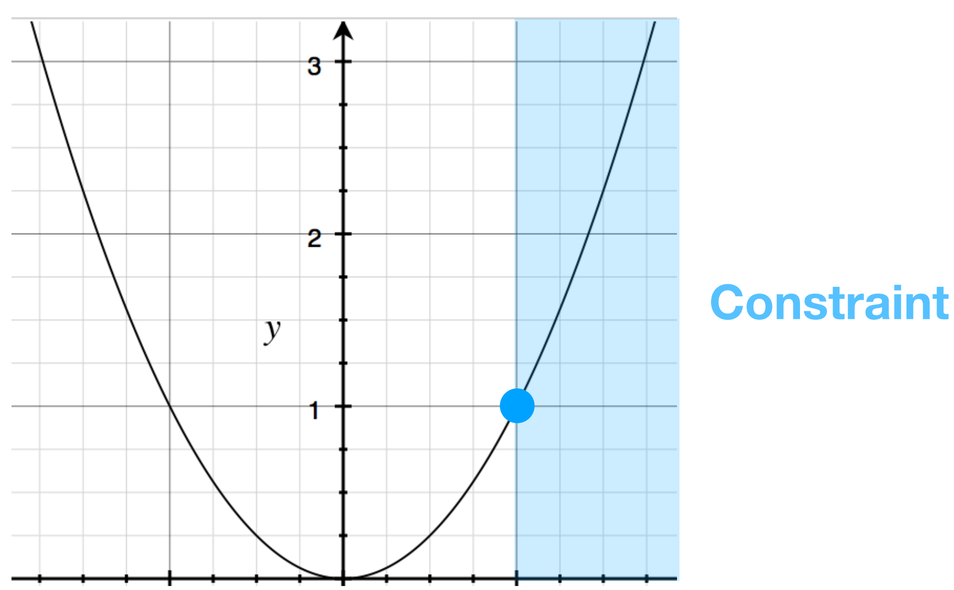


Constrained optimization

- Things become more complicated when we put a constraint on the optimization variables.

- What if we want to minimize *f* subject to the **inequality constraint** that $x \geq 1$?

- The solution no longer occurs at a critical point of *f*.



- The minimum of *f*, constrained s.t. $x \geq 1$, is at $x=1$.

Constrained optimization methods

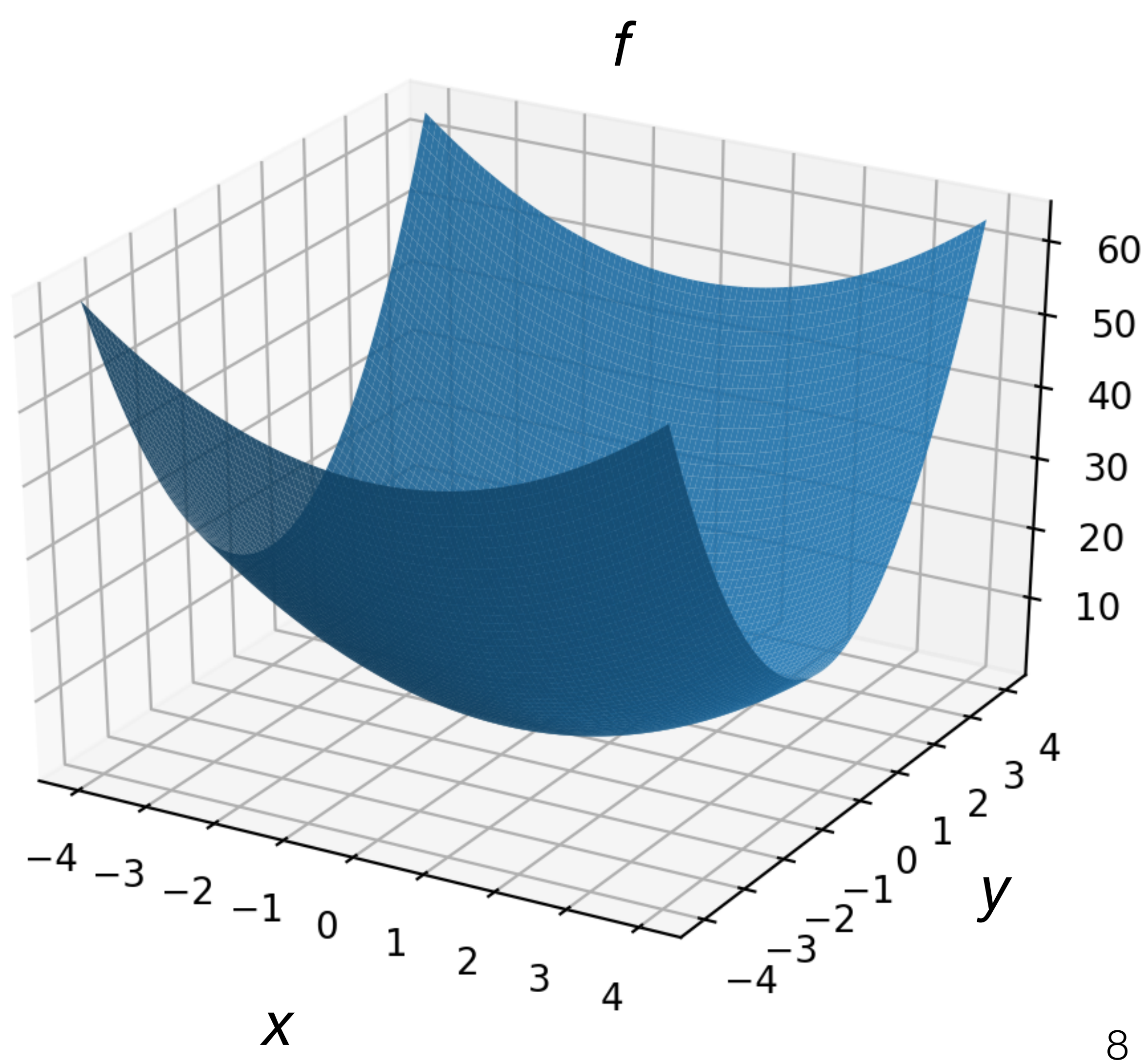
- A variety of techniques exist for solving constrained optimization problems.
- Many of these are applicable when the objective function *f* is convex.
- Two widely used techniques:
    - Lagrange multipliers
    - Karush-Kuhn-Tucker (KKT) optimality conditions

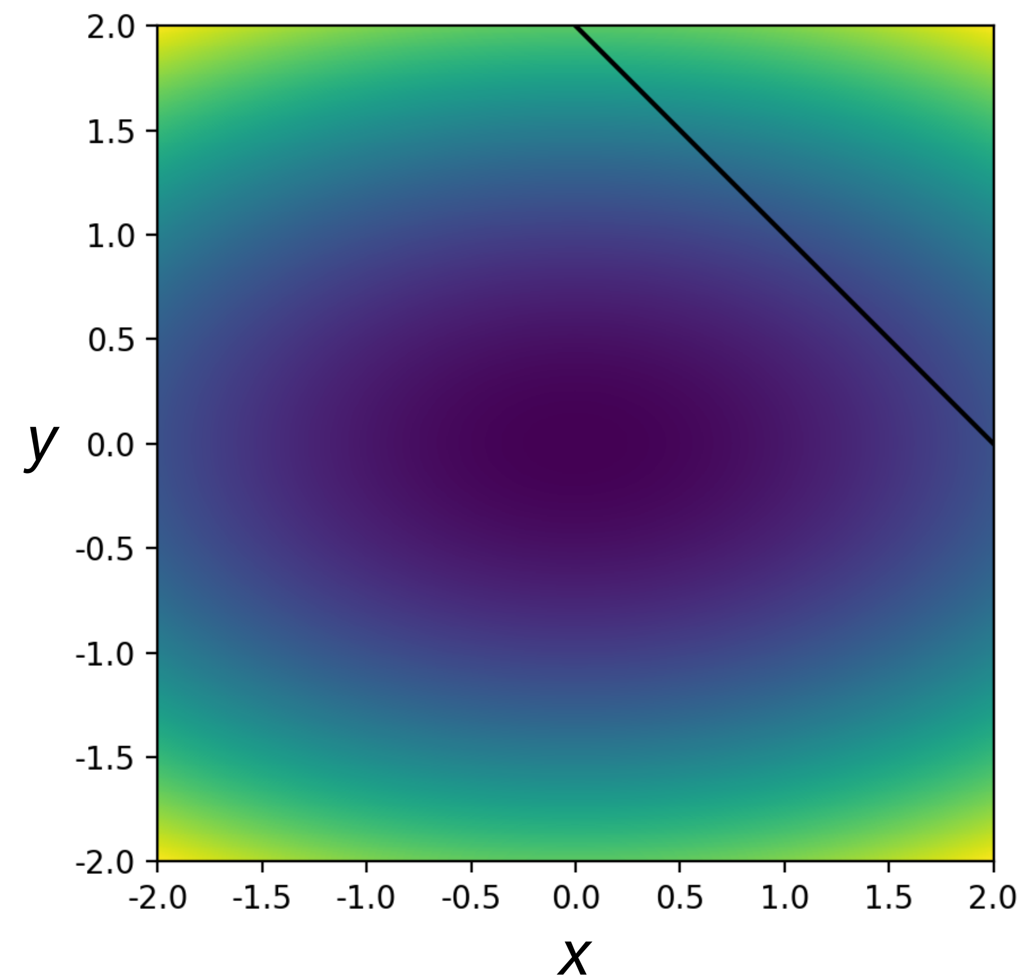
Lagrange multipliers

Lagrange multipliers

- Lagrange multipliers are useful for solving optimization problems involving **equality constraints**, e.g., minimize:

$$f(x,y) \quad = \quad x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2$$





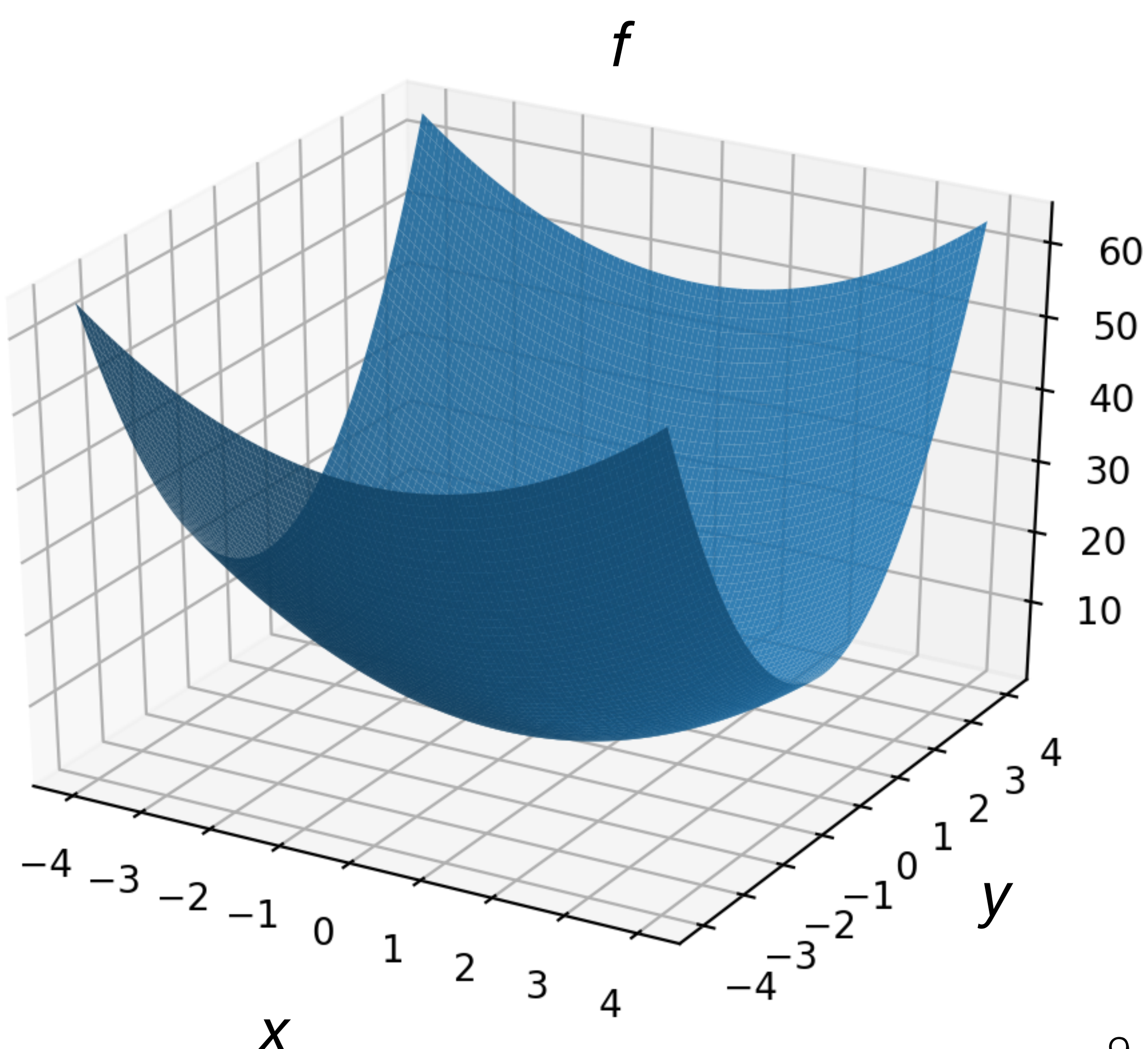
Lagrange multipliers

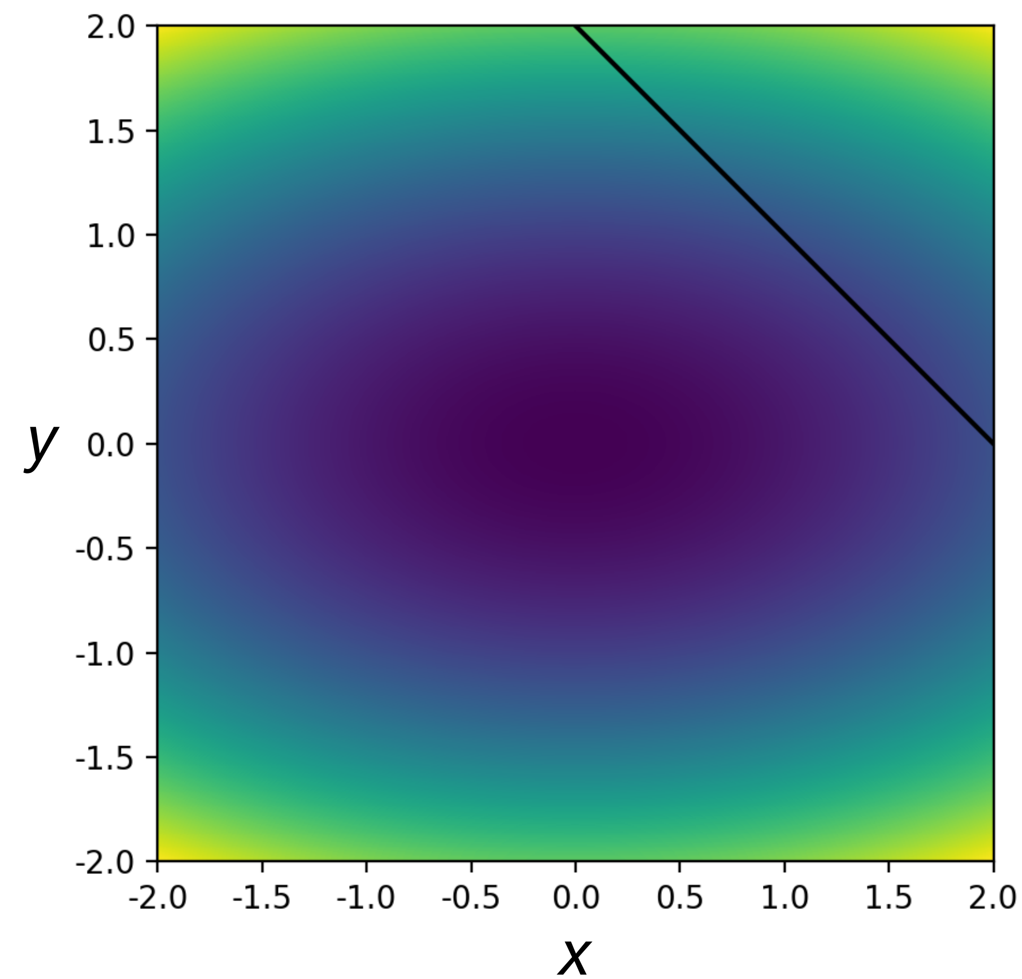
- Lagrange multipliers are useful for solving optimization problems involving **equality constraints**, e.g., minimize:

$$f(x,y) \quad = \quad x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2$$

Objective function

Equality constraint





Lagrange multipliers

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

$$g(x, y) = \boxed{?}$$

Lagrange multipliers

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

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

Lagrange multipliers

- To solve the constrained optimization problem, we define the Lagrangian function *L* in terms of:
  - The original optimization variables.
  - The Lagrange multiplier(s) *α* (one for each constraint).
- For one constraint *g*, we have:

$$L(x, y, \alpha) = f(x, y) + \alpha g(x, y)$$

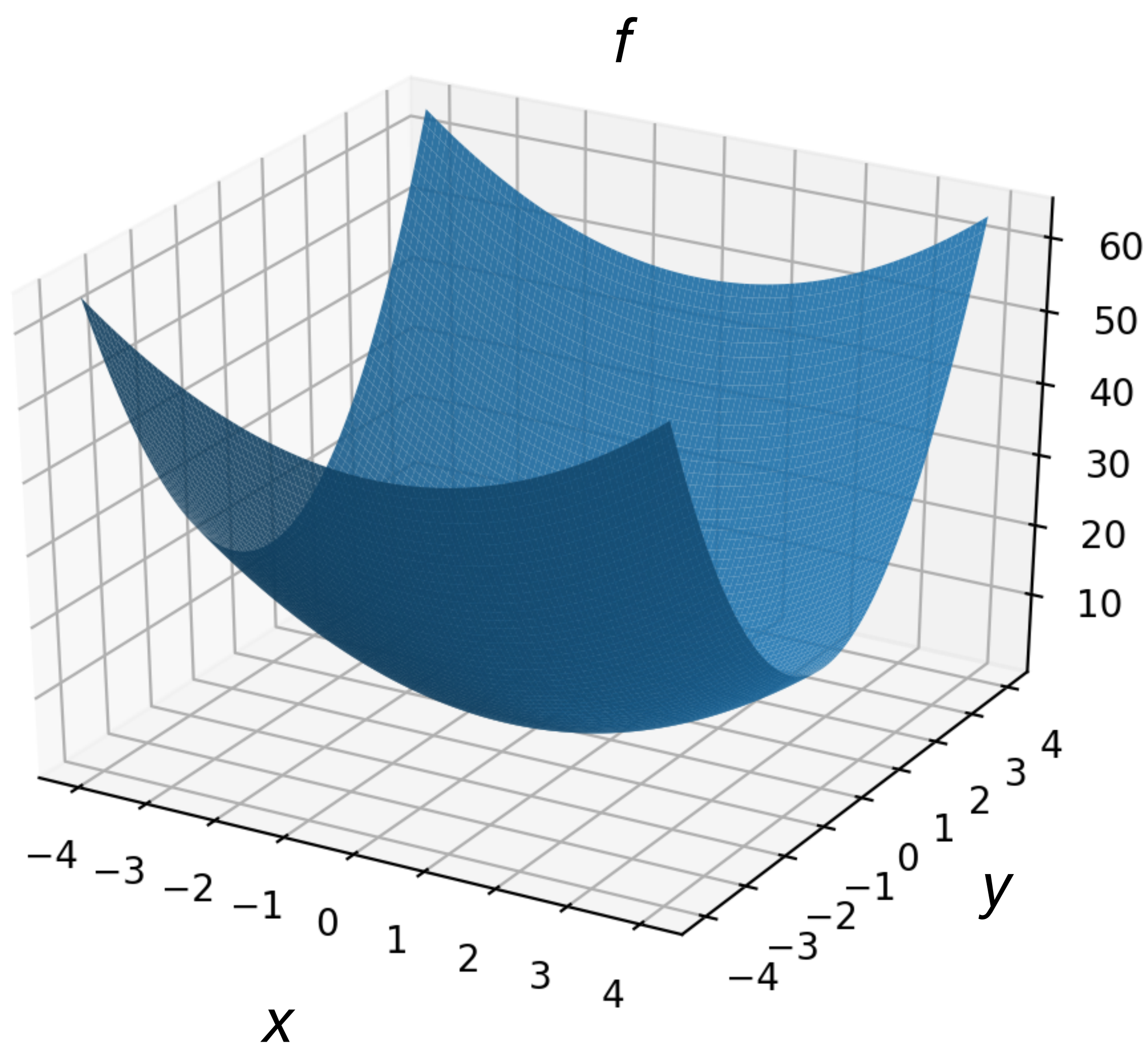
Lagrange multipliers

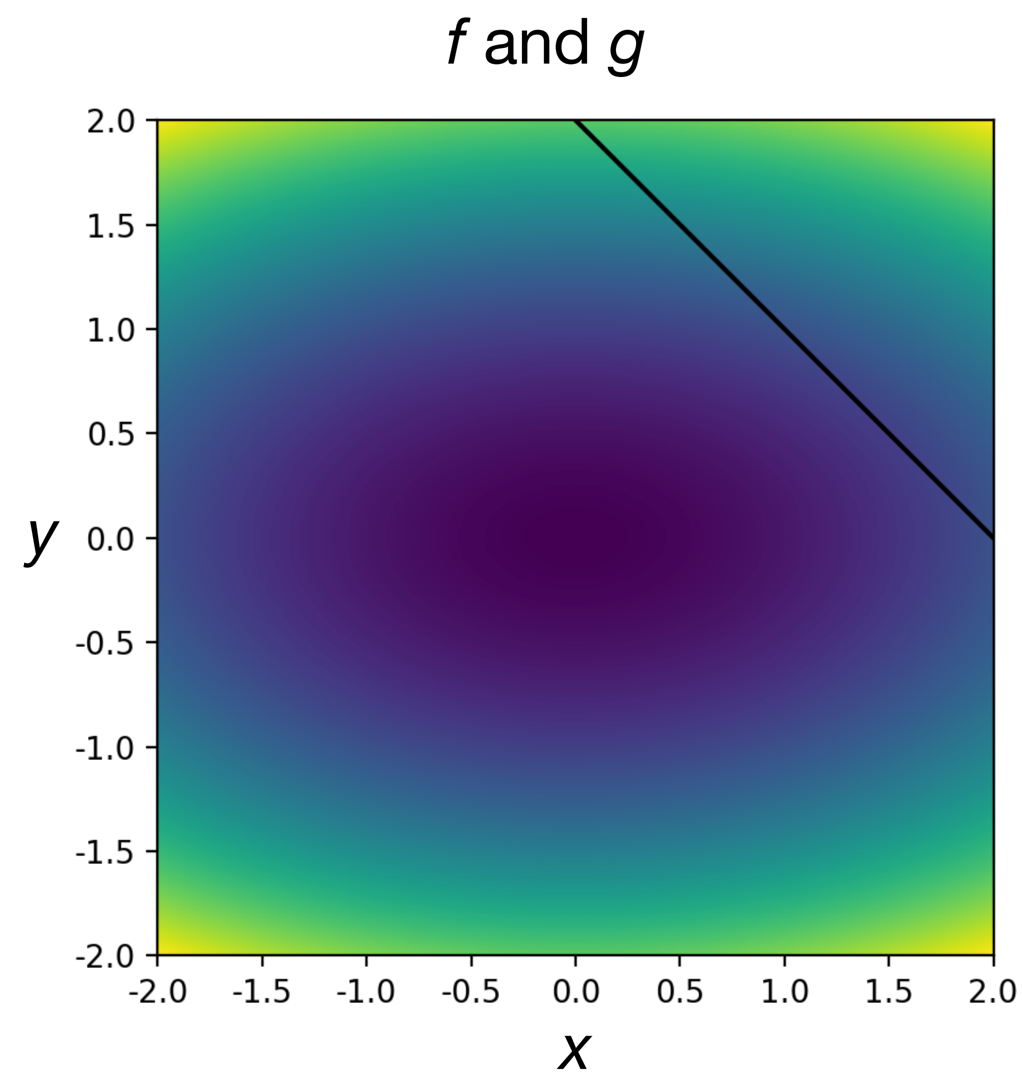
- The solution occurs at a critical point of *L*, i.e., where the derivative of *L* with respect to *x*, *y*, and *α* = 0.

$$
\begin{aligned}
L(x, y, \alpha) &= f(x, y) + \alpha g(x, y) \\
\frac{\partial L}{\partial x} &= 0 \\
\frac{\partial L}{\partial y} &= 0 \\
\frac{\partial L}{\partial \alpha} &= 0
\end{aligned}
$$

Example

$$f(x, y) \quad = \quad x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2$$





Example

$$\begin{aligned} f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\ L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \end{aligned}$$

Example

$$\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0
\end{aligned}$$

Example

$$
\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y
\end{aligned}
$$

Example

$$
\begin{aligned}
f(x, y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x, y, \alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y
\end{aligned}
$$

Example

$$
\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y \\
3y + y - 2 &= 0
\end{aligned}
$$

Example

$$
\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y \\
3y + y - 2 &= 0 \\
4y &= 2
\end{aligned}
$$

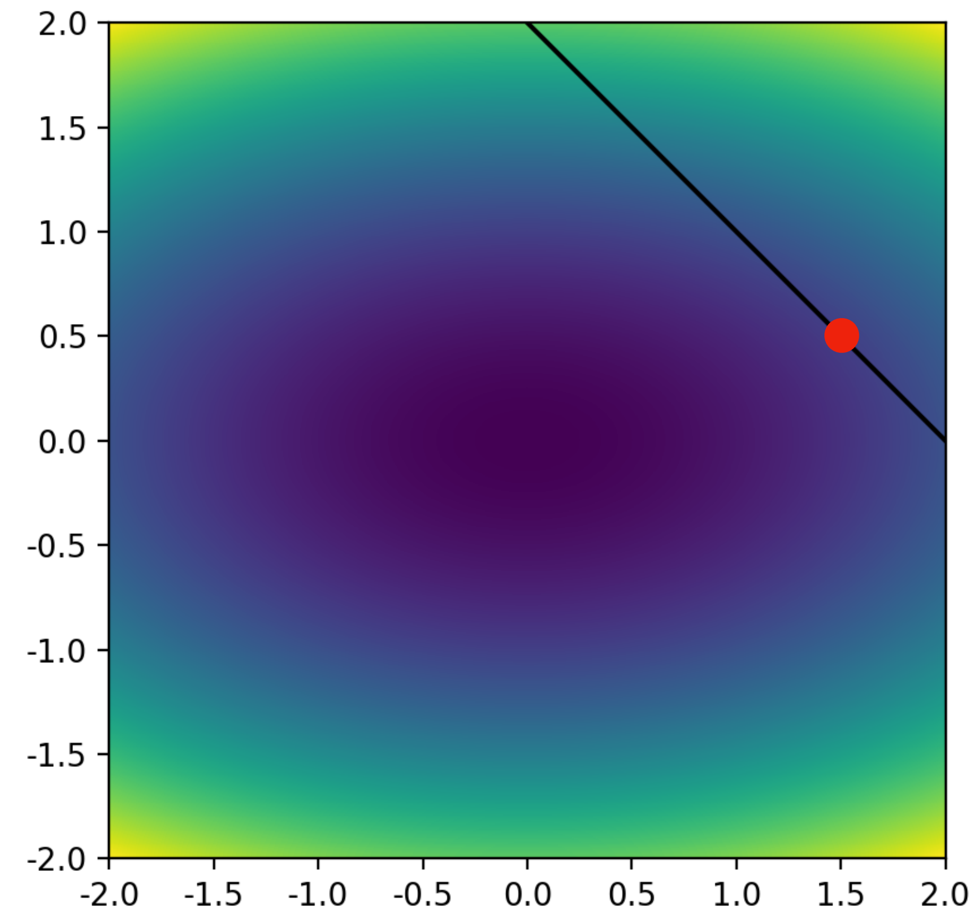
Jacob Whitehill, WPI

Example

$$
\begin{aligned}
f(x, y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x, y, \alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y \\
3y + y - 2 &= 0 \\
4y &= 2 \\
y &= 1/2
\end{aligned}
$$

Example

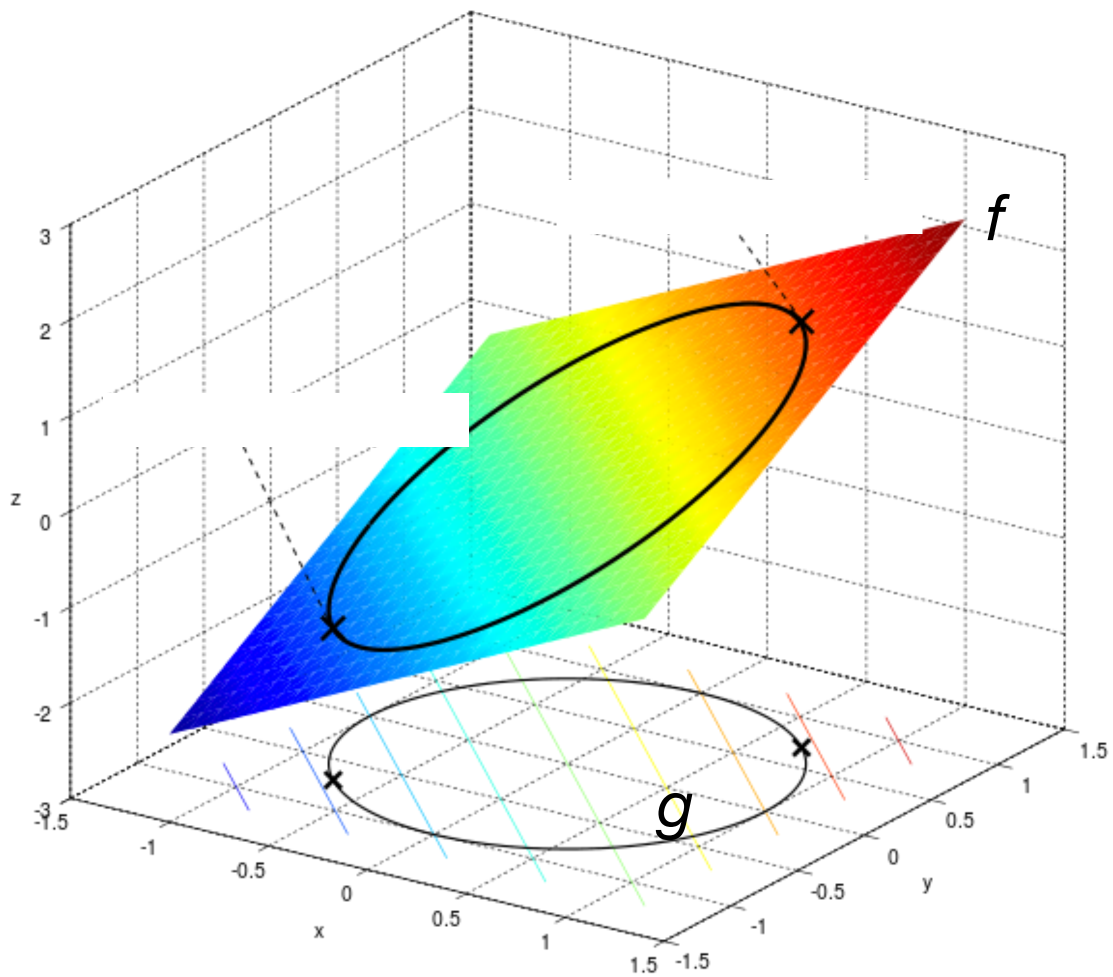
$$
\begin{aligned}
f(x,y) &= x^2 + 3y^2 \quad \text{subject to} \quad x + y = 2 \\
L(x,y,\alpha) &= x^2 + 3y^2 + \alpha(x + y - 2) \\
\frac{\partial L}{\partial x} &= 2x + \alpha = 0 \\
\frac{\partial L}{\partial y} &= 6y + \alpha = 0 \\
\frac{\partial L}{\partial \alpha} &= x + y - 2 = 0 \\
2x &= 6y \\
x &= 3y \\
3y + y - 2 &= 0 \\
4y &= 2 \\
y &= 1/2 \\
x &= 3/2
\end{aligned}
$$



Exercise

- Minimize:

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



Jacob Whitehill, WPI

Exercise

- Minimize:

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

Exercise

- Minimize:

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

Exercise

- Minimize:

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

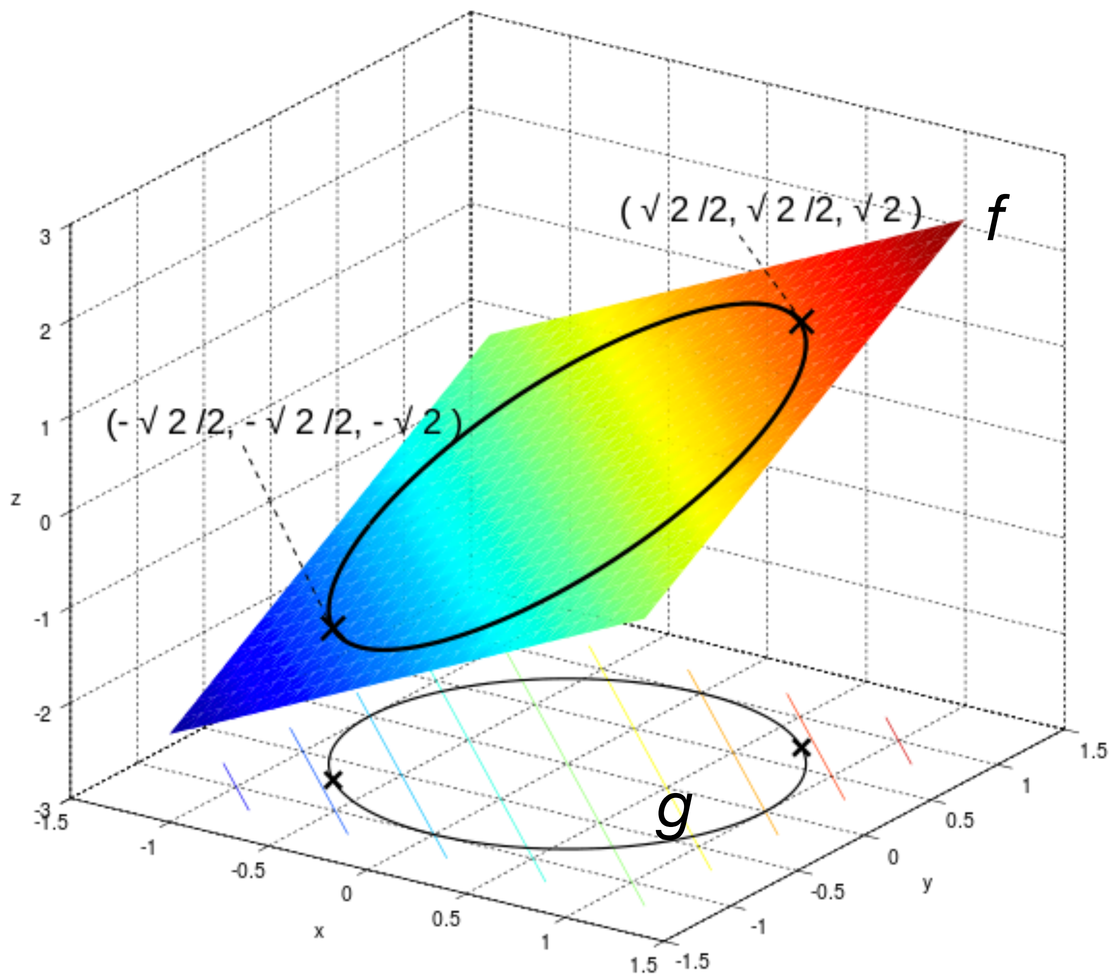
Exercise

- Minimize:

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

Exercise

- Try $x = y = +1/\sqrt{2}$: $f(+1/\sqrt{2}, +1/\sqrt{2}) = +2/\sqrt{2} = +\sqrt{2}/2$ **Maximum**
- Try $x = y = -1/\sqrt{2}$: $f(-1/\sqrt{2}, -1/\sqrt{2}) = -2/\sqrt{2} = -\sqrt{2}/2$ **Minimum**



Jacob Whitehill, WPI

KKT multipliers

Lagrange multipliers

- A generalization of Lagrange multipliers, which also handles inequality constraints, are KKT conditions.

- We define the optimization problem with:
  - The original optimization variables.
  - The Lagrange multiplier(s) α (one for each constraint).

- Note that either of the following Lagrangian formulations will work (since the value of α can compensate):

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) - \alpha g(\mathbf{w})$$

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) + \alpha g(\mathbf{w})$$

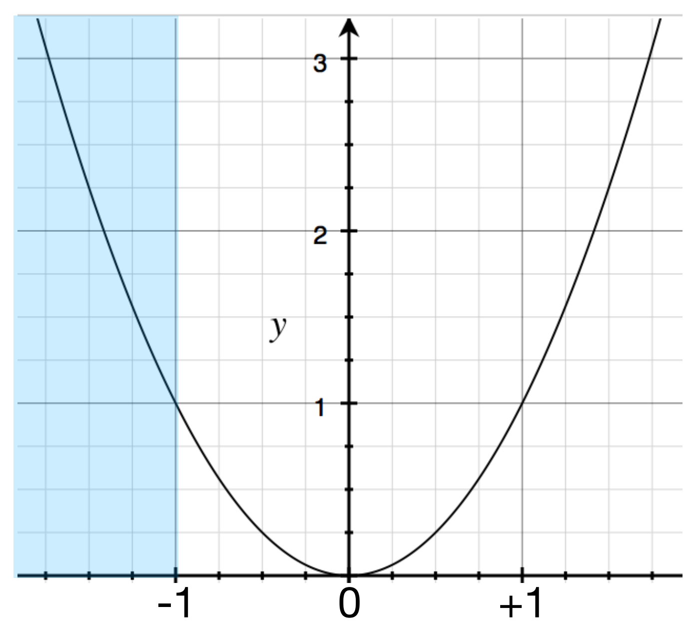
- However, with SVMs, the convention is:

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) - \alpha g(\mathbf{w})$$

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function *g*.

- Suppose we wish to minimize *f* subject to $g(x) \leq 0$:

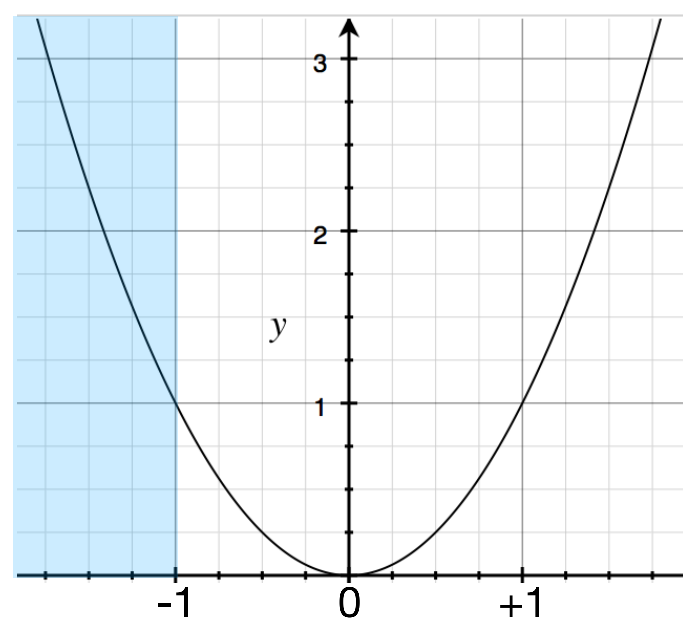


$g(x) =$?

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function *g*.

- Suppose we wish to minimize *f* subject to $g(x) \leq 0$:

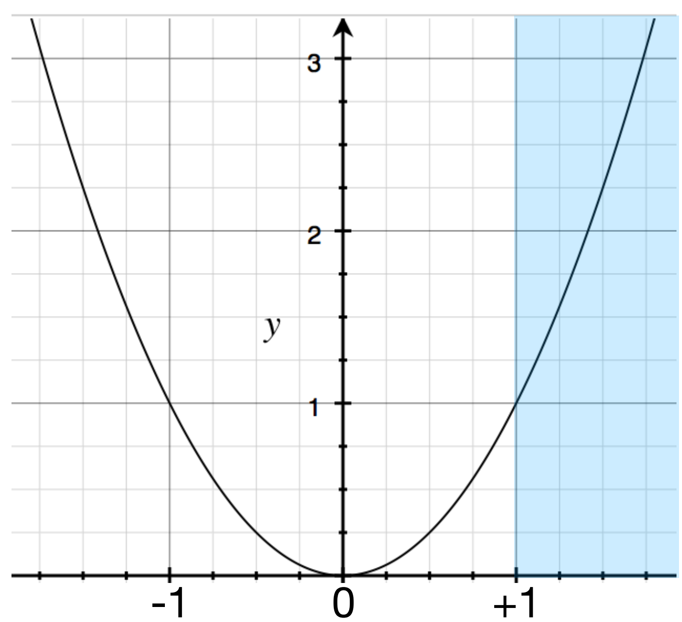


$g(x)=x+1$

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function *g*.

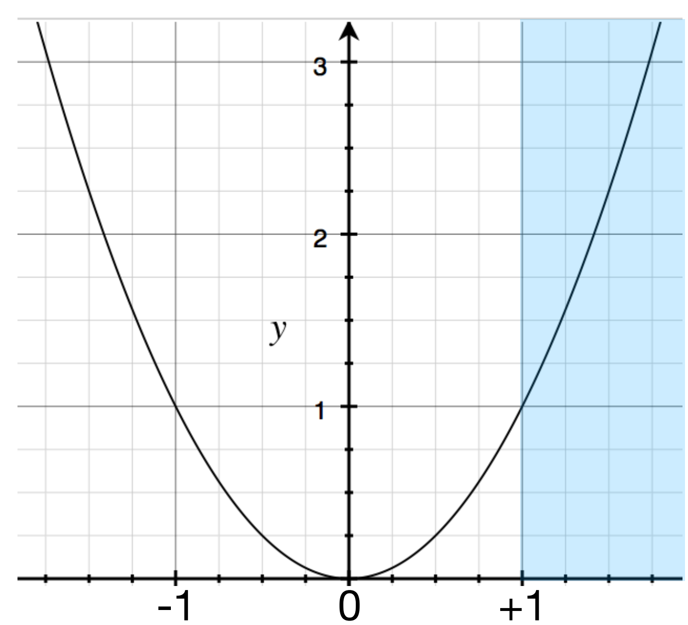
- Suppose we wish to minimize *f* subject to $g(x) \leq 0$:



$g(x)=$?

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function *g*.
- Suppose we wish to minimize *f* subject to $g(x) \leq 0$:

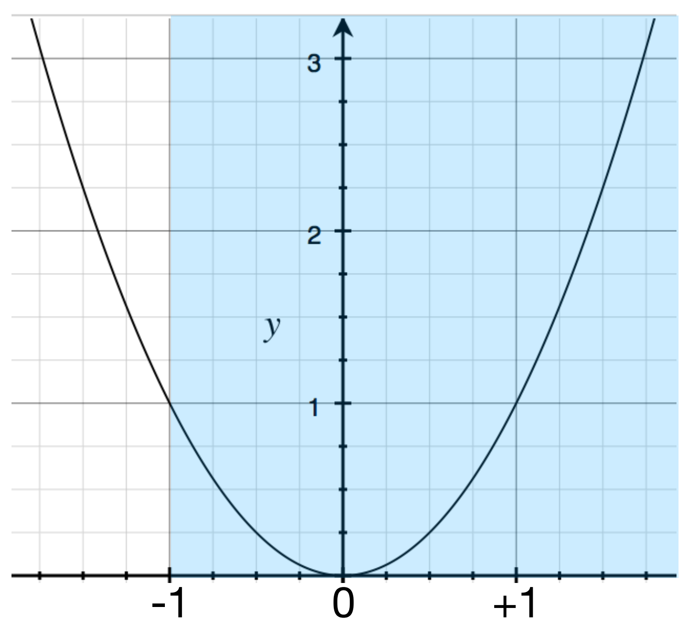


$g(x)=1-x$

Jacob Whitehill, WPI

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function *g*.
- Suppose we wish to minimize *f* subject to $g(x) \leq 0$:



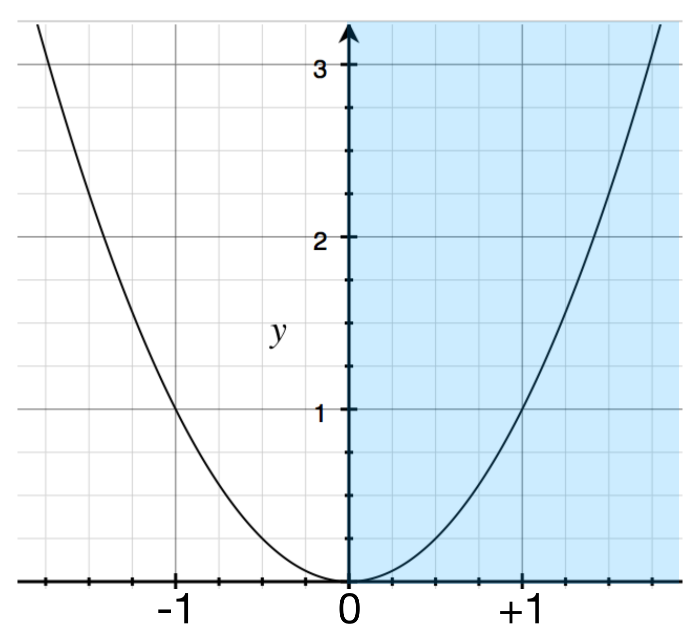
$g(x) = -1 - x$

Jacob Whitehill, WPI

Karush-Kuhn-Tucker (KKT) conditions

- As with Lagrange multipliers, we encode each constraint as a function g.

- Suppose we wish to minimize f subject to $g(x) \leq 0$:



$g(x) = -x$

Jacob Whitehill, WPI

Karush-Kuhn-Tucker (KKT) conditions

- Similarly as with Lagrange multipliers, with KKT conditions we also use a set of “multipliers” α (one for each constraint), sometimes known as **dual variables**.

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) - \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

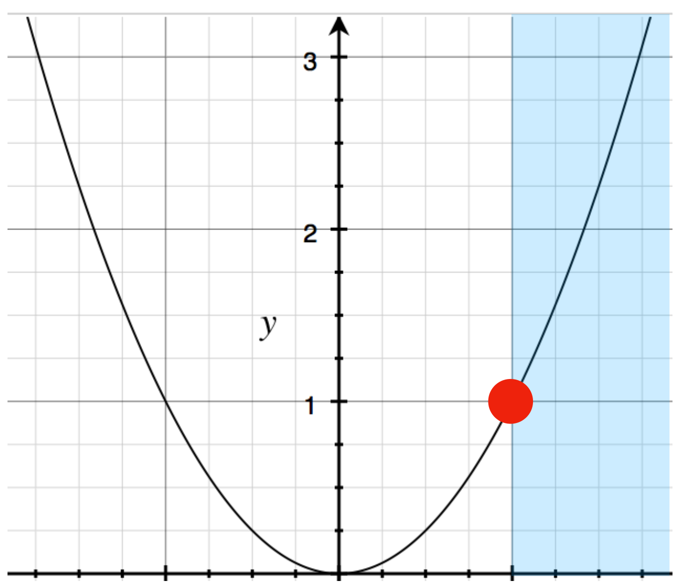
Karush-Kuhn-Tucker (KKT) conditions

- Similarly as with Lagrange multipliers, with KKT conditions we also use a set of "multipliers" α (one for each constraint), sometimes known as **dual variables**.

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) - \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

- Key points:

1. With *inequality* constraints, we require that each $\alpha_i \geq 0$.
2. At optimal solution:
    - $\alpha_i > 0$ if the constraint is **active**.



$g(x)=x$

Jacob Whitehill, WPI

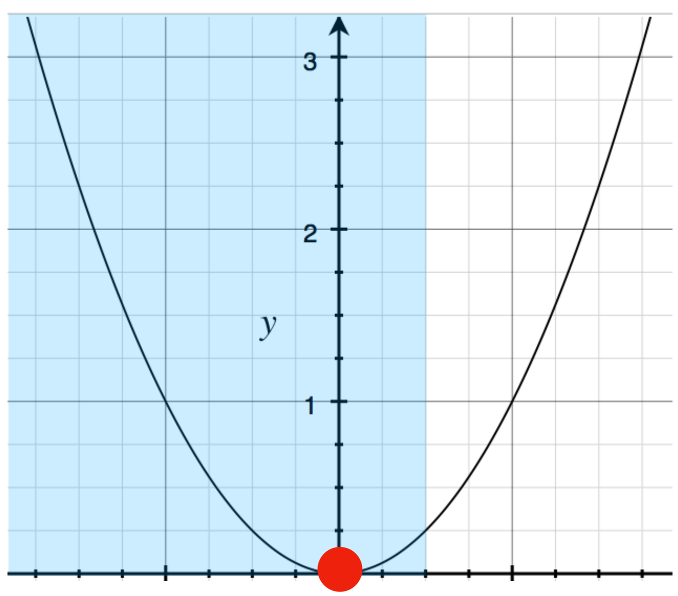
Karush-Kuhn-Tucker (KKT) conditions

- Similarly as with Lagrange multipliers, with KKT conditions we also use a set of “multipliers” α (one for each constraint), sometimes known as **dual variables**.

$$L(\mathbf{w}, \alpha) = f(\mathbf{w}) - \sum_{i=1}^{n} \alpha_i g_i(\mathbf{w})$$

- Key points:

1. With *inequality* constraints, we require that each $\alpha_i \geq 0$.
2. At optimal solution:
    - $\alpha_i > 0$ if the constraint is **active**.
    - $\alpha_i = 0$ if the constraint is **inactive**.



$g(x)=x-1/2$

Support vector machines

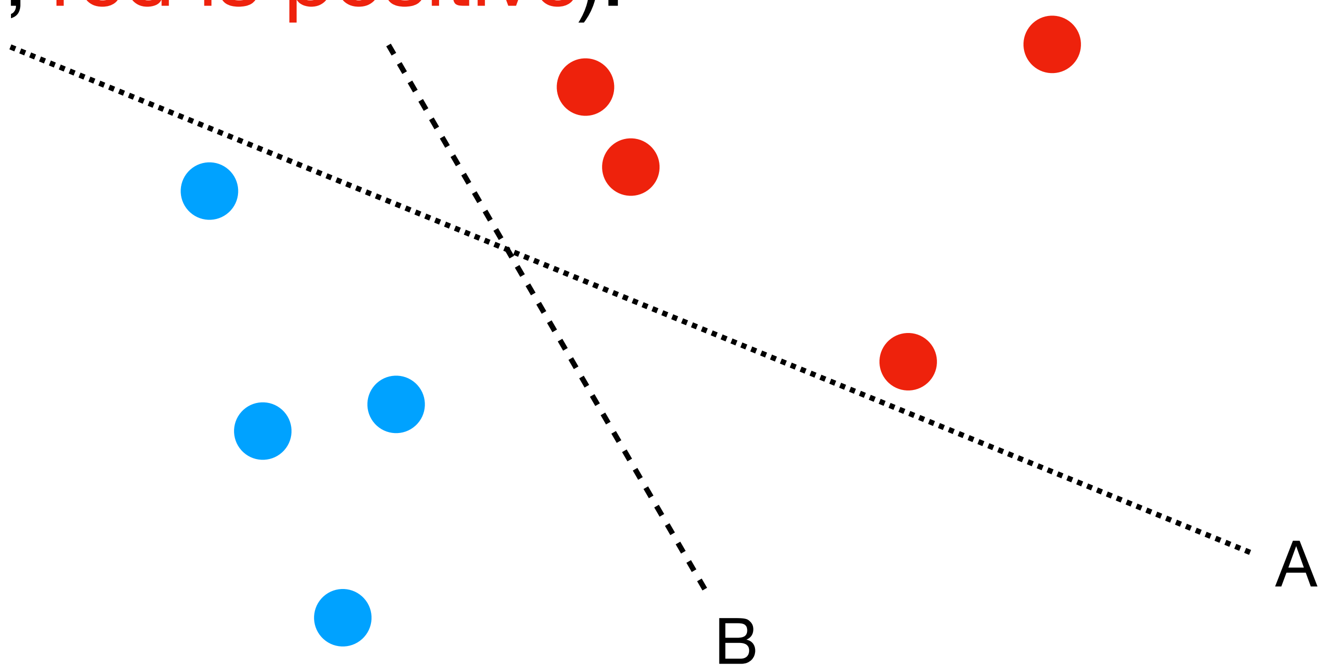
Support vector machines

- **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).

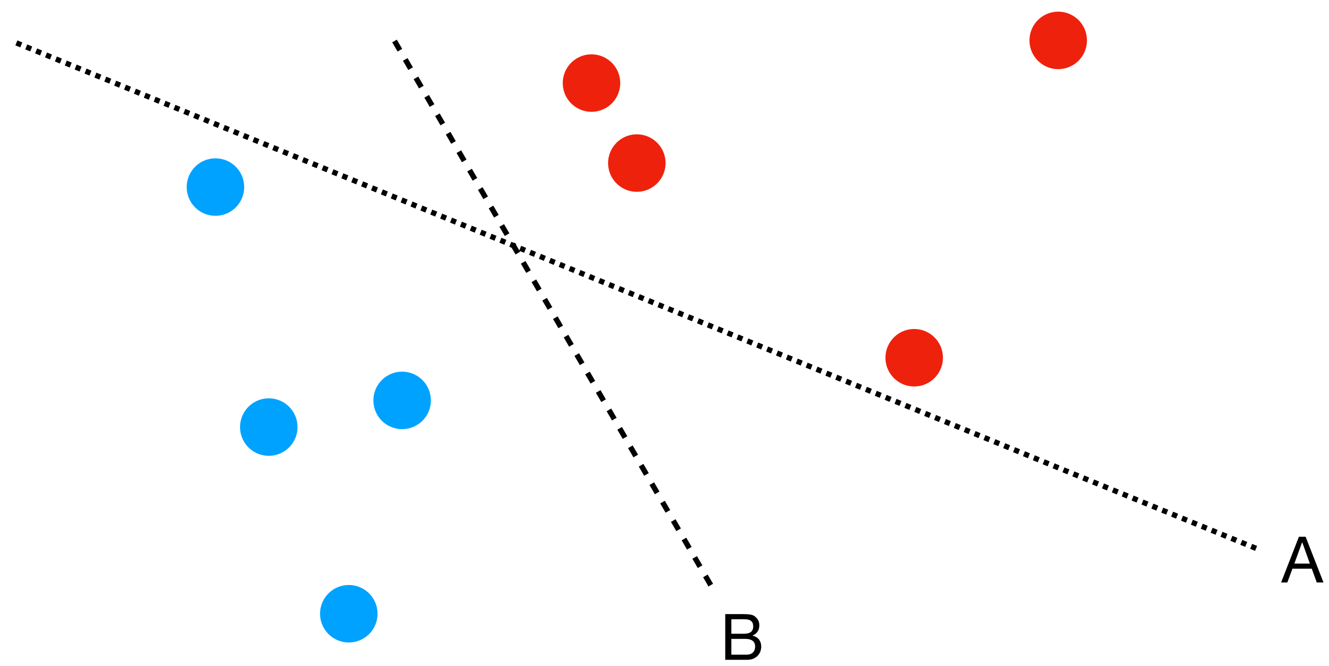
Support vector machines

- Suppose we have the following set of training data (blue is negative, red is positive):



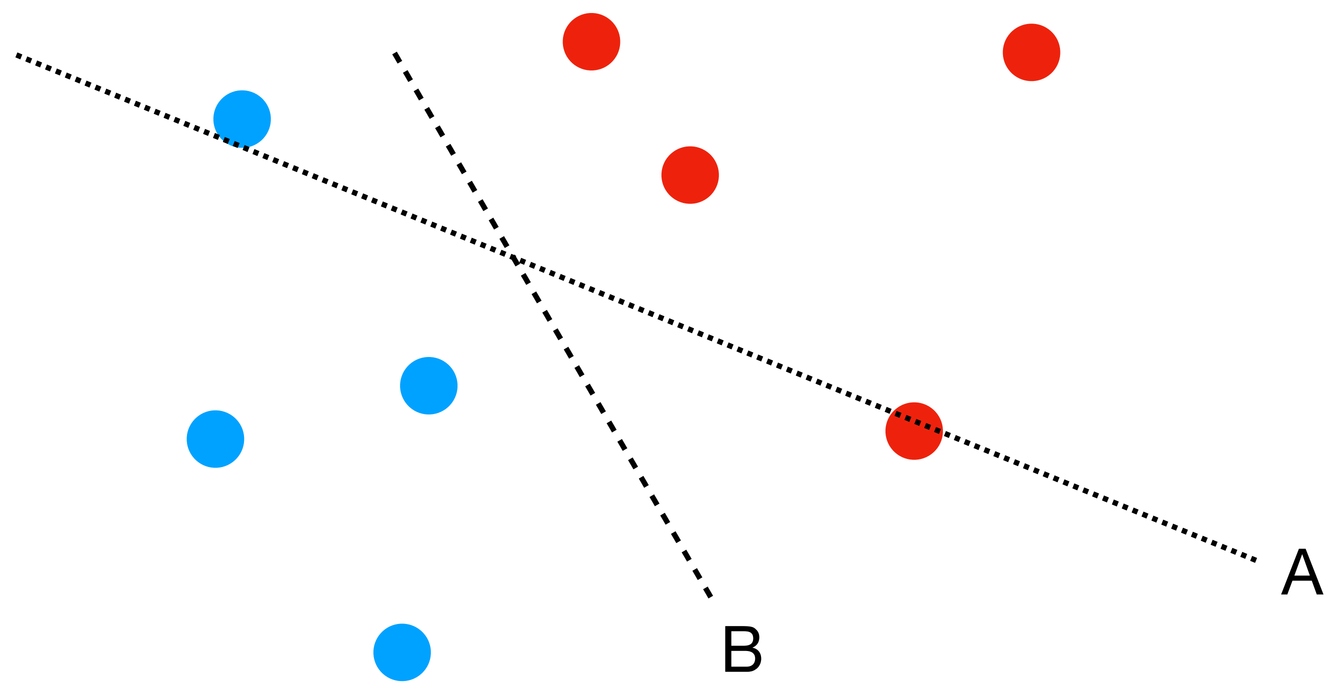
- Examples above the line will be classified as positive; examples below the line will be classified as negative.
- Intuitively, which line (or **hyperplane** in higher dimensions) would likely perform better on *testing* data, and why?

Support vector machines



- B is farther from any of the data points than A is — it has a bigger “margin”.

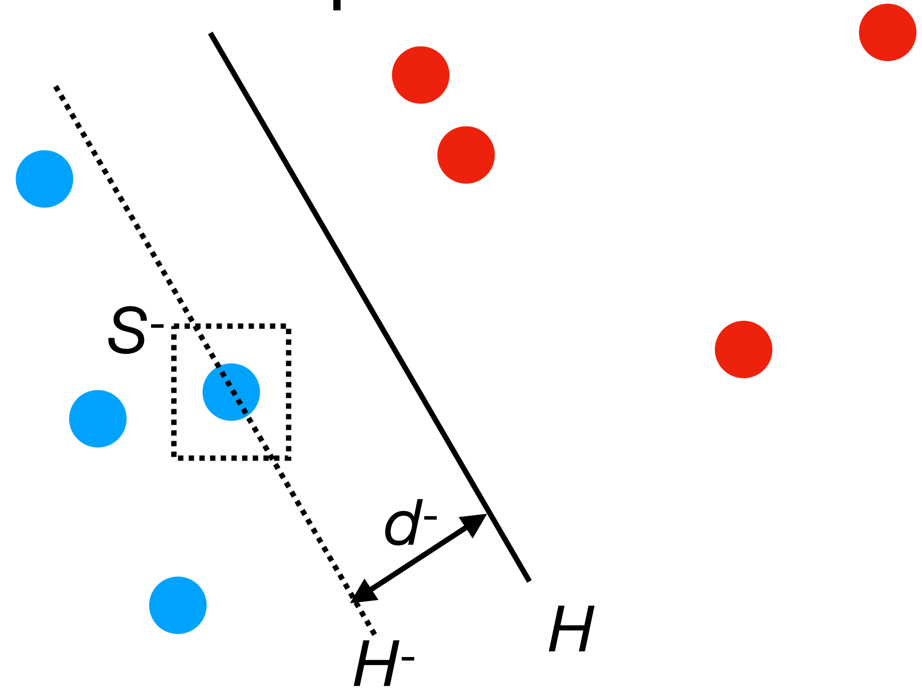
Support vector machines



- B is farther from any of the data points than A is — it has a bigger “margin”.
- If we “jitter” the data slightly, then B will still perfectly separate the two classes, whereas A will not.

Support vector machines

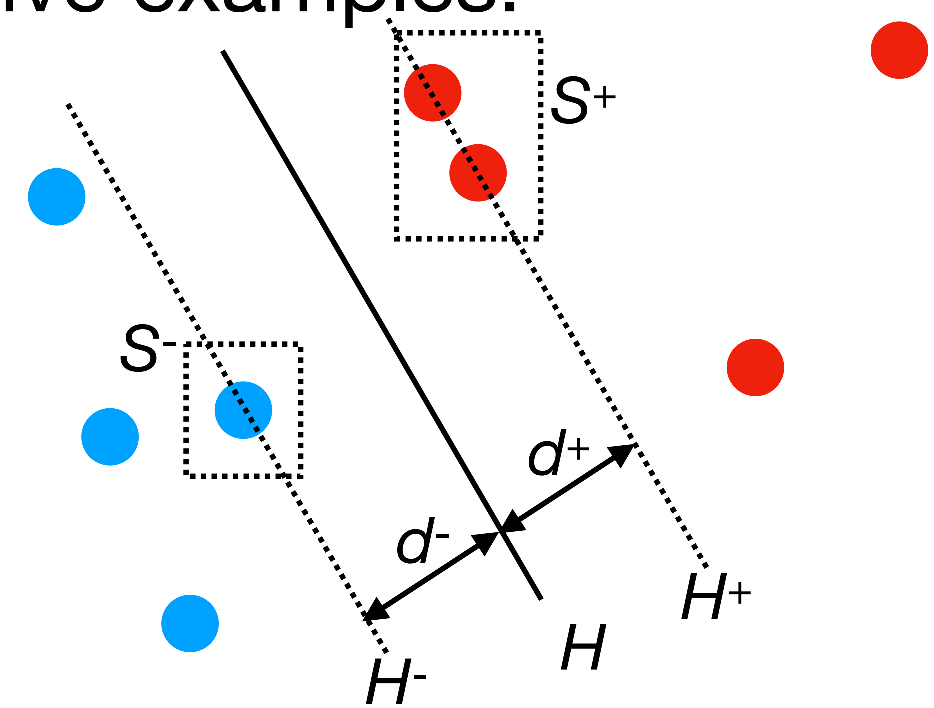
- For any hyperplane H that perfectly separates the positive from the negative examples:



- 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

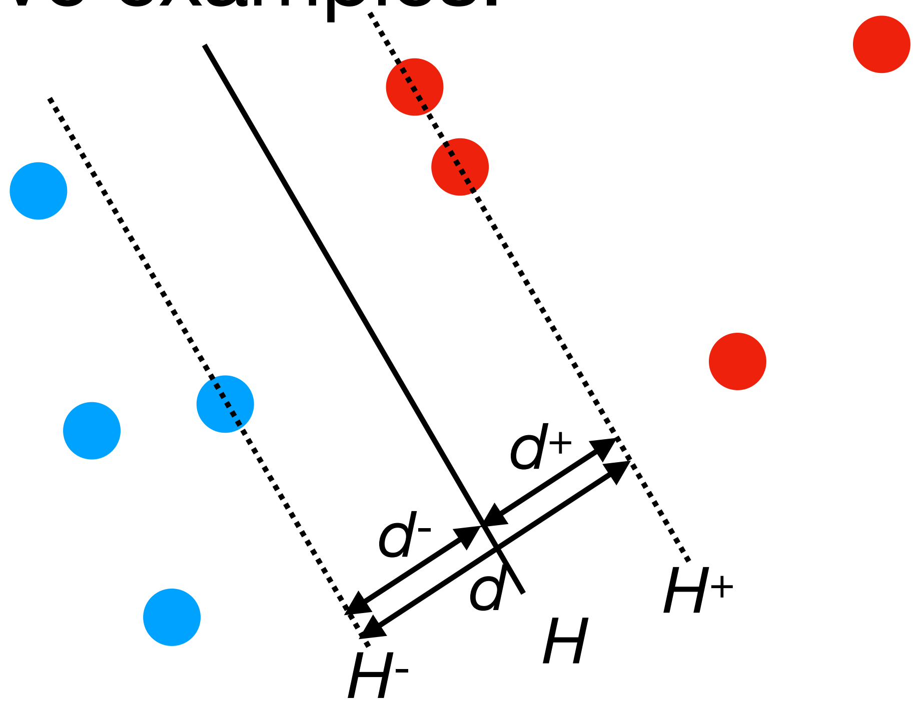
- For any hyperplane H that perfectly separates the positive from the negative examples:



- 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:



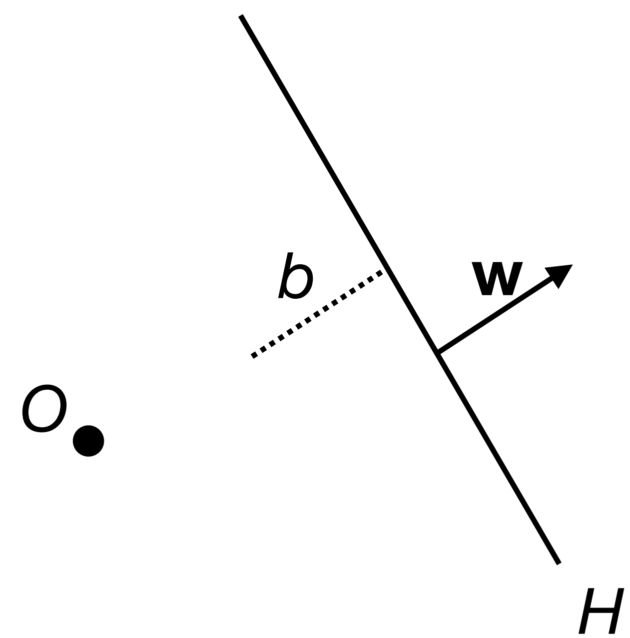
- Let d denote the **margin** — the sum of d^+ and d^-.
- The optimization objective of SVMs is to find a separating hyperplane H that **maximizes** d.

Hyperplanes

Hyperplanes

- Informally, a hyperplane is the generalization of a “plane” into higher-dimensional spaces. It splits the ambient space into two “halves”.
- In 1-D, a hyperplane is a point.
- In 2-D, a hyperplane is a line.
- In 3-D, a hyperplane is a plane.
- In 4-D, …

Defining a hyperplane



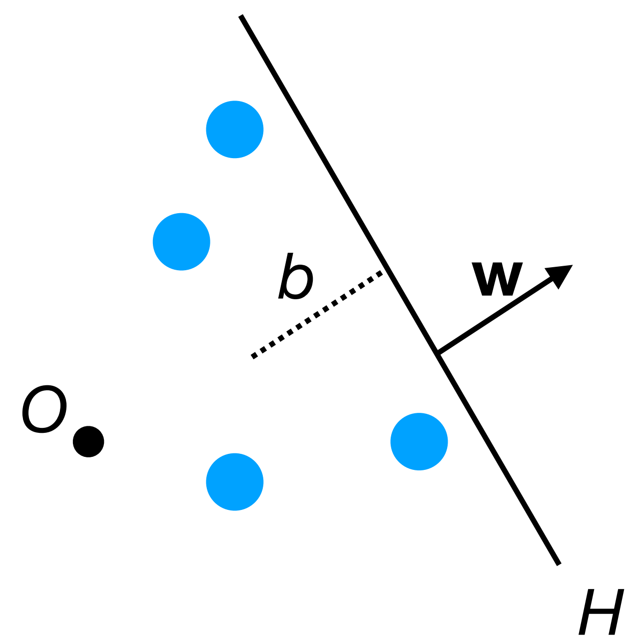
- A **hyperplane** is defined by a normal vector **w** ($\perp$ to H) and a bias b that is proportional to the distance to the origin.

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

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

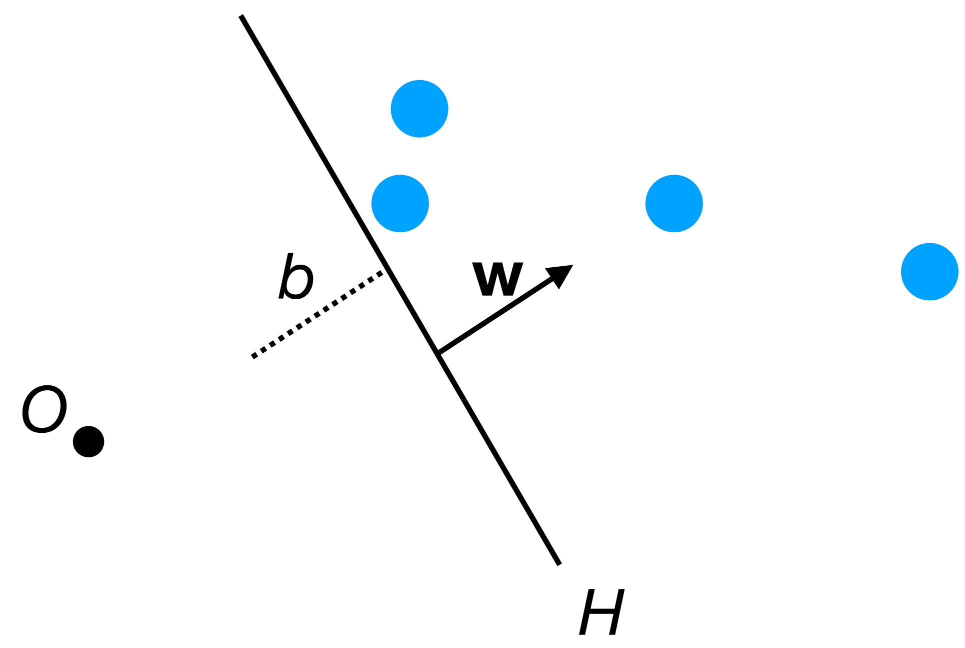
Jacob Whitehill, WPI

Defining a hyperplane



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

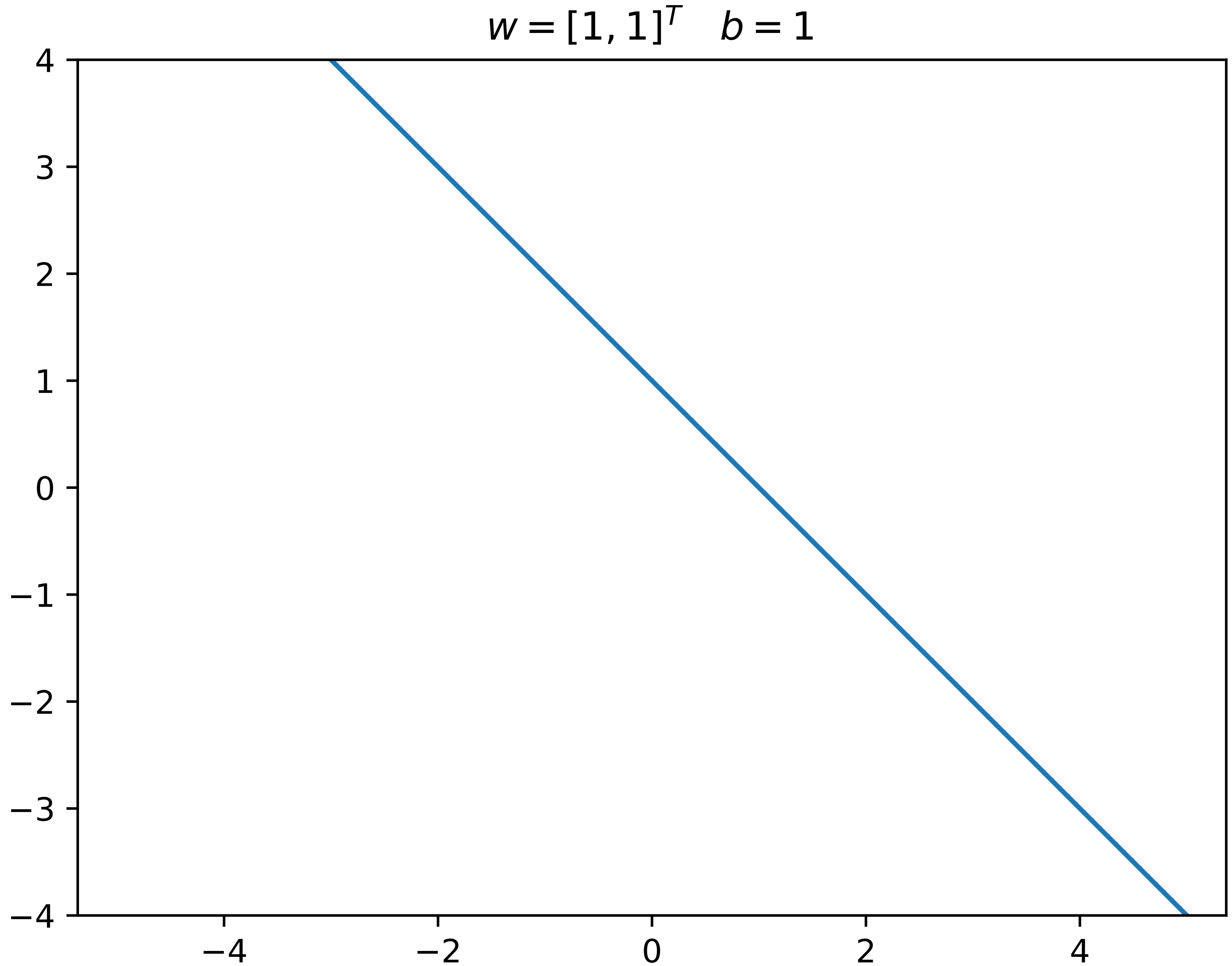
Defining a hyperplane



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

Hyperplane examples

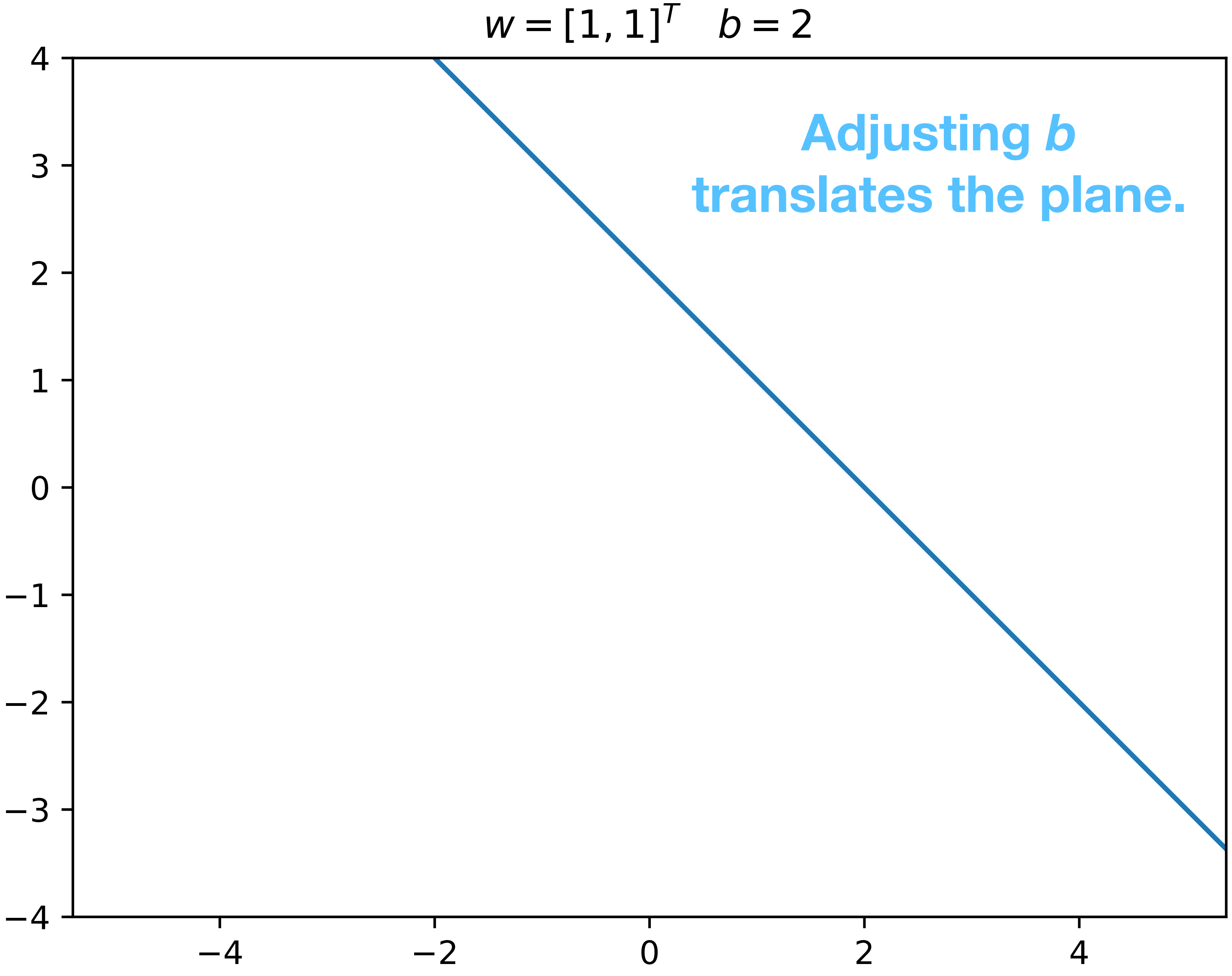
$$H = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{w} + b = 0\}$$



Jacob Whitehill, WPI

Hyperplane examples

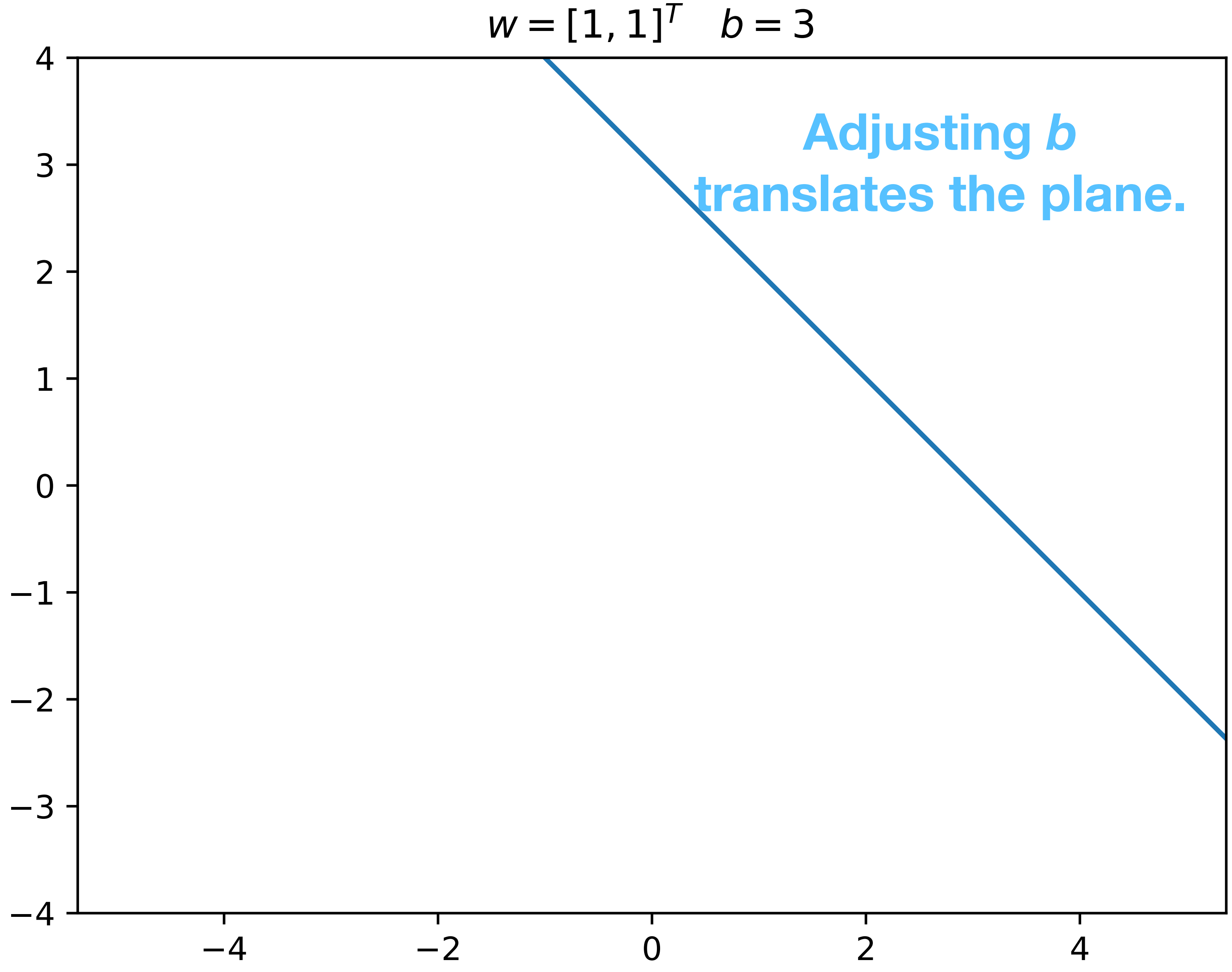
$$H = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{x}^\top \mathbf{w} + b = 0\}$$



Jacob Whitehill, WPI

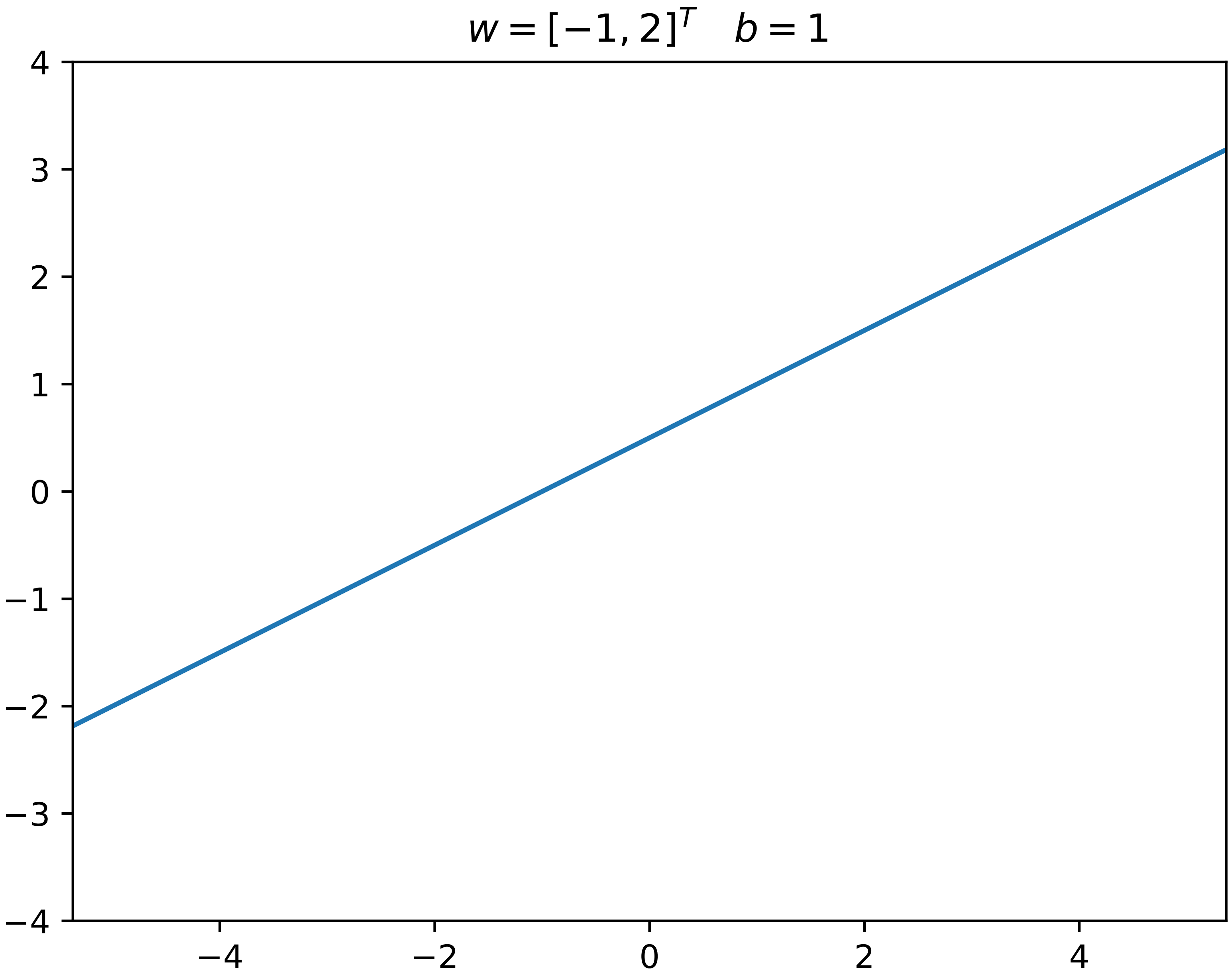
Hyperplane examples

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



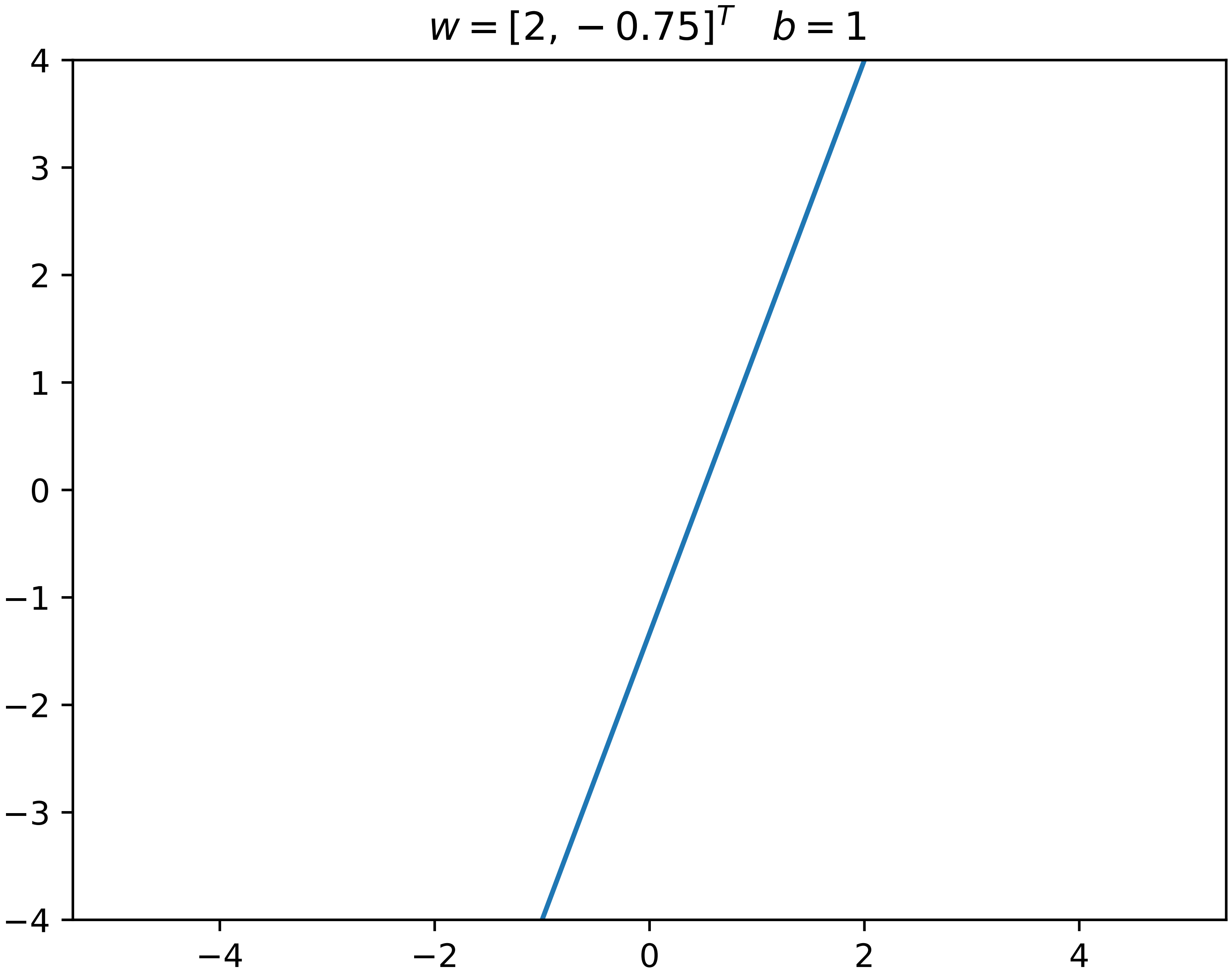
Hyperplane examples

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



Hyperplane examples

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



Hyperplane examples

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