

Today

- Recap yesterday
- Logistic regression: using regression tools for classification
- Neural network basics

Yesterday

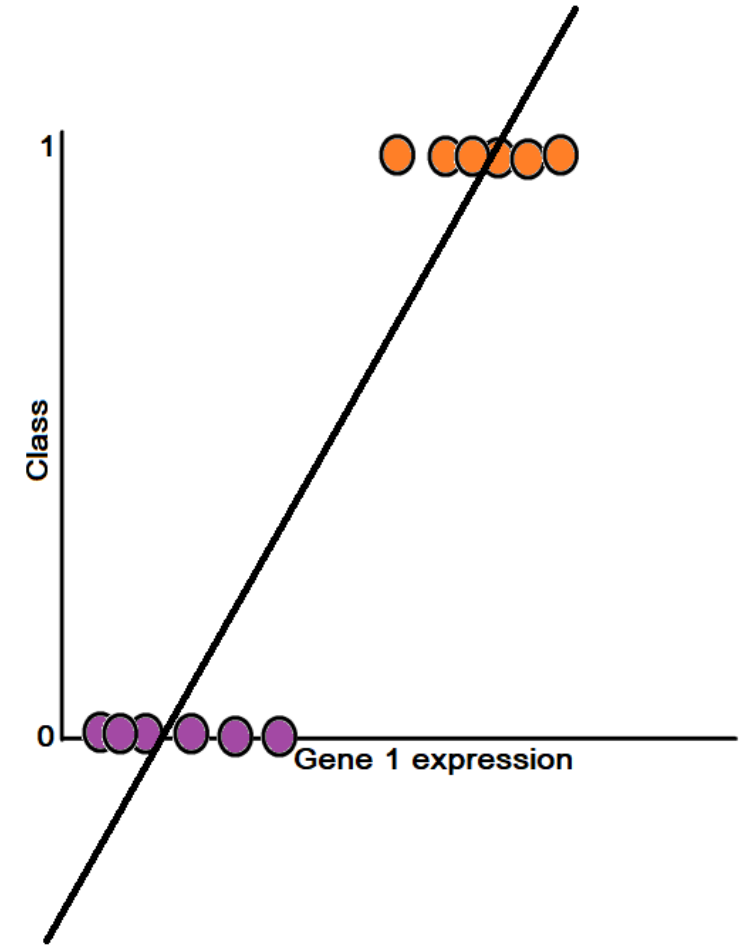
- Cost function: (differentiable) function that shows how wrong an estimate is for given parameters.
- Gradient descent: one common way to minimise the cost function automatically, i.e. to get optimal parameters
- Linear regression: very simple model that assumes that value to predict is linear combination of input features.
- Overfitting and underfitting, bias and variance: want our model to work well for unseen data. Need just enough model freedom given the complexity of our problem. How:
 - Cross-validation to measure ability to generalise + get best hyperparameters
 - Use learning curves to diagnose bias vs. variance

Logistic regression

- Use regression-like framework for classification

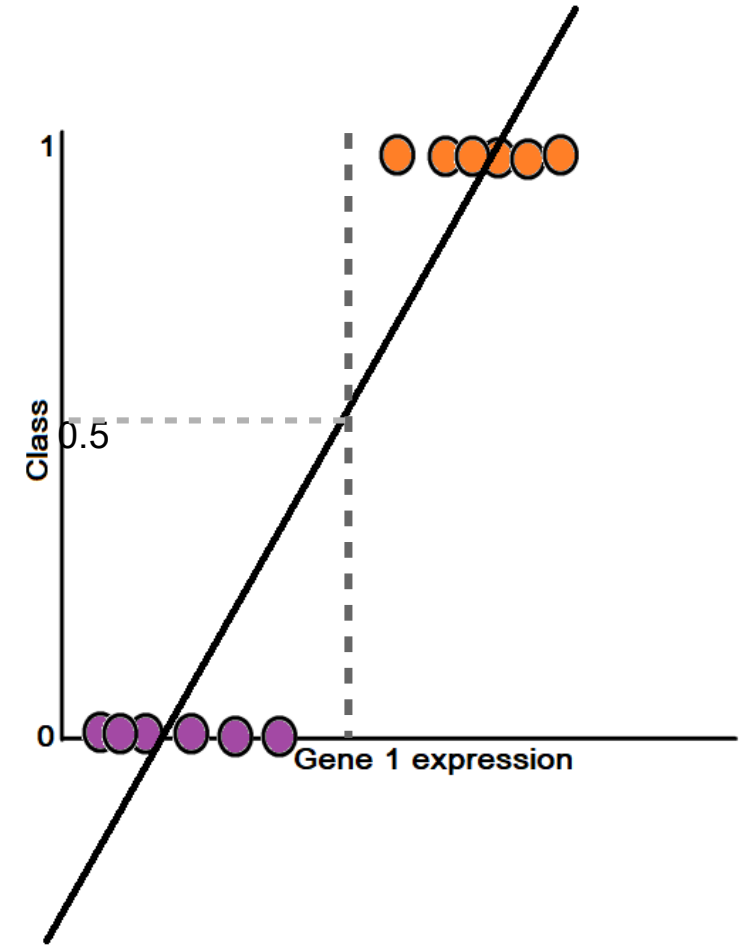
Logistic regression

- Naïve idea:
Train a linear regression. If
Class ≥ 0.5 , predict class 1.
Otherwise, class 0.



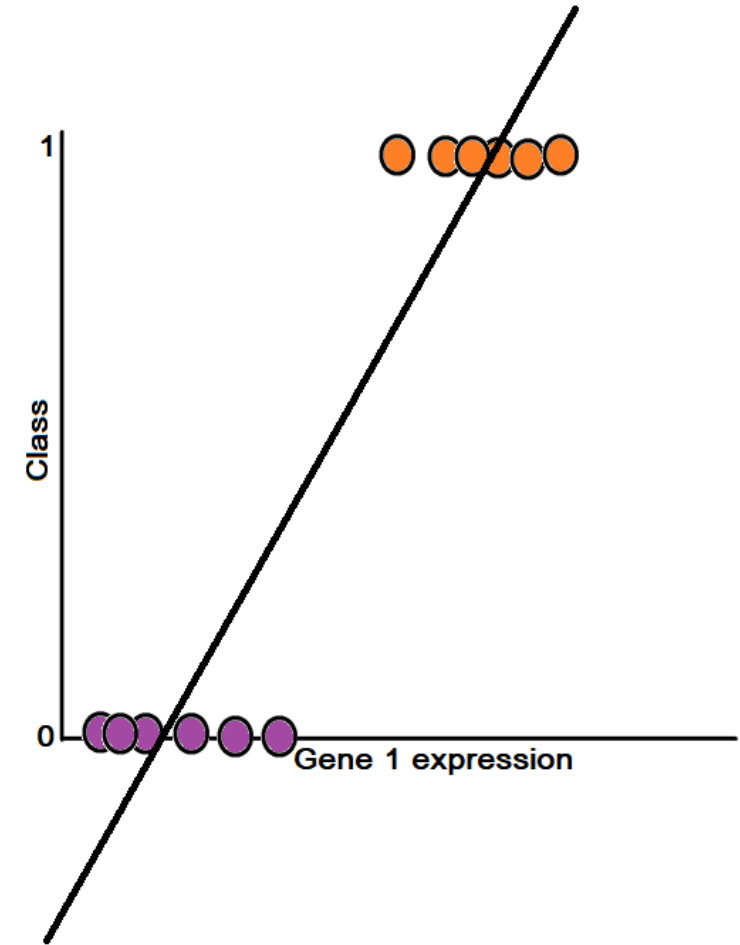
Logistic regression

- Naïve idea:
Train a linear regression. If
Class ≥ 0.5 , predict class 1.
Otherwise, class 0.



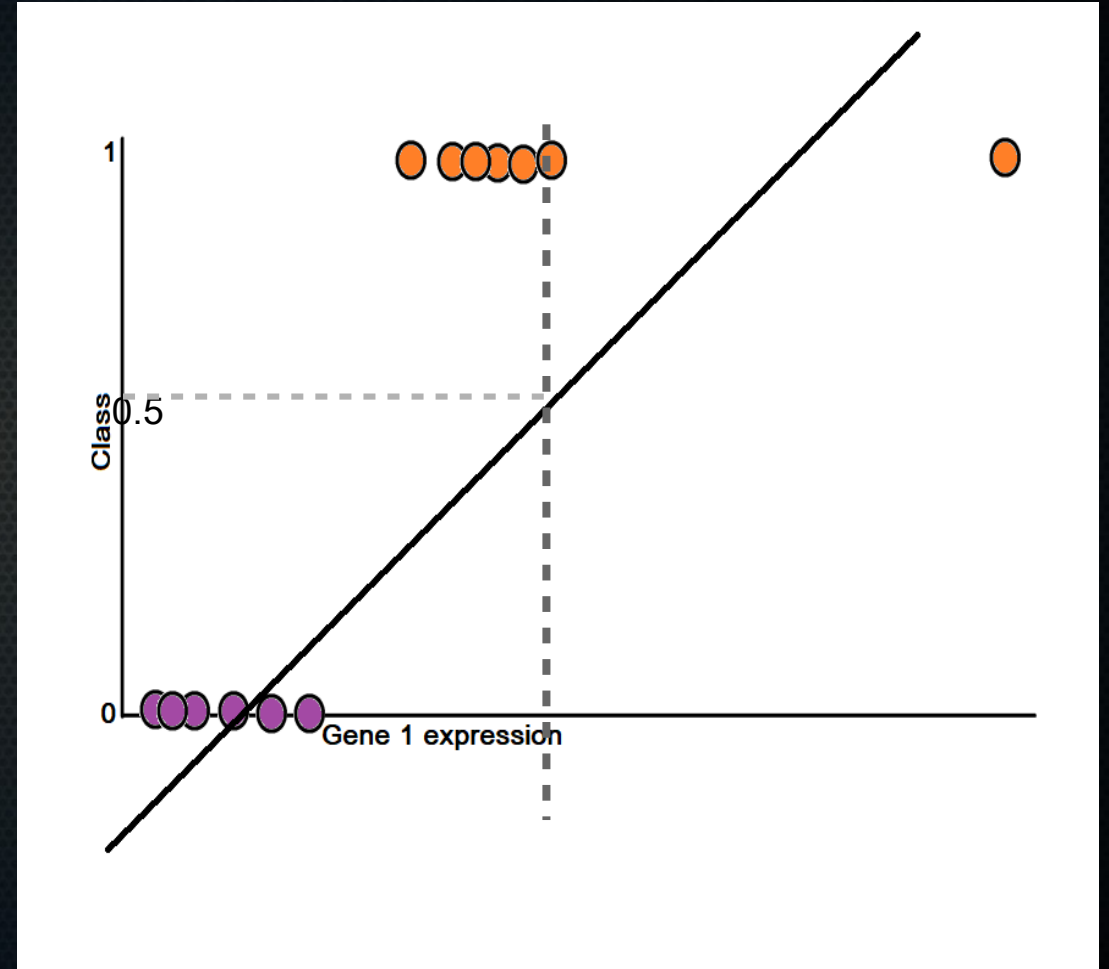
Logistic regression

- Naïve idea:
Train a linear regression. If $\text{Class} \geq 0.5$, predict class 1. Otherwise, class 0.
- Problems:
 - You can predict class > 1 and < 0 , while that is not possible in reality.



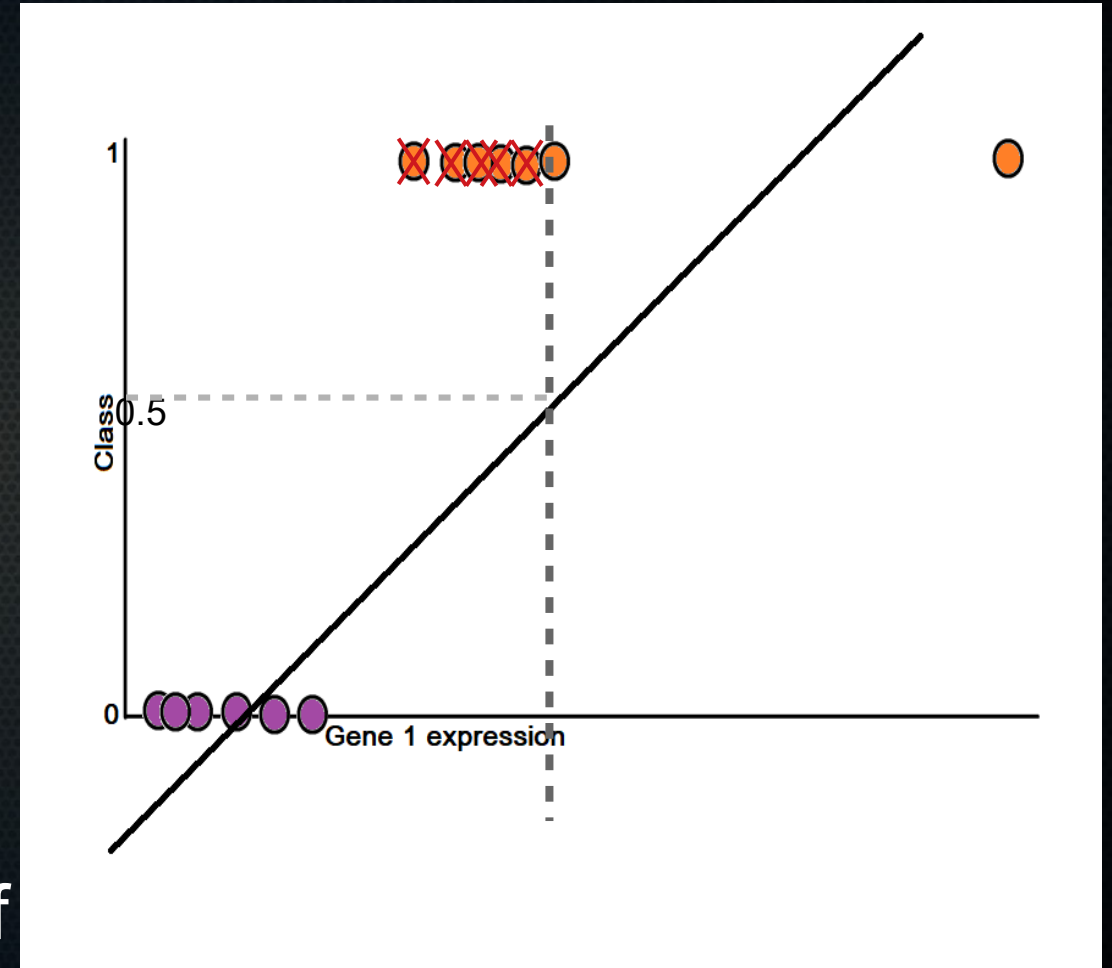
Logistic regression

- Naïve idea:
Train a linear regression. If $\text{Class} \geq 0.5$, predict class 1. Otherwise, class 0.
- Problems:
 - You can predict class > 1 and < 0 , while that is not possible in reality.
 - This example seemed to work, but quickly breaks down \rightarrow



Logistic regression

- Naïve idea:
Train a linear regression. If $\text{Class} \geq 0.5$, predict class 1. Otherwise, class 0.
- Problems:
 - You can predict class > 1 and < 0 , while that is not possible in reality.
 - This example seemed to work, but quickly breaks down \rightarrow get what is basically confirmation of hypothesis, but perform worse!



Logistic regression

- What we want:
 - Use the information that we only have two classes, 0 or 1.
 - Hypothesis function should output only numbers between 0 or 1.

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x \longrightarrow [0.5 \quad 3 \quad -1.5] \cdot \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix}$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x \quad \longrightarrow \quad \underbrace{[0.5 \quad 3 \quad -1.5]}_{\text{Learned parameters (theta 0 – theta 2)}} \cdot \underbrace{\begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix}}_{\text{Features for one sample (x0 = 1, intercept term, 2 data-derived features x1 and x2)}}$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x \longrightarrow [0.5 \quad 3 \quad -1.5] \cdot \begin{bmatrix} 1 \\ 3 \\ 8 \end{bmatrix} = 0.5 \cdot 1 + 3 \cdot 3 - 1.5 \cdot 8 = -2.5$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x$$

- Change that to the following:

$$h_{\theta}(x) = g(\theta^T \cdot x)$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x$$

- Change that to the following:

$$h_{\theta}(x) = g(\theta^T \cdot x) \longrightarrow g(z) = \frac{1}{1 + e^{-z}}$$

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x$$

- Change that to the following:

$$h_{\theta}(x) = g(\theta^T \cdot x) \longrightarrow g(z) = \frac{1}{1 + e^{-z}}$$

- What does that look like?

Sigmoid or logistic function

- Before, our hypothesis function was of the form:

$$h_{\theta}(x) = \theta^T \cdot x$$

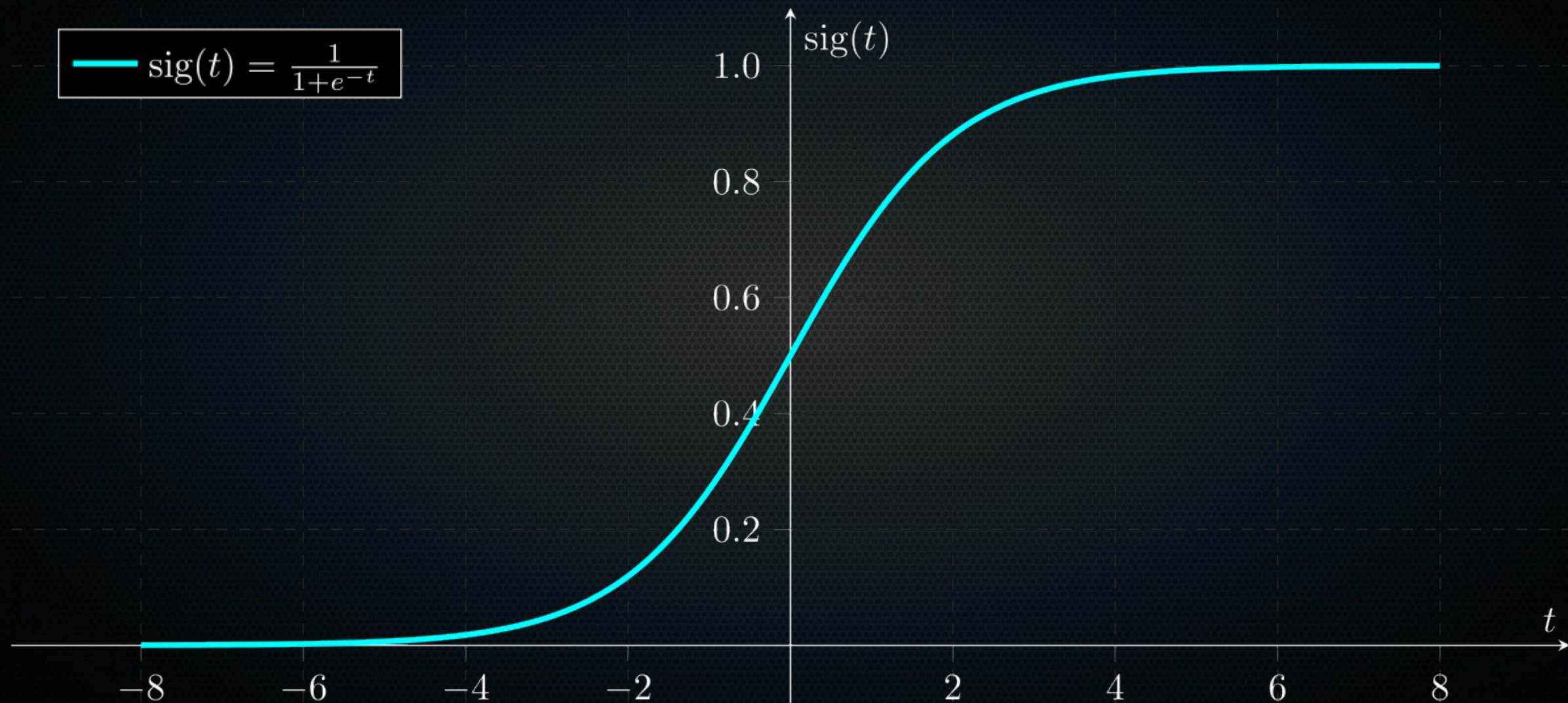
- Change that to the following:

$$h_{\theta}(x) = g(\theta^T \cdot x) \longrightarrow g(z) = \frac{1}{1 + e^{-z}}$$

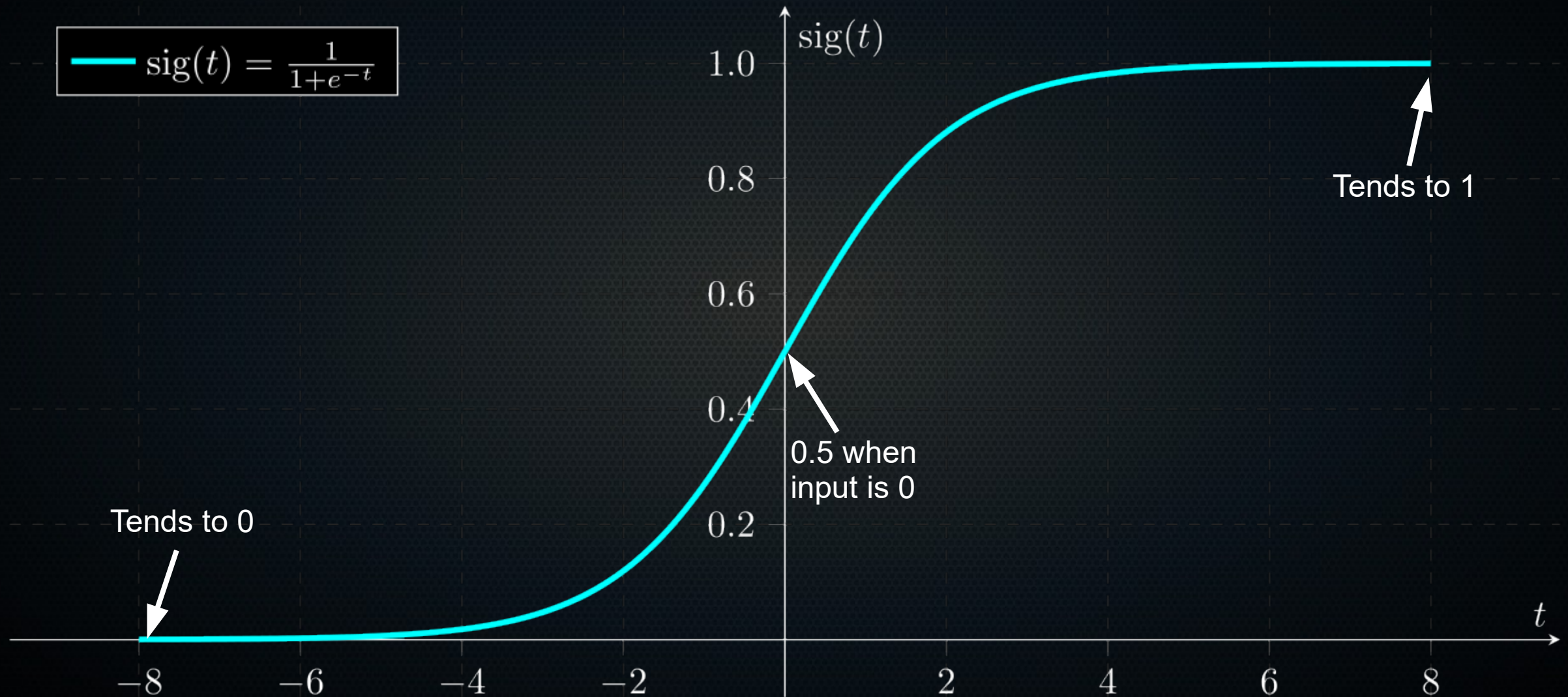
- What does that look like? $z \rightarrow \infty, e^{-z} \rightarrow 0$

$$z \rightarrow -\infty, e^{-z} \rightarrow \infty$$

What does the sigmoid function look like?

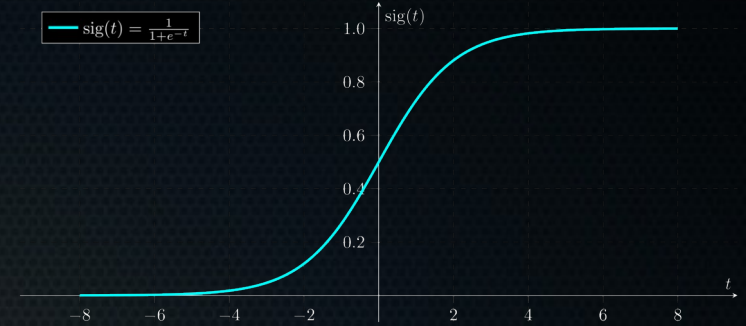


What does the sigmoid function look like?



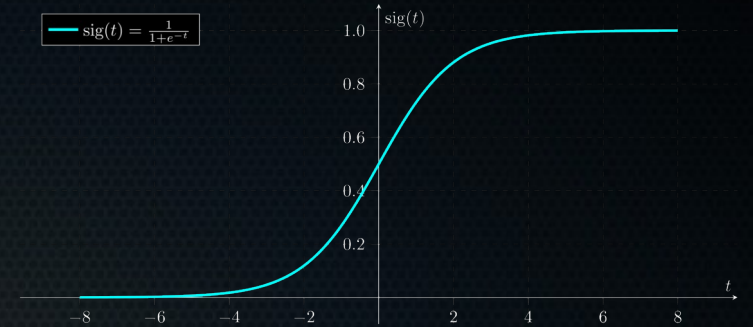
Sigmoid or logistic function

- How do we work with this? $h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}}$



Sigmoid or logistic function

- How do we work with this? $h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}}$
 - Interpret outcome of $h_{\theta}(x)$ as probability that class = 1 given the features.



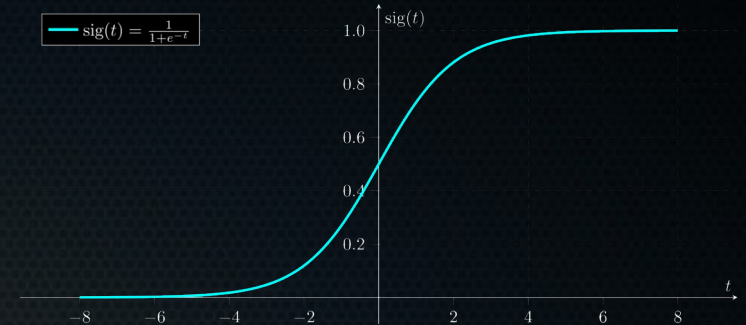
Sigmoid or logistic function

- How do we work with this? $h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}}$

- Interpret outcome of $h_{\theta}(x)$ as probability that class = 1 given the features. Example:

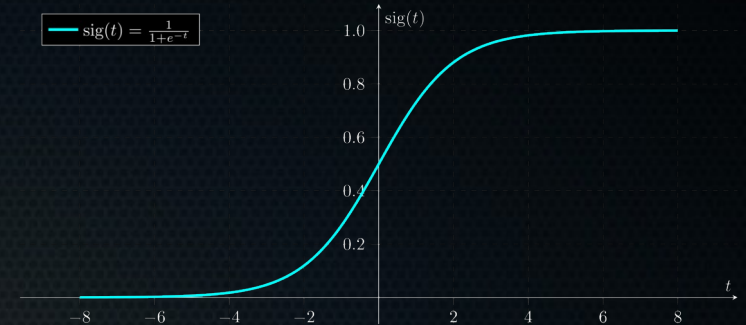
$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{Tumor size} \\ \text{Neovascularisation level} \end{bmatrix}$$

$h_{\theta}(x) = 0.8 \longrightarrow$ 80% chance of tumor being malignant



Sigmoid or logistic function

- How do we work with this? $h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}}$



- Interpret outcome of $h_{\theta}(x)$ as probability that class = 1 given the features. Example:

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{Tumor size} \\ \text{Neovascularisation level} \end{bmatrix}$$

$h_{\theta}(x) = 0.8 \longrightarrow$ 80% chance of tumor being malignant (class 1)
100% - 80% \rightarrow 20 % chance of being benign (class 0)

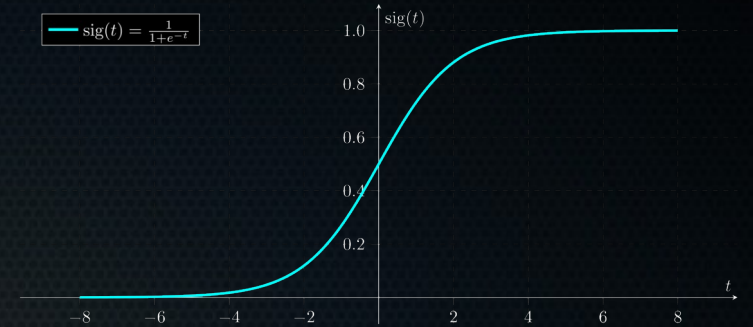
Sigmoid or logistic function

▪ How do we work with this? $h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}}$

- Interpret outcome of $h_{\theta}(x)$ as probability that class = 1 given the features.
- Formally:

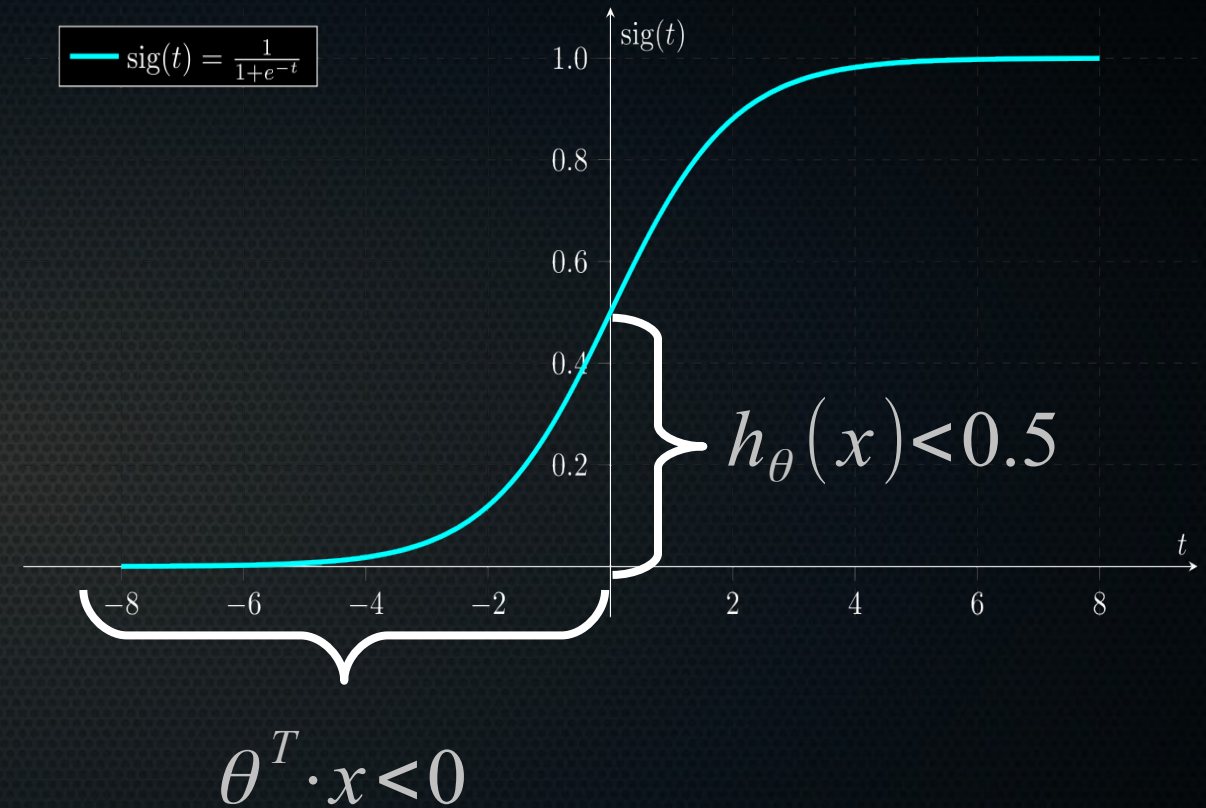
$$h_{\theta}(x) = \frac{1}{1 + e^{-(\theta^T \cdot x)}} = p(y=1|x;\theta)$$

$$p(y=0|x;\theta) = 1 - h_{\theta}(x)$$



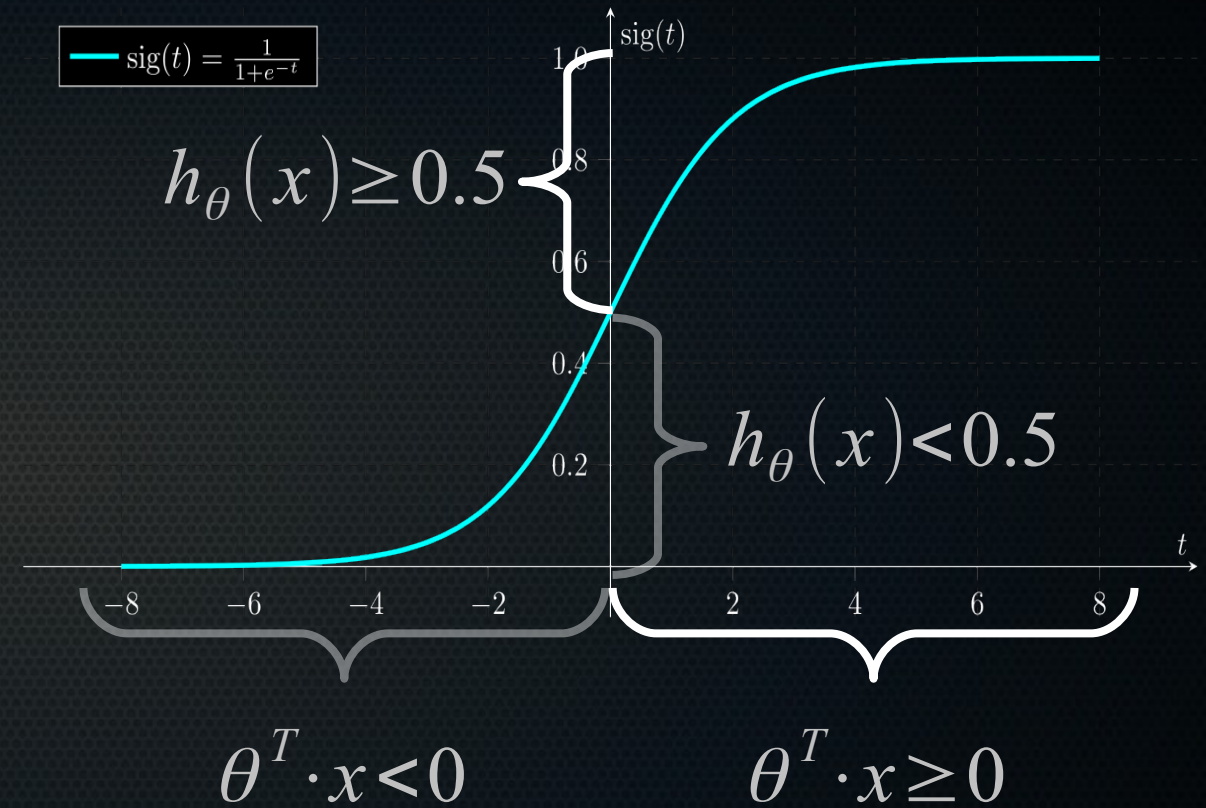
Decision boundary

- Threshold:



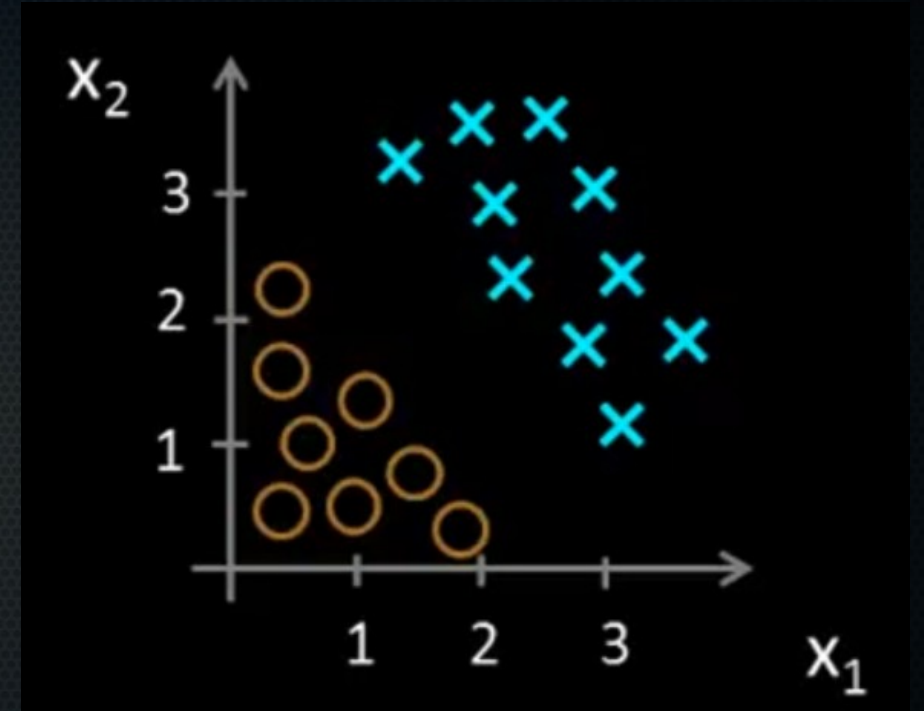
Decision boundary

- Threshold:



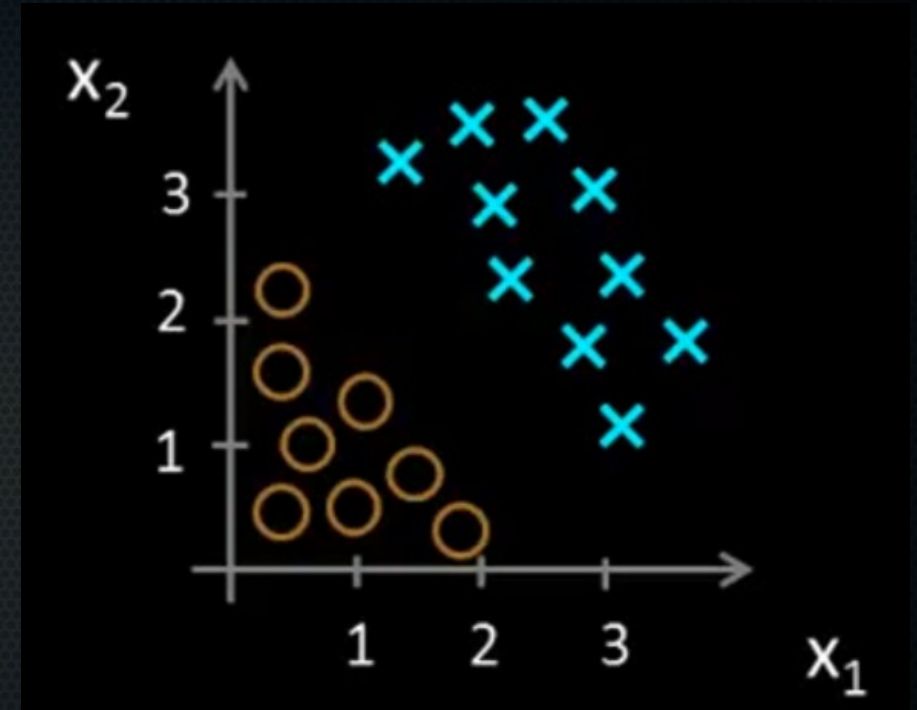
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$



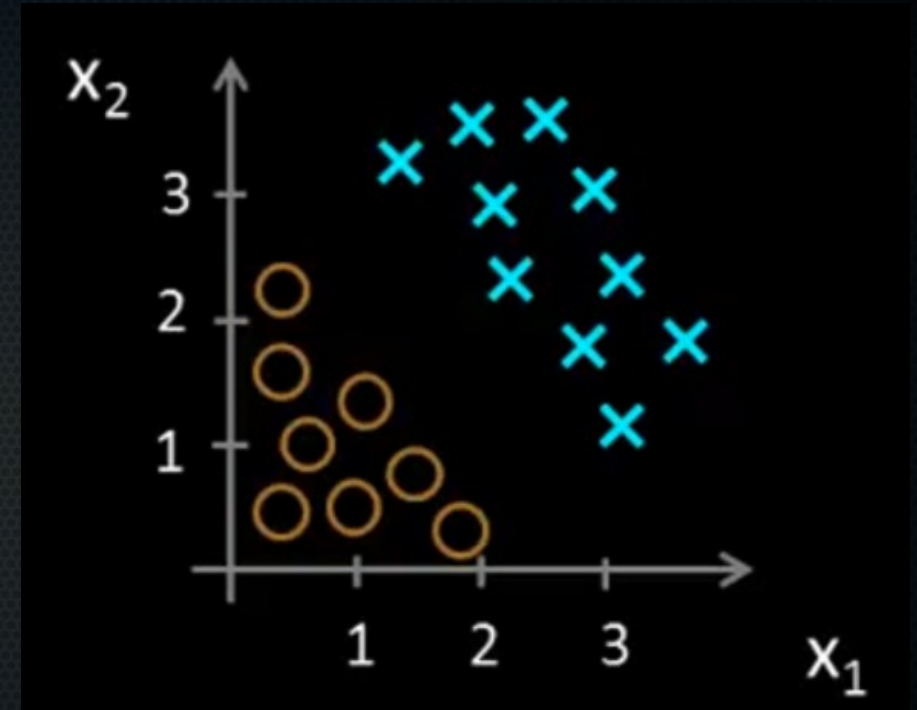
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$



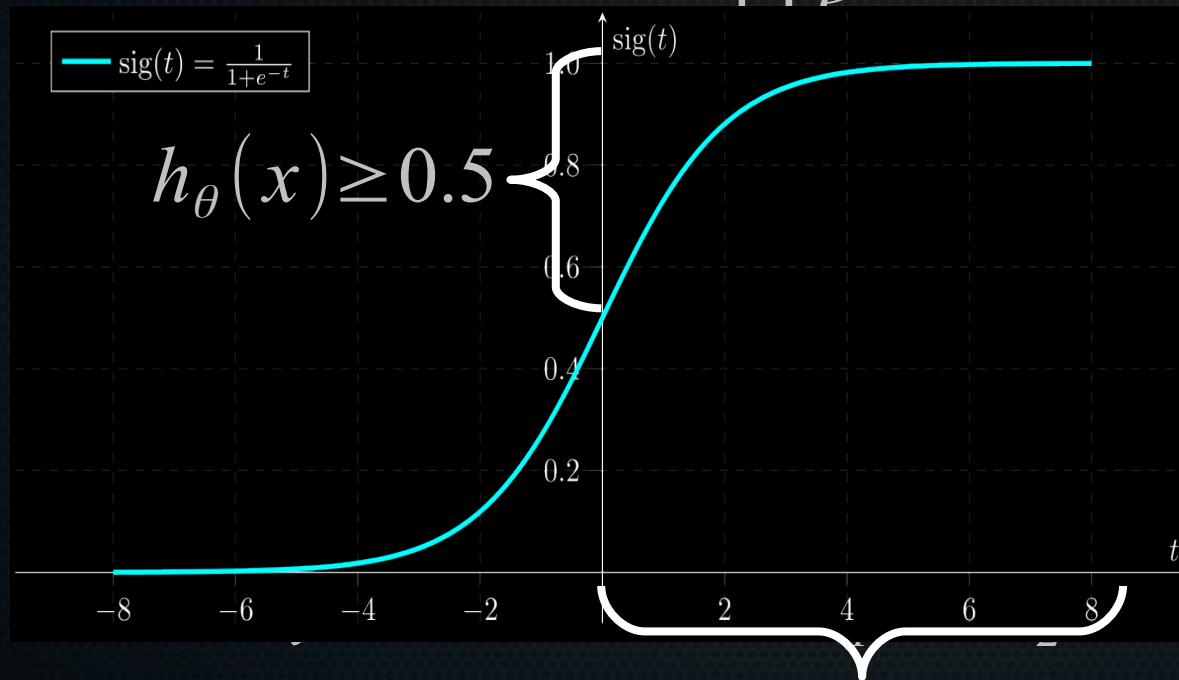
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
 $y = 1$ if $-3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$

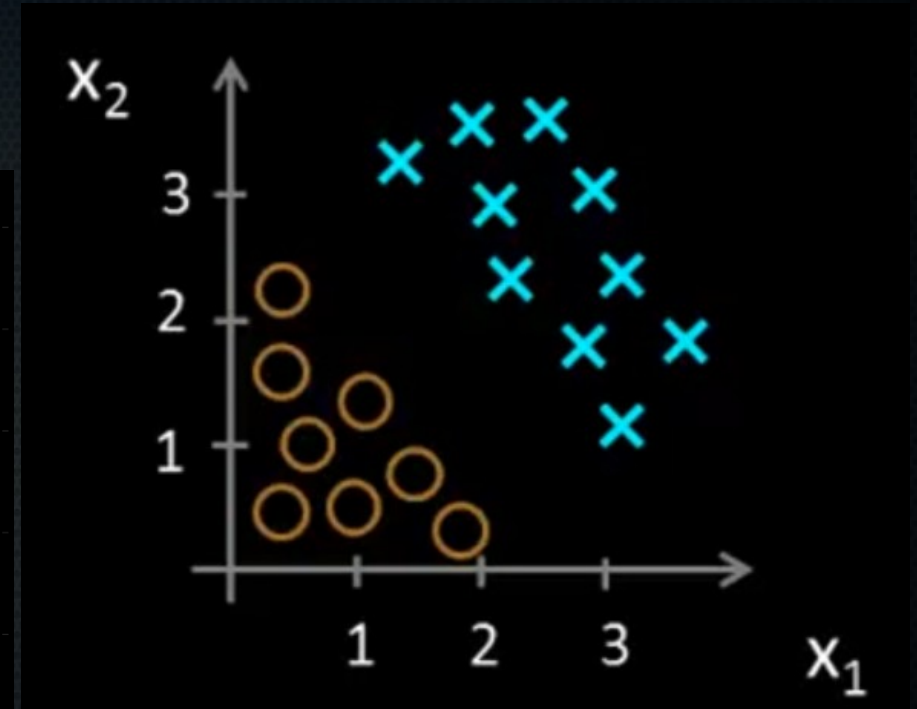


Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$

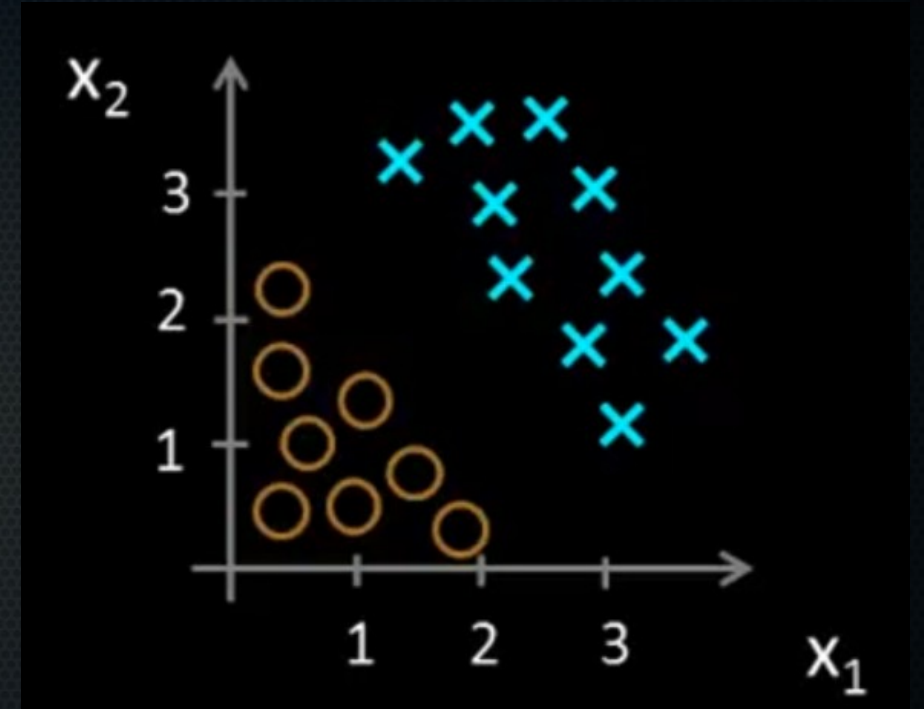


$$\theta^T \cdot x \geq 0$$



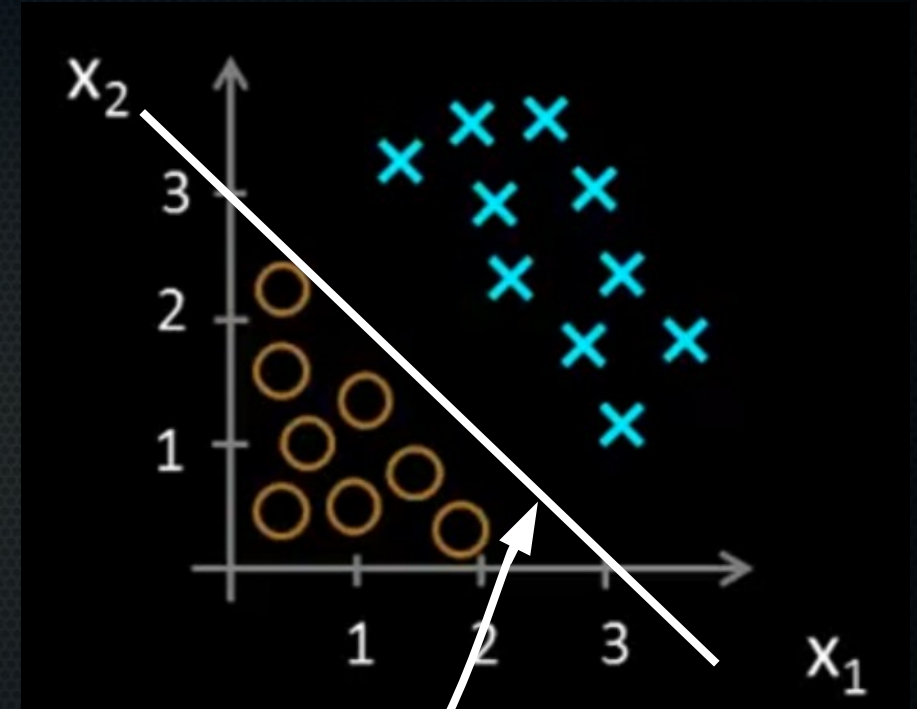
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
 $y = 1$ if $-3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$
 $-3 \cdot 1 + \cancel{1 \cdot x_1} + \cancel{1 \cdot x_2} \geq 0$
 $x_1 + x_2 = 3$



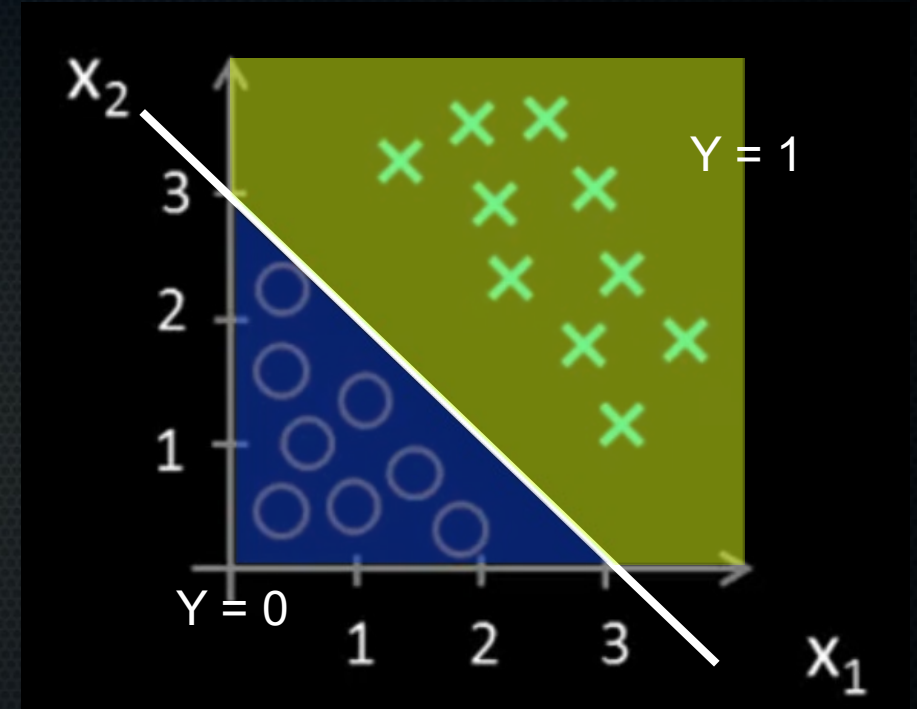
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
 $y = 1$ if $-3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$
 $-3 \cdot 1 + \cancel{1 \cdot x_1} + \cancel{1 \cdot x_2} \geq 0$
 $x_1 + x_2 = 3$



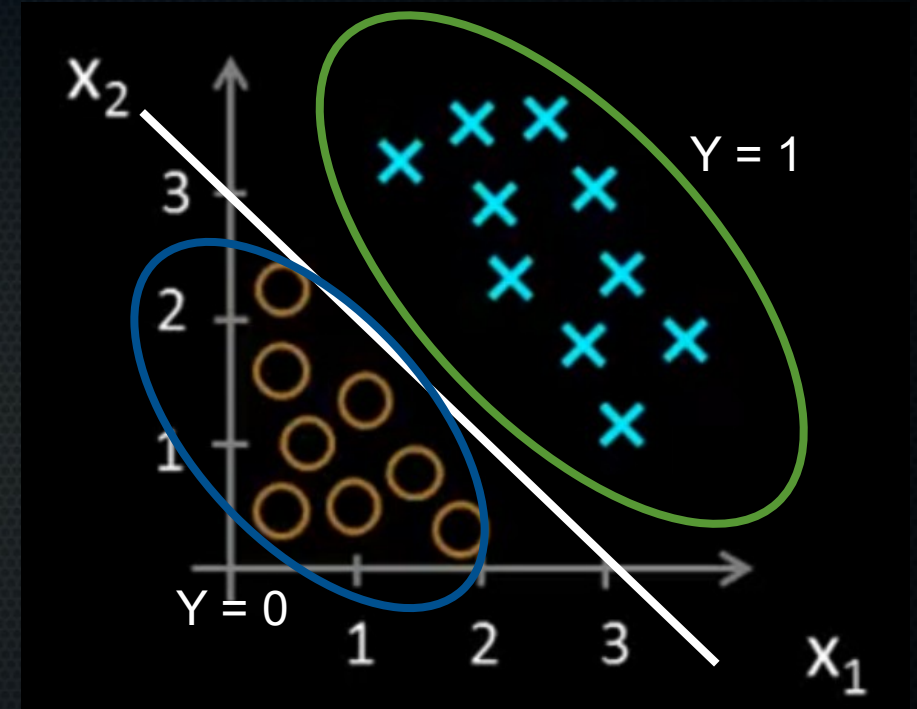
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
 $y = 1$ if $-3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$



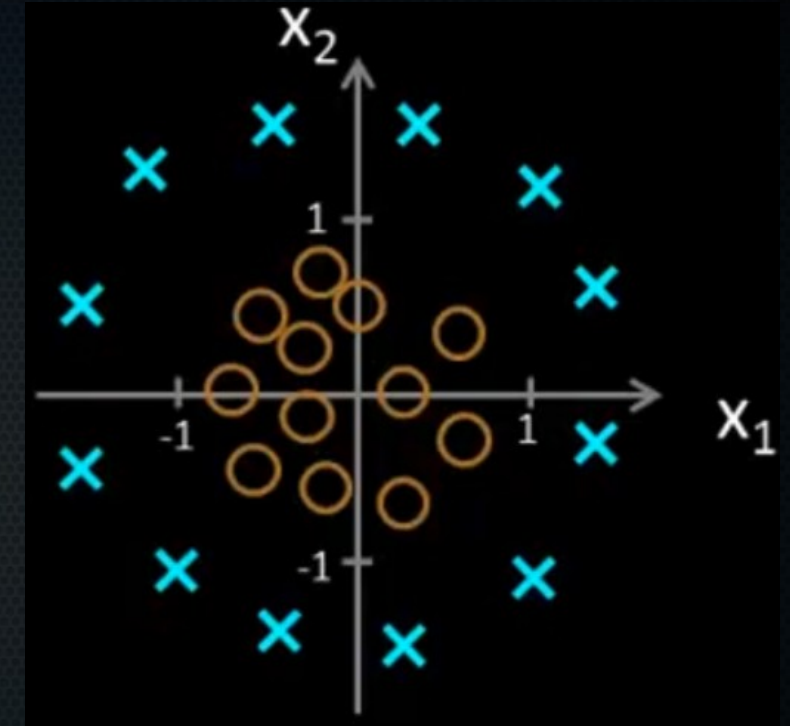
Decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$
 $y = 1$ if $-3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$



Decision boundary

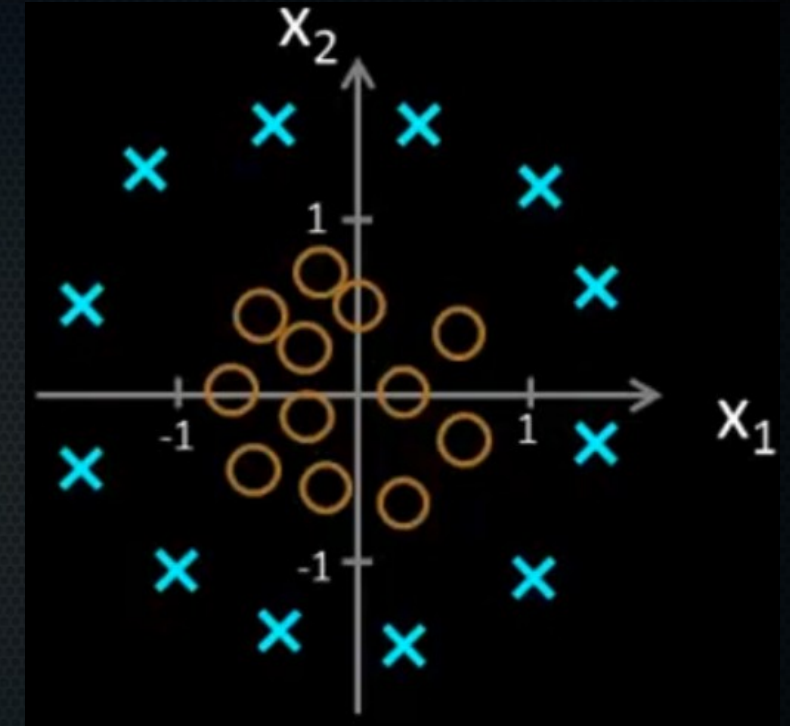
- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 \cdot x_0, \theta_1 \cdot x_1, \theta_2 \cdot x_2)$
 $\theta = \begin{bmatrix} -3 \\ 1 \\ 1 \end{bmatrix}$



$$y = 1 \text{ if } -3 \cdot 1 + 1 \cdot x_1 + 1 \cdot x_2 \geq 0$$

Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \theta_3 x_1^2, \theta_4 x_2^2)$
- Add two polynomial features



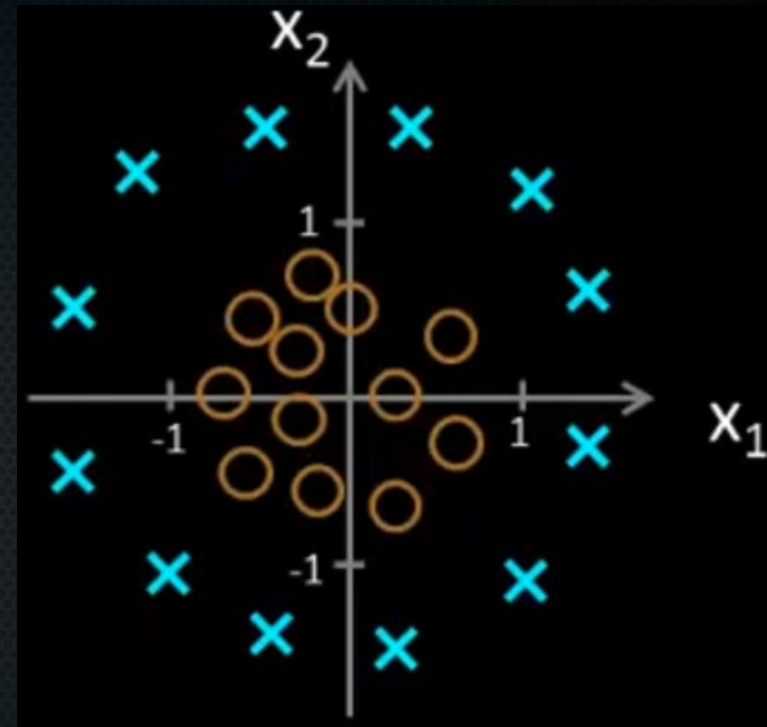
Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$

$$h_{\theta}(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \theta_3 x_1^2, \theta_4 x_2^2)$$

- Add two polynomial features

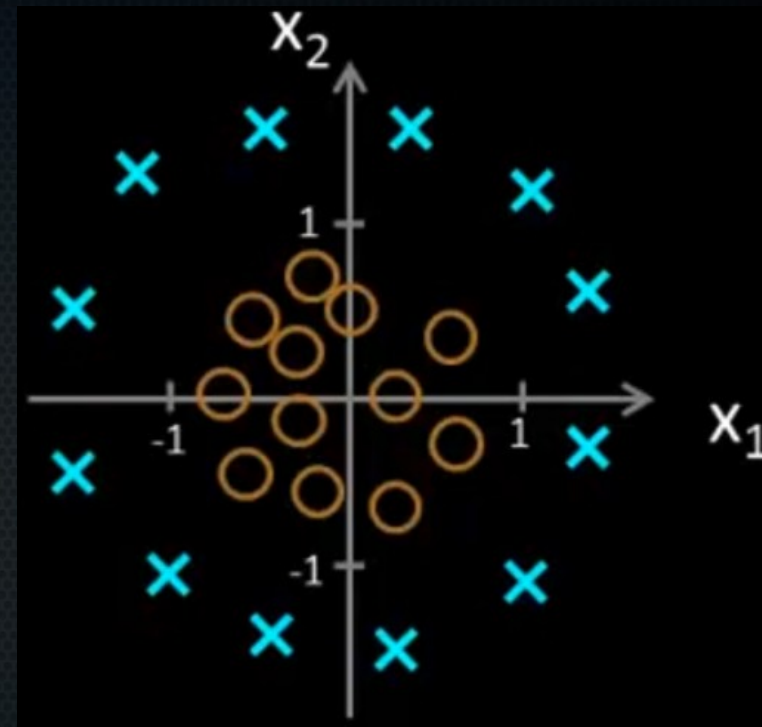
$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$



Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \theta_3 x_1^2, \theta_4 x_2^2)$
- Add two polynomial features

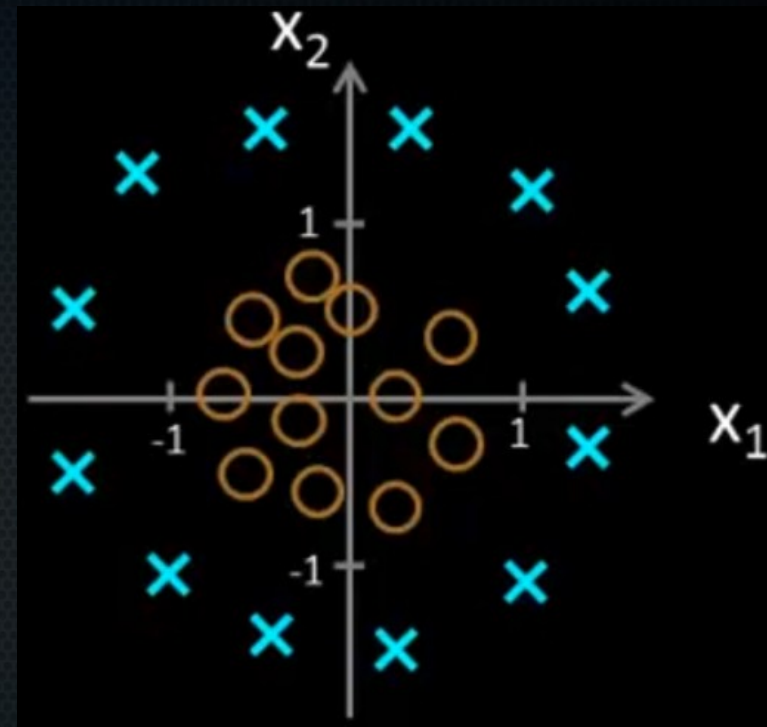
$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \longrightarrow -1 + x_1^2 + x_2^2 \geq 0$$



Non-linear decision boundary

- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \theta_3 x_1^2, \theta_4 x_2^2)$
- Add two polynomial features

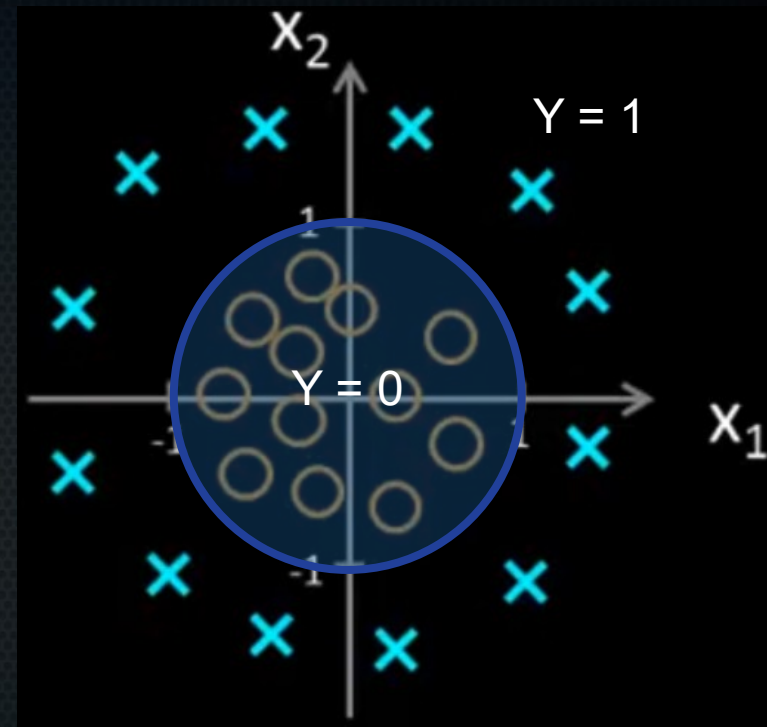
$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \longrightarrow -1 + x_1^2 + x_2^2 \geq 0$$
$$\downarrow$$
$$x_1^2 + x_2^2 \geq 1$$



Non-linear decision boundary

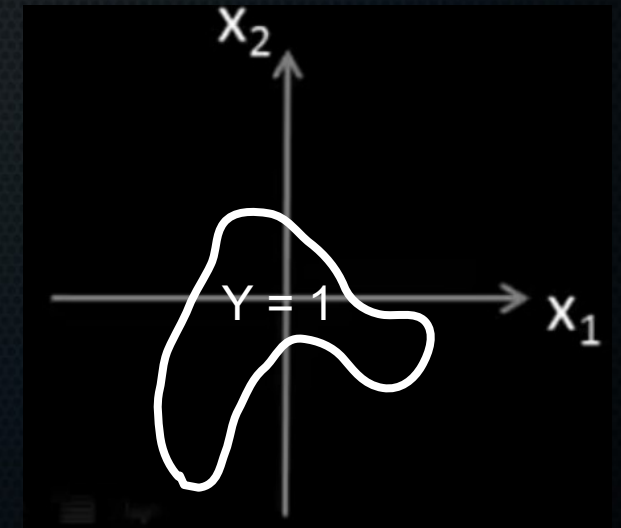
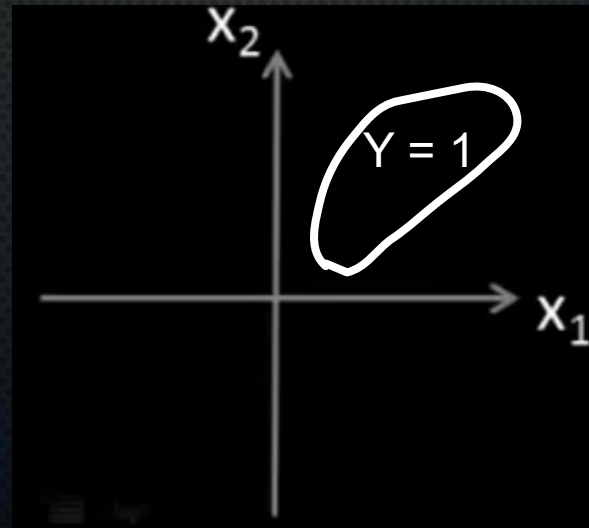
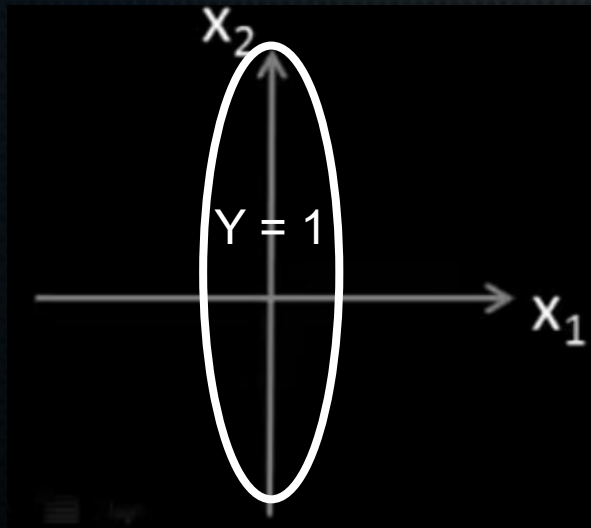
- How does it look? $g(z) = \frac{1}{1 + e^{-z}}$
 $h_{\theta}(x) = g(\theta_0 x_0, \theta_1 x_1, \theta_2 x_2, \theta_3 x_1^2, \theta_4 x_2^2)$
- Add two polynomial features

$$\theta = \begin{bmatrix} -1 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix} \longrightarrow -1 + x_1^2 + x_2^2 \geq 0$$
$$\downarrow$$
$$x_1^2 + x_2^2 \geq 1$$



Non-linear decision boundary

- How does it look?
- If you add more and higher-order polynomial features, you can get complex boundaries:



So how do we get theta's?

- Need a cost function

So how do we get theta's?

- Need a cost function

- Before: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

So how do we get theta's?

- Need a cost function

- Before: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

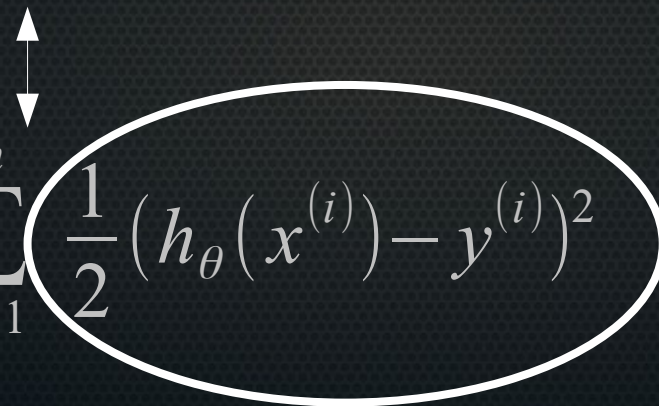


$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

So how do we get theta's?

- Need a cost function

- Before: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \left(\frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2 \right)$$


$$\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$$

So how do we get theta's?

- Need a cost function

- Before: $J(\theta) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$

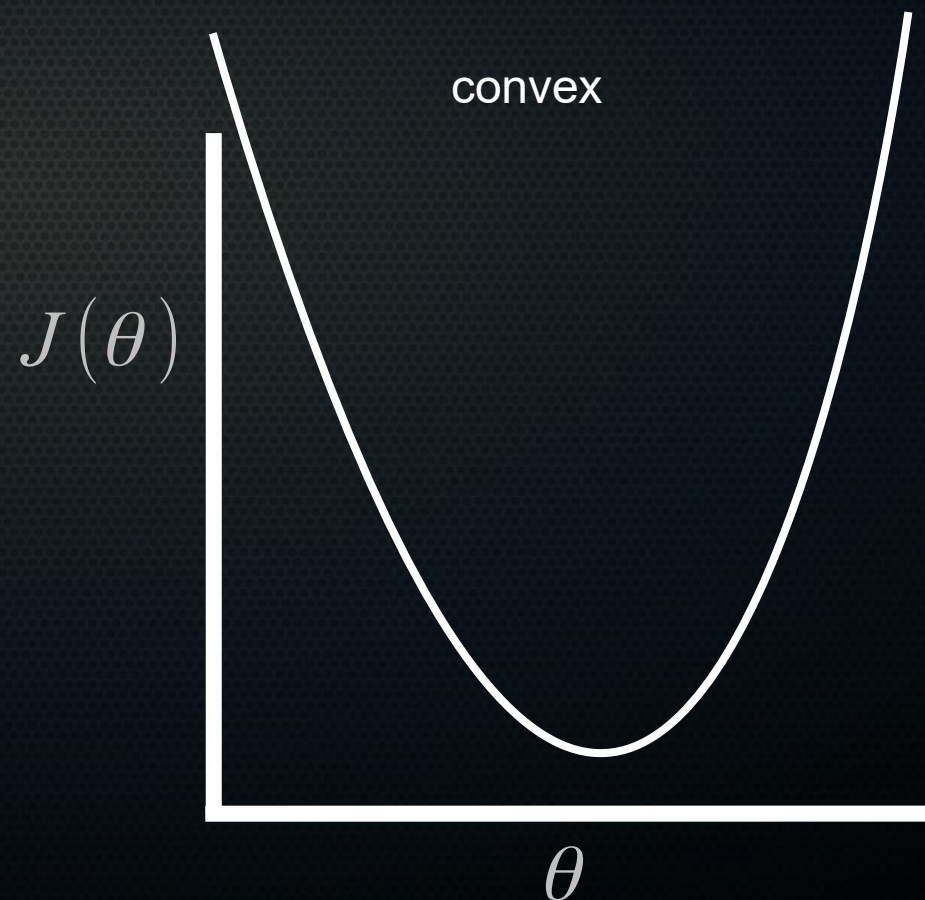
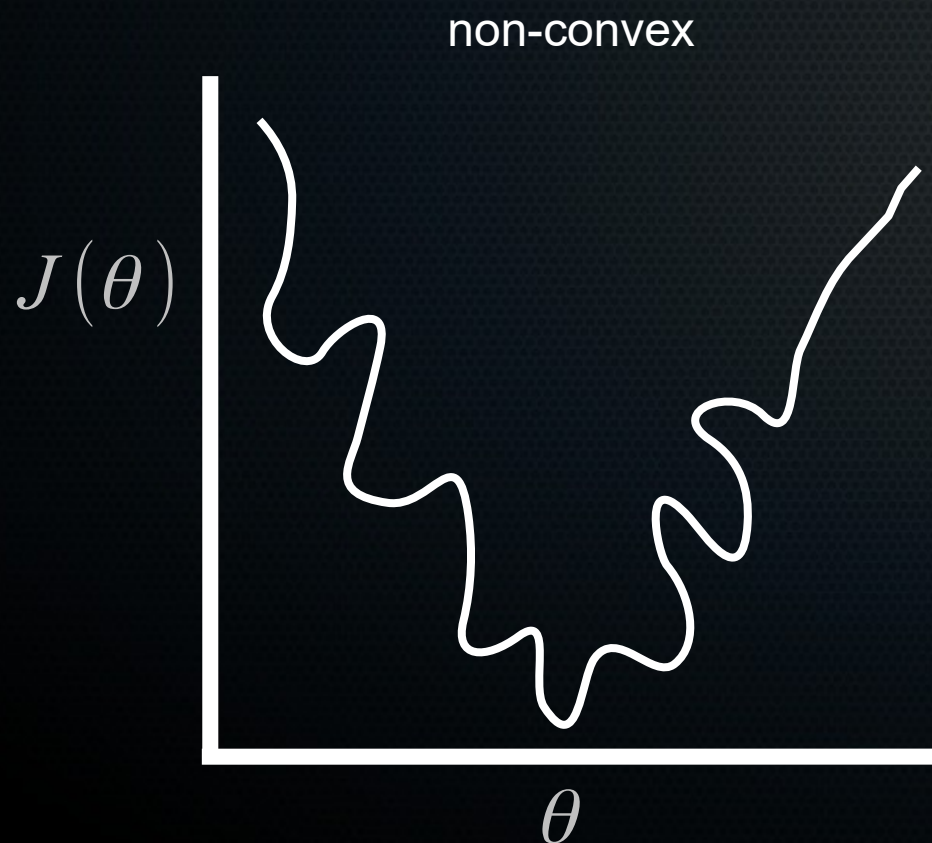
$$\begin{array}{l} \updownarrow \\ J(\theta) = \frac{1}{m} \sum_{i=1}^m \left\{ \frac{1}{2} (h_{\theta}(x^{(i)}) - y^{(i)})^2 \right\} \\ \text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2 \end{array} \left. \vphantom{\begin{array}{l} J(\theta) = \frac{1}{m} \sum_{i=1}^m \\ \text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2 \end{array}} \right\} J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^{(i)})$$

So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? → not convex

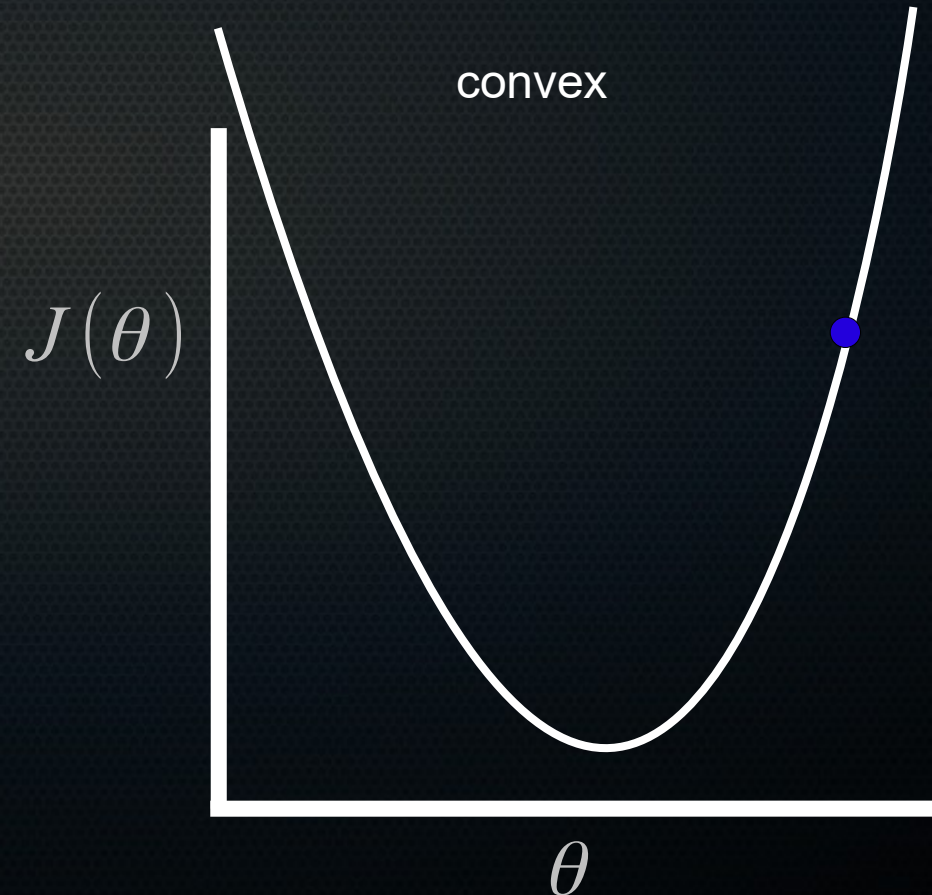
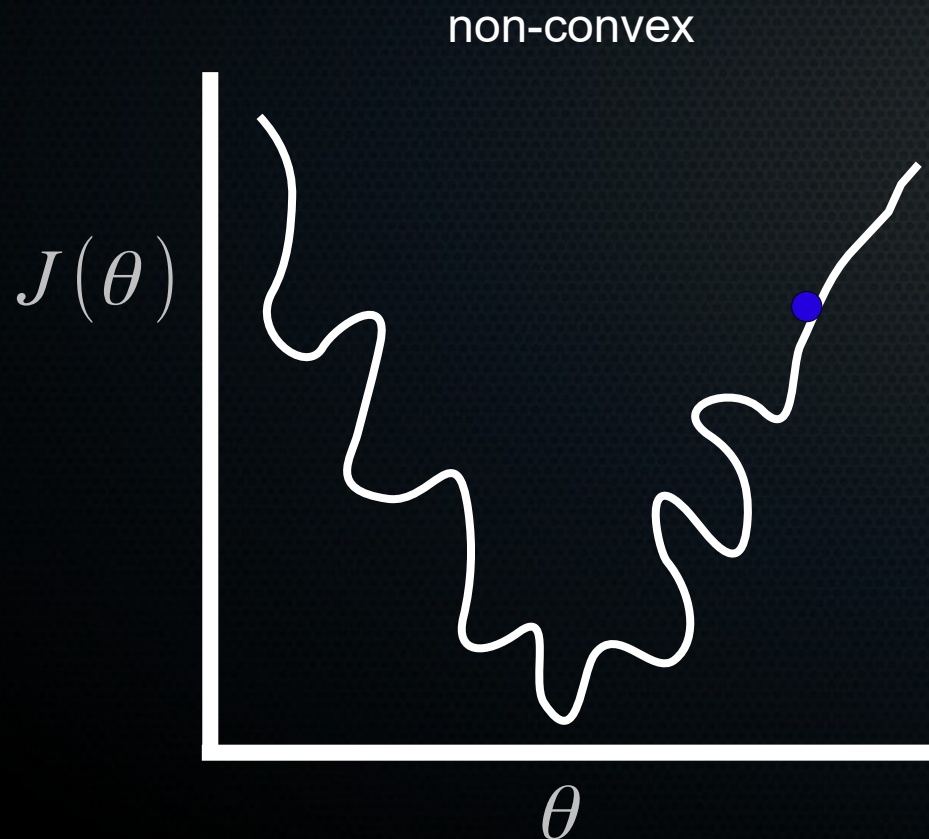
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? → not convex



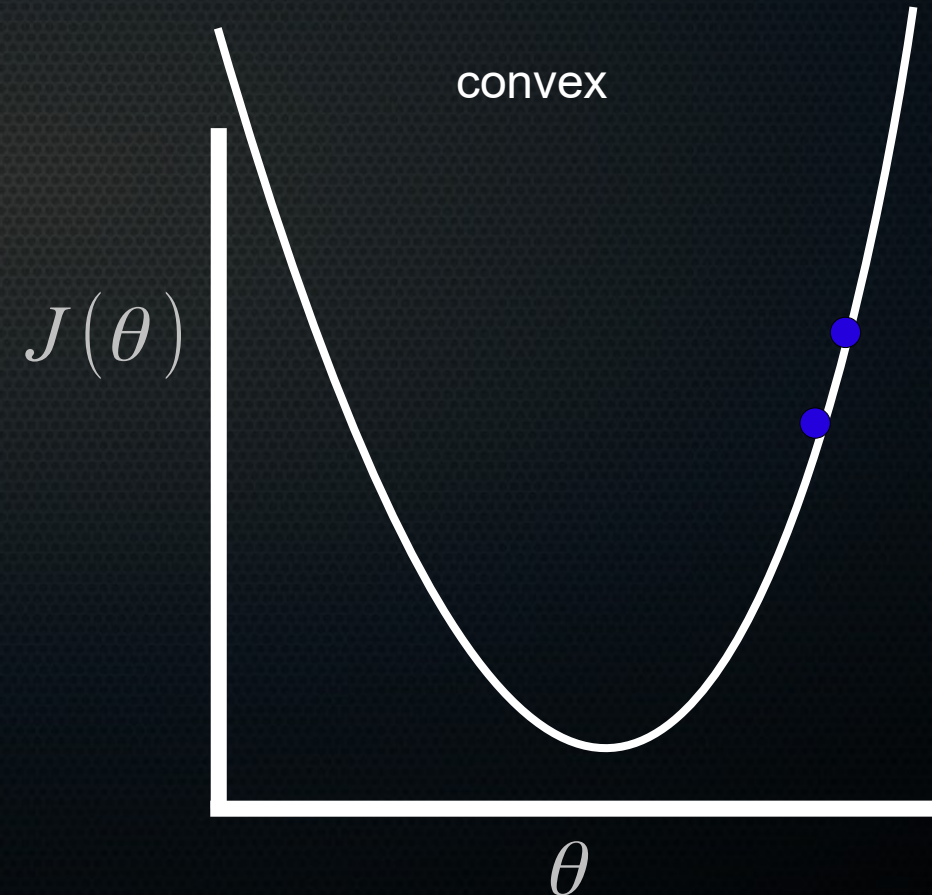
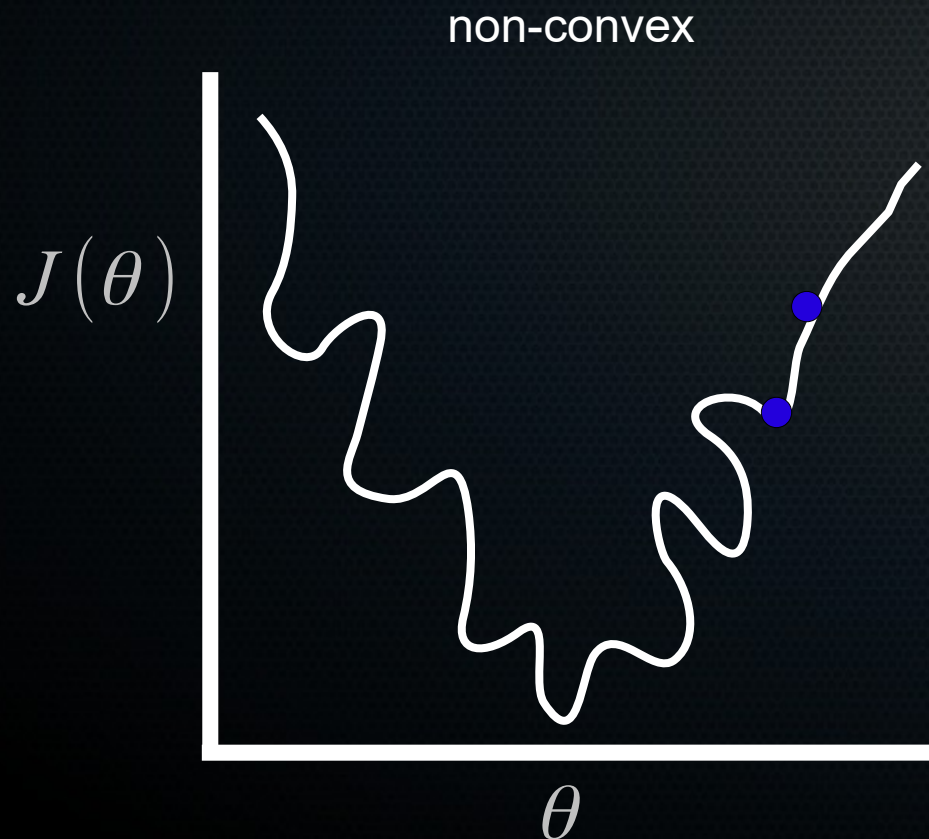
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? \rightarrow not convex



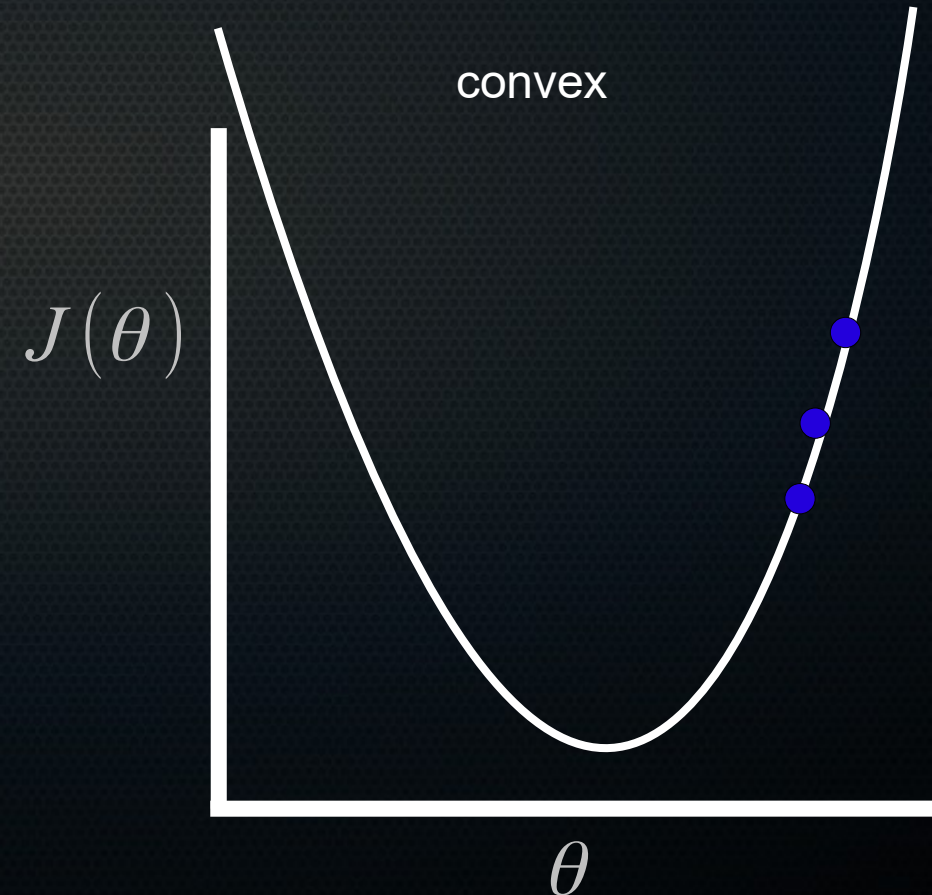
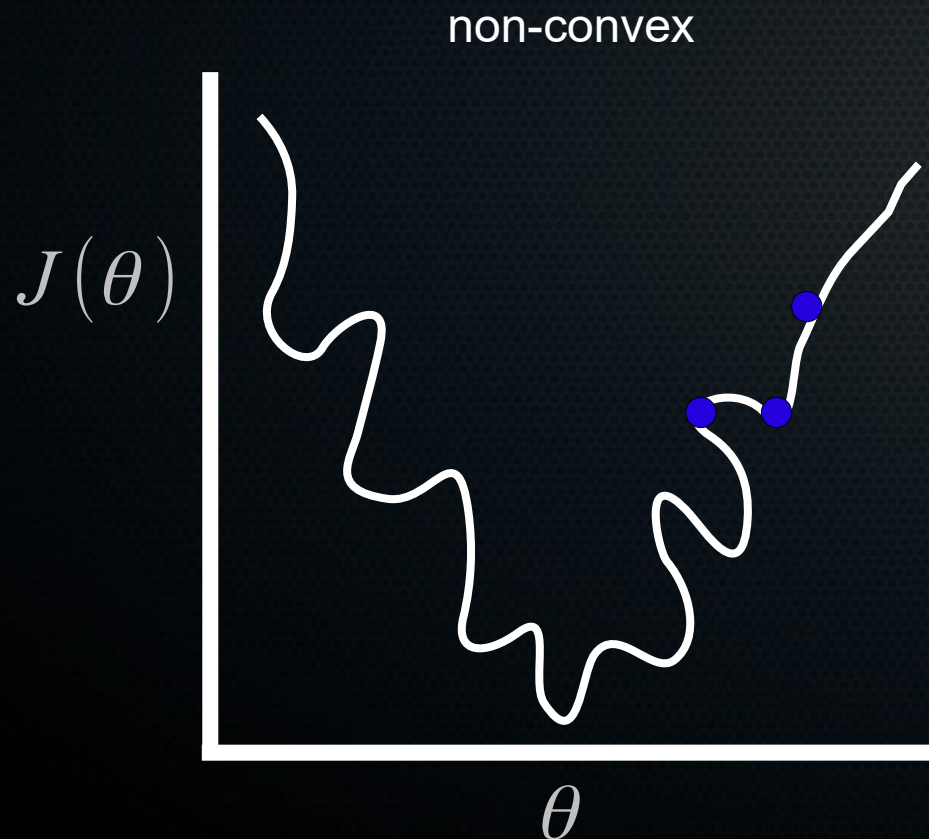
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? → not convex



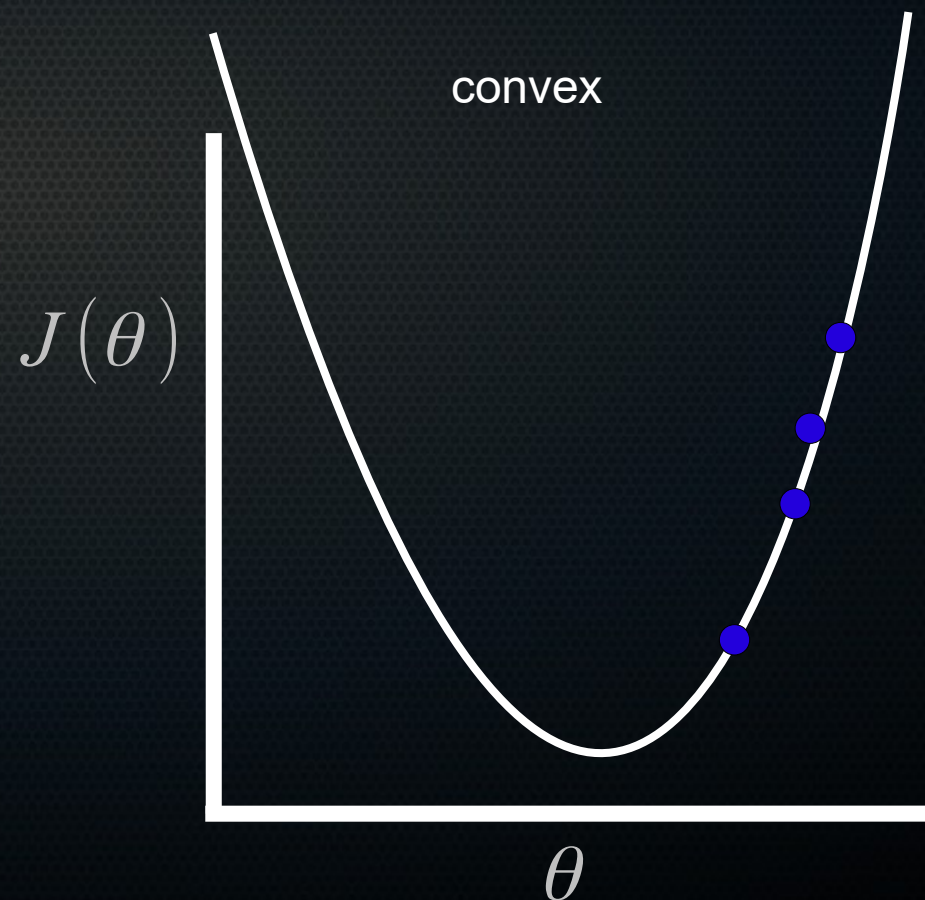
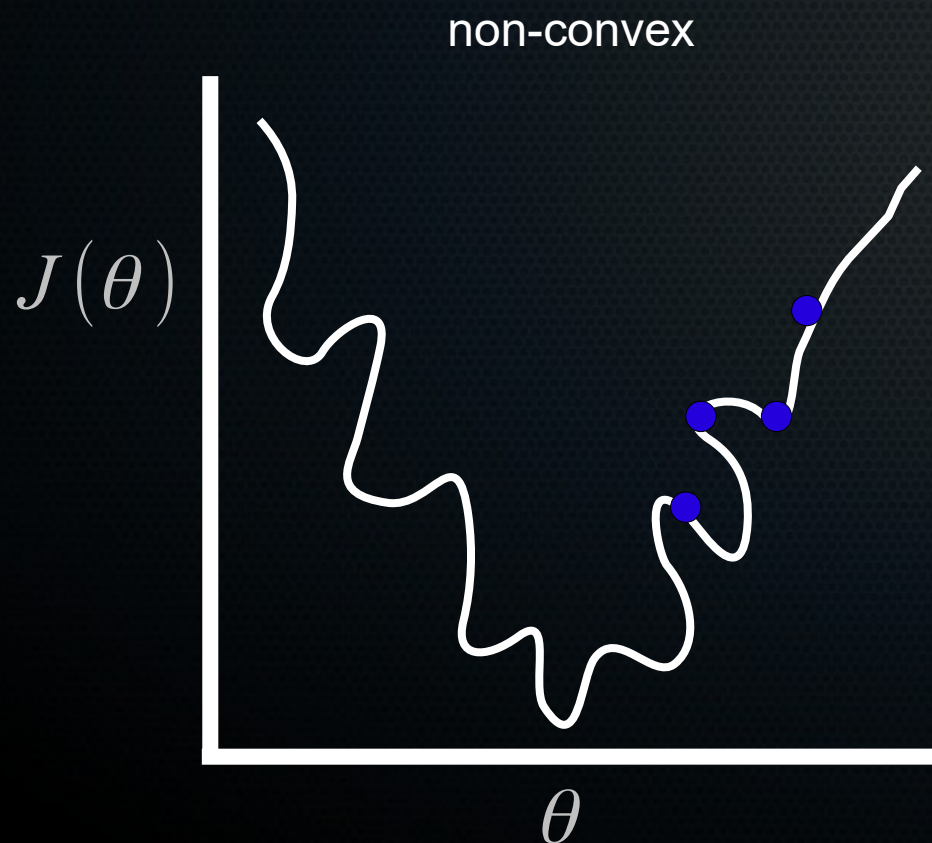
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? → not convex



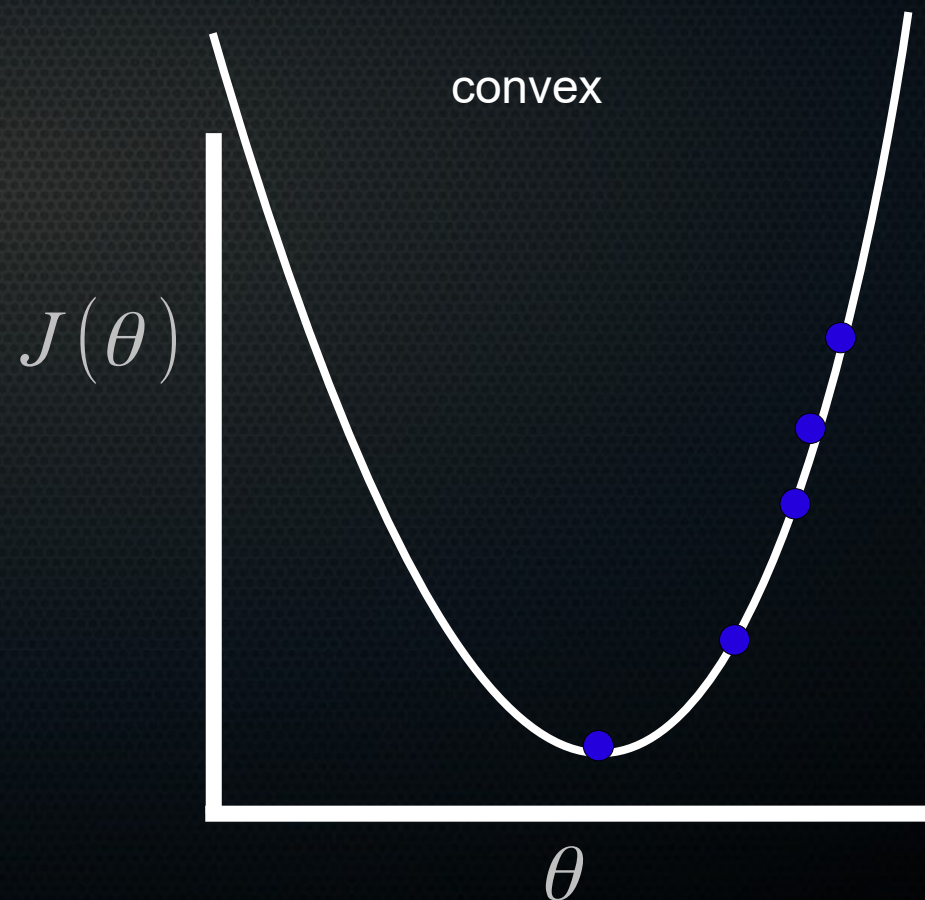
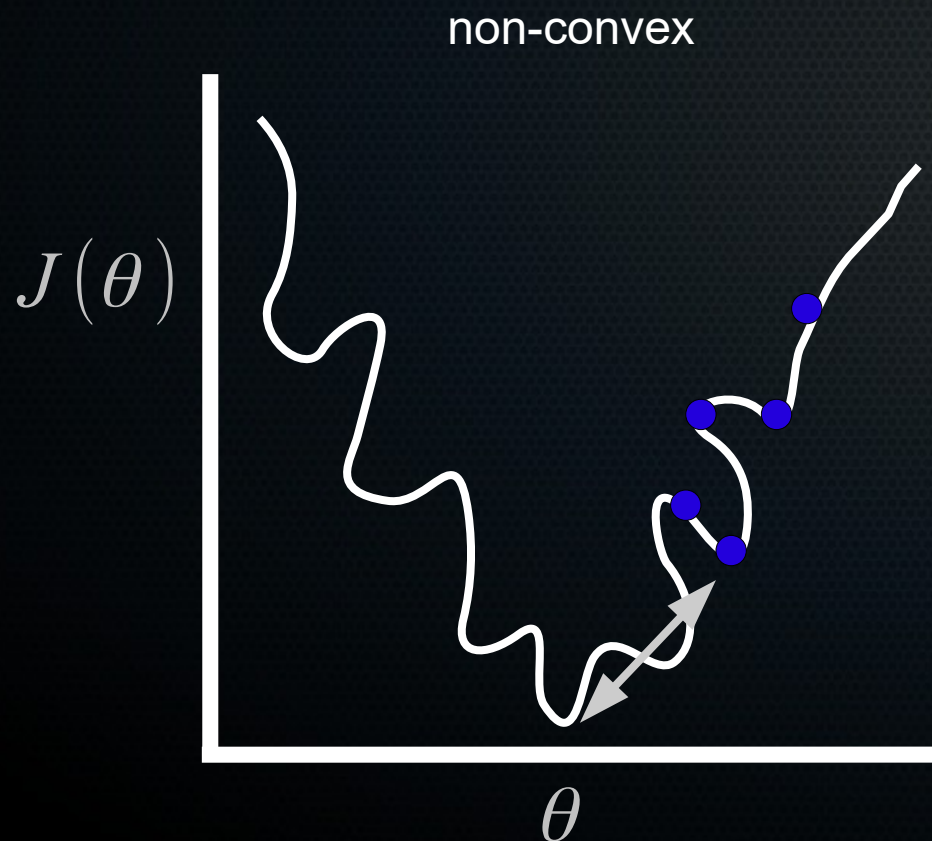
So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? \rightarrow not convex



So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$ $\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$
- Why not MSE? → not convex



So how do we get theta's?

- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$
- What then?

$$\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$$

So how do we get theta's?

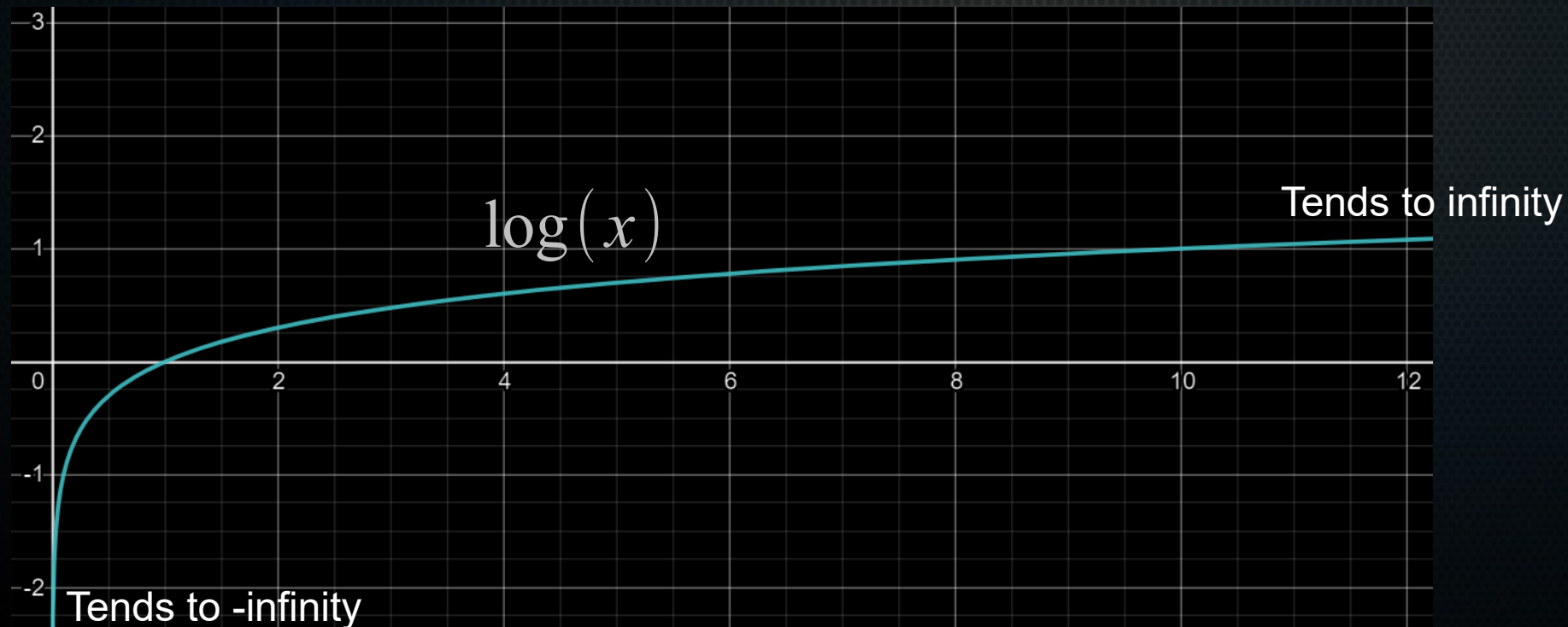
- Need a cost function $J(\theta) = \frac{1}{m} \sum_{i=1}^m \frac{1}{2} \text{Cost}(x^i)$ ~~$\text{Cost}(x) = \frac{1}{2} (h_{\theta}(x) - y)^2$~~
- What then?

$$\text{Cost}(x) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1-h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

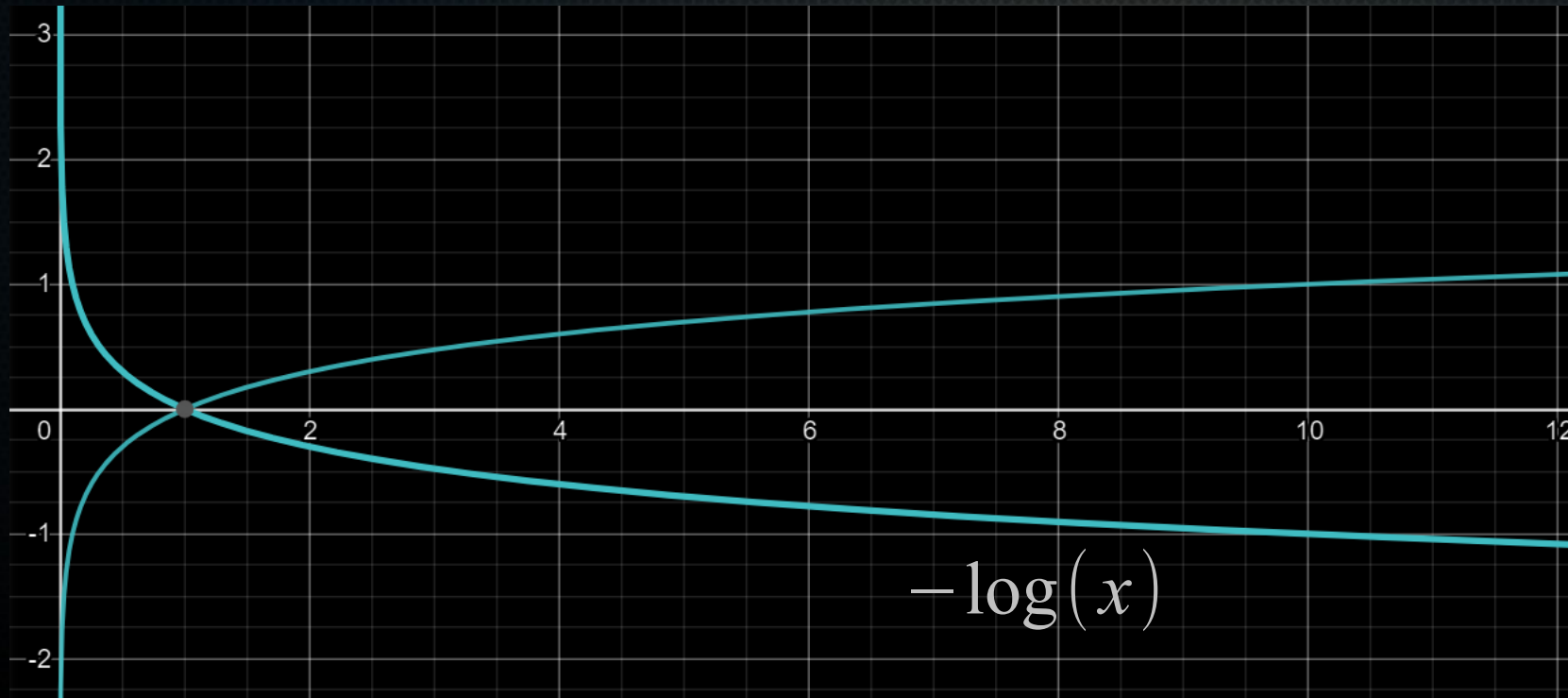
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1-h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

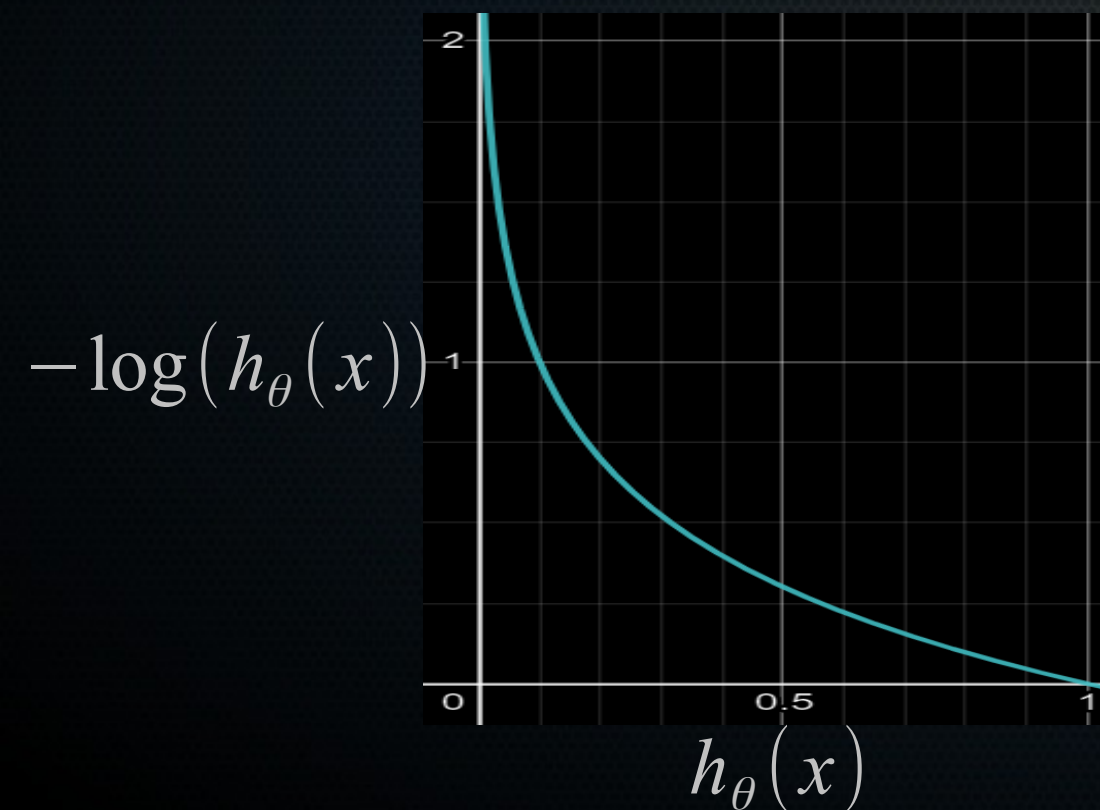
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1-h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

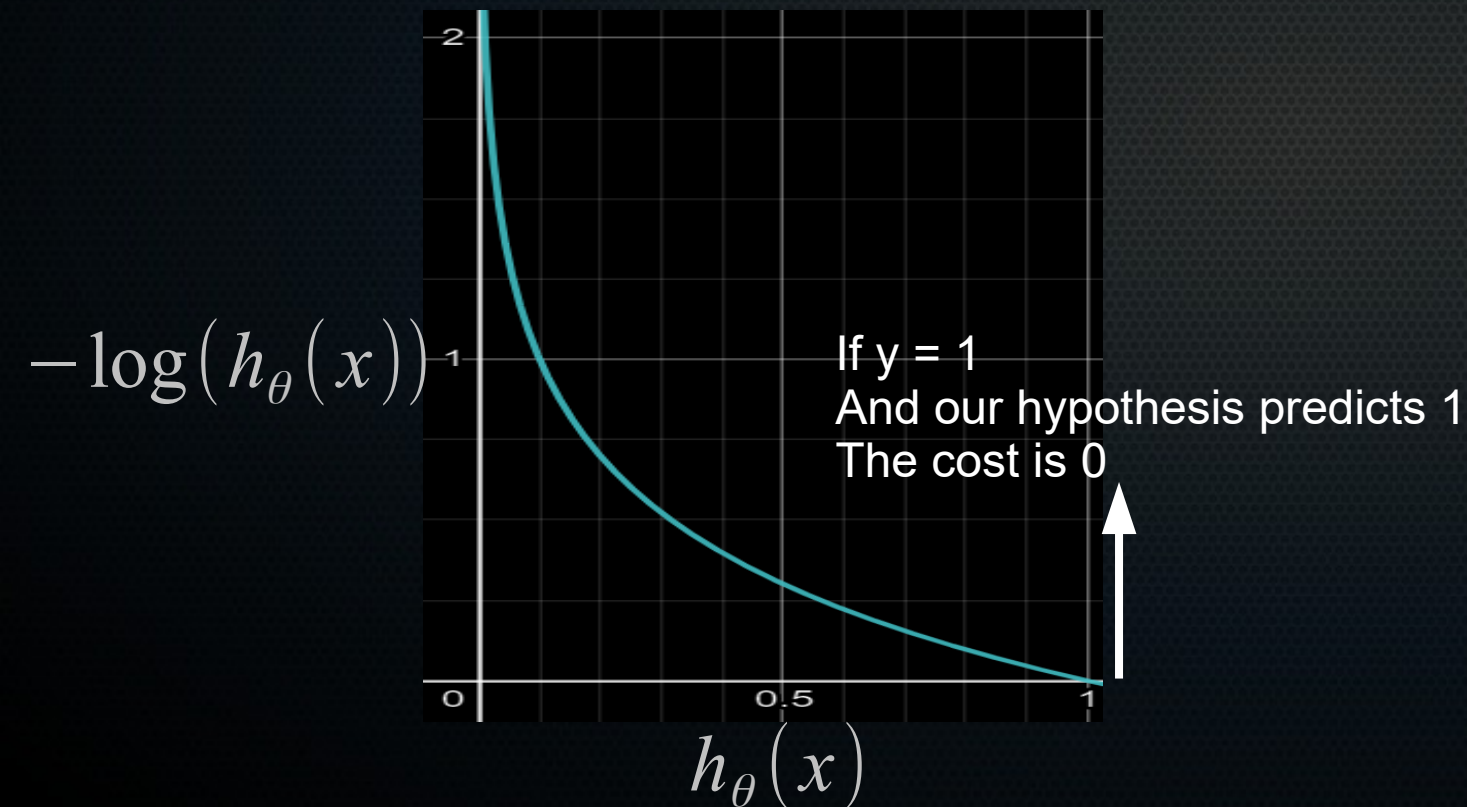
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

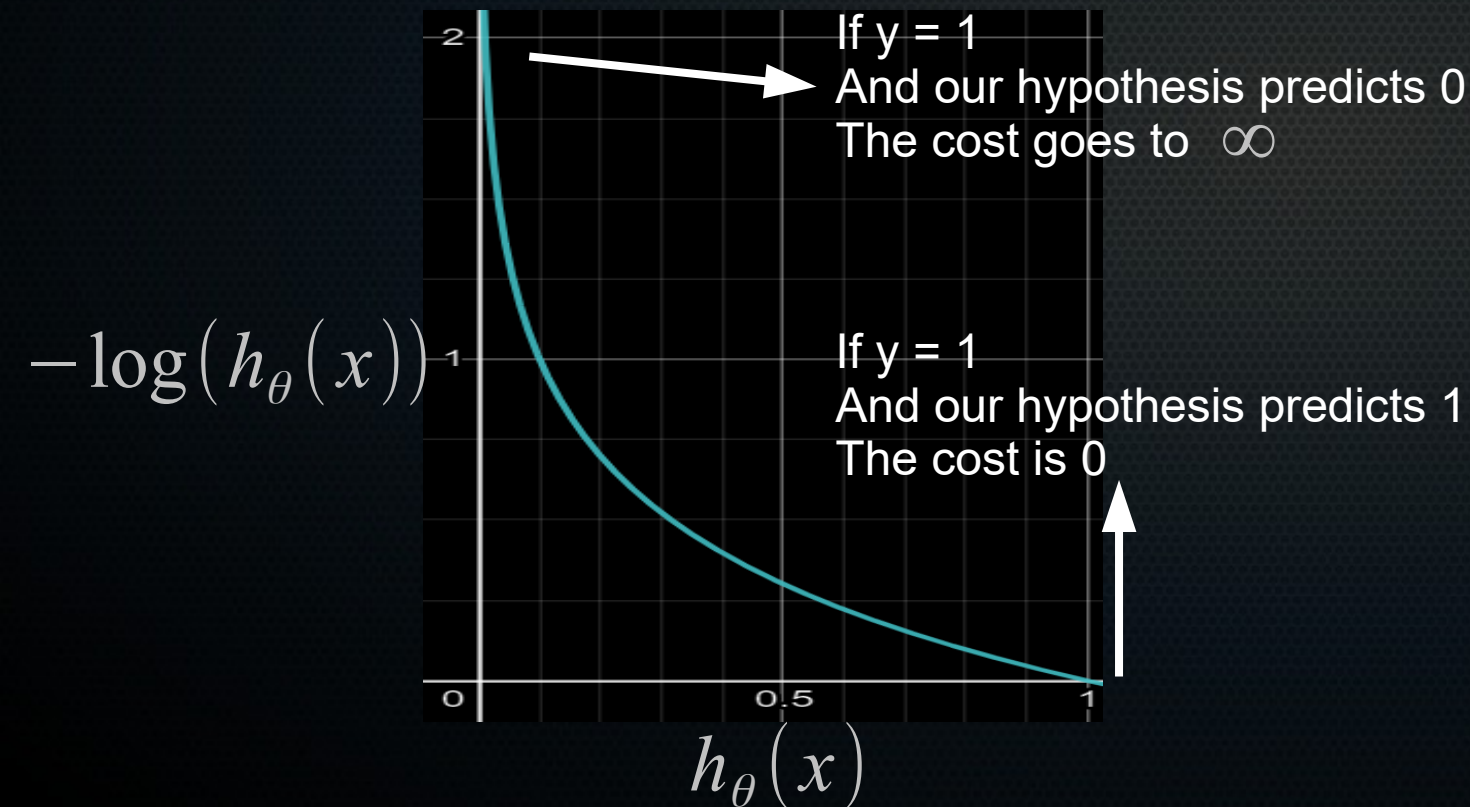
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

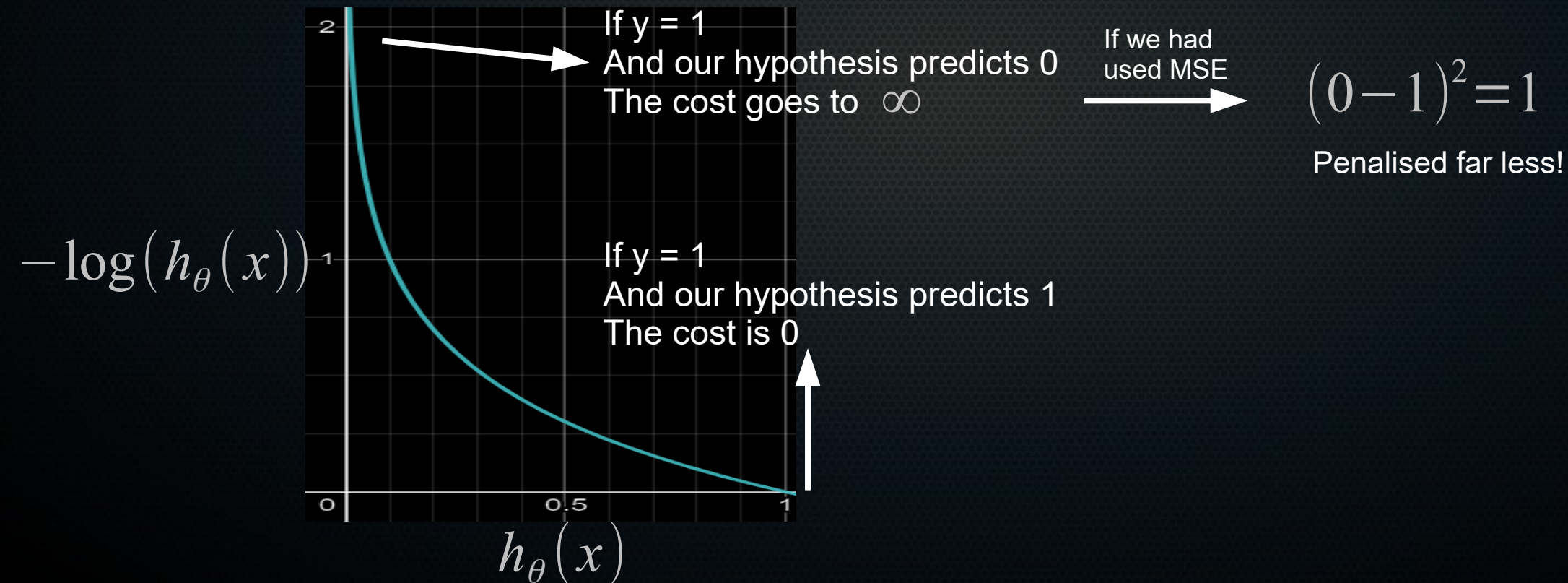
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y=1 \\ -\log(1-h_{\theta}(x)) & \text{if } y=0 \end{cases}$$

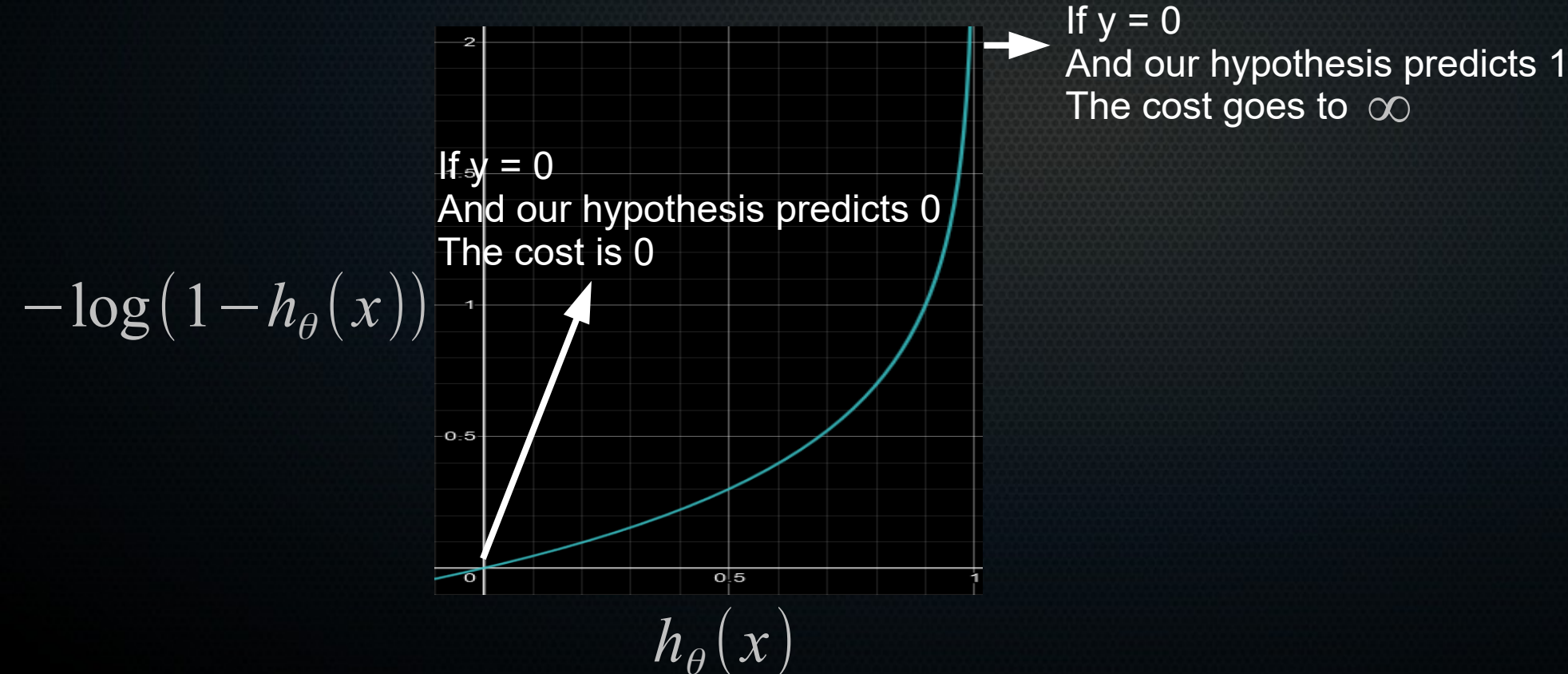
$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



What does this function look like?

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



Simplified notation

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$



$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$

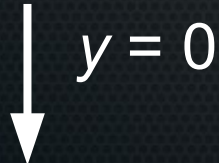
Simplified notation

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$



$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$



$$-0 \cdot \log(h_\theta(x)) - (1-0) \cdot \log(1-h_\theta(x))$$

Simplified notation

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$

$$y = 0$$

$$-0 \cdot \log(h_\theta(x)) - (1-0) \cdot \log(1-h_\theta(x))$$

$$-\log(1-h_\theta(x))$$

Simplified notation

$$\text{Cost}(x) = \begin{cases} -\log(h_\theta(x)) & \text{if } y=1 \\ -\log(1-h_\theta(x)) & \text{if } y=0 \end{cases}$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x))$$

$$y = 1$$

$$-1 \cdot \log(h_\theta(x)) - (1-1) \cdot \log(1-h_\theta(x))$$

$$-\log(h_\theta(x))$$

Putting it all together

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x)) \quad J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^{(i)})$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m -y^{(i)} \cdot \log(h_\theta(x^{(i)})) - (1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)}))$$

Putting it all together

$$\text{Cost}(x) = -y \cdot \log(h_\theta(x)) - (1-y) \cdot \log(1-h_\theta(x)) \quad J(\theta) = \frac{1}{m} \sum_{i=1}^m \text{Cost}(x^i)$$

$$J(\theta) = \frac{1}{m} \sum_{i=1}^m \left[-y^{(i)} \cdot \log(h_\theta(x^{(i)})) - (1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)})) \right]$$



$$J(\theta) = -\frac{1}{m} \sum_{i=1}^m \left[y^{(i)} \cdot \log(h_\theta(x^{(i)})) + (1-y^{(i)}) \cdot \log(1-h_\theta(x^{(i)})) \right]$$

Optimising the cost function

- Same form as for linear regression (only hypothesis function differs!)

$$\frac{\partial}{\partial \theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^m ((h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)})$$

$$\theta_j := \theta_j - \frac{\alpha}{m} \sum_{i=1}^m ((h_{\theta}(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)})$$

$$h_{\theta}(x^{(i)}) = \frac{1}{1 + e^{-\theta^T * x}}$$

Summary

- By using the sigmoid function as a transformation of normal regression and interpreting the output as a chance of being 0 or 1 we can do classification.
- Only the form of our hypothesis function is different
- Need a different cost function: should be smooth, and give logical values for large errors.

Break for practical
