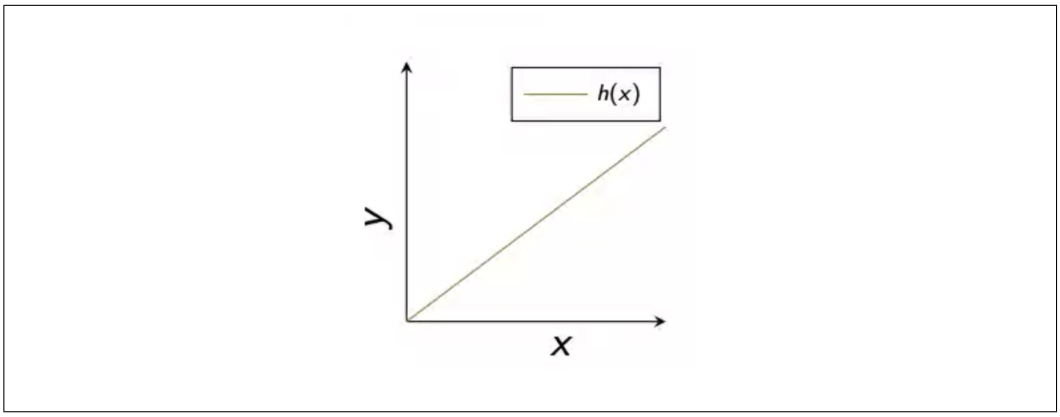
# 1 Linear Regression Contd.

$\theta_i$'s: Parameters

How to choose $\theta_i$'s?

Hypothesis: $h_\theta(x) = \theta_1 x$

Parameters: $\theta_1$

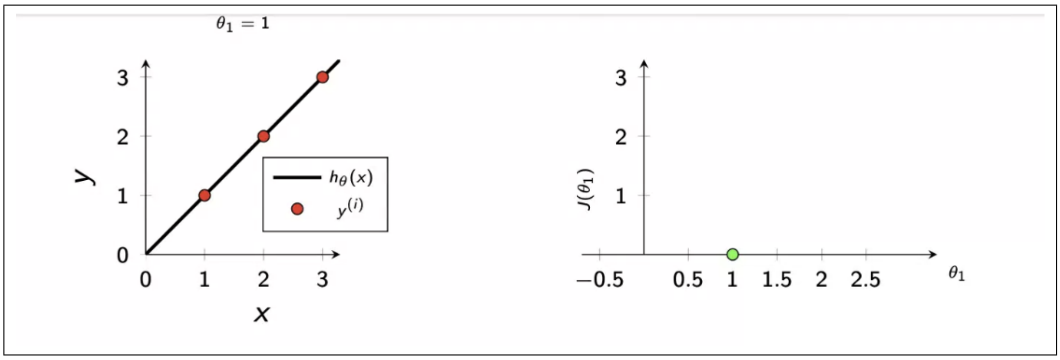



Cost function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left( h_\theta\left(x^{(i)}\right) - y^{(i)} \right)^2$$

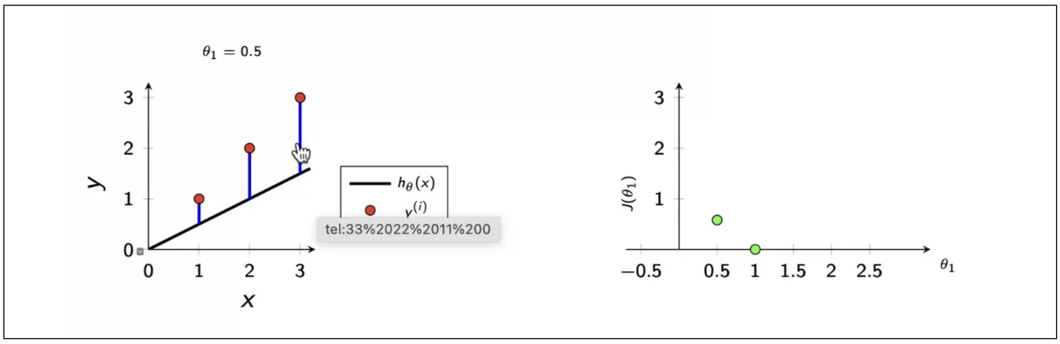
Goal: $\min_{\theta_1} J(\theta_1)$

$h_\theta(x)$: for fixed $\theta_1$, this is a function of $x$, and $J(\theta_1)$: function of the parameter $\theta_1$.



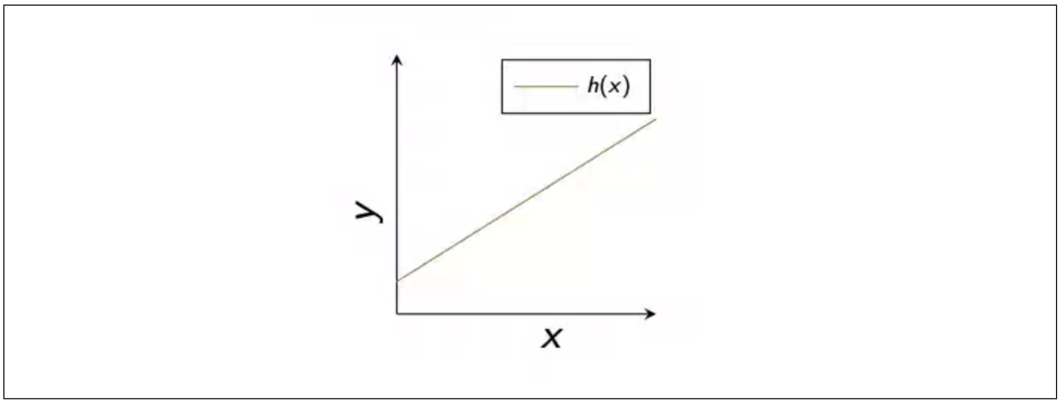
$$
\begin{aligned}
J\left(\theta_1\right) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 \\
&= \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_1 x^{(i)} - y^{(i)}\right)^2 \\
&= \frac{1}{2m} \sum_{i=1}^{m} \left(0^2 + 0^2 + 0^2\right) = 0
\end{aligned}
$$

$h_\theta(x)$: for fixed $\theta_1$, this is a function of $x$, and $J(\theta_1)$: function of the parameter $\theta_1$.




Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: $\theta_0, \theta_1$ and Goal: $\min_{\theta_0,\theta_1} J\left(\theta_0, \theta_1\right)$




Cost function:

$$
J\left(\theta_0, \theta_1\right) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2
$$

## 1.1 Gradient descent approach

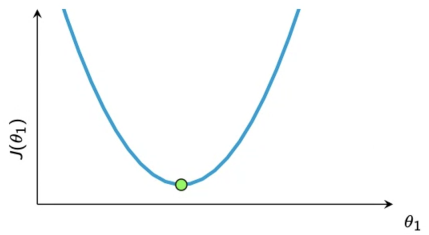
We have some function $J(\theta_0, \theta_1)$

and we want to solve $\min_{\theta_1} J(\theta_0, \theta_1)$

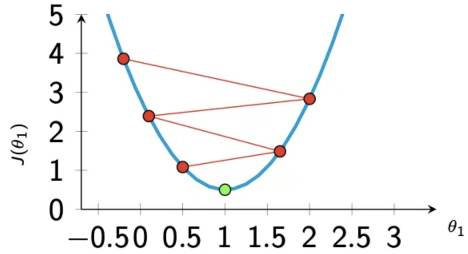
- Start with some $\theta_0, \theta_1$
- Keep changing $\theta_0, \theta_1$ to reduce $J(\theta_0, \theta_1)$ until we hopefully end up at a minimum.

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If $\alpha$ is too small, gradient descent can be slow.




If alpha is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.




Gradient descent can converge to a local minimum, even with the learning rate $\alpha$ fixed.

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

As we approach a local minimum, gradient descent will automatically take smaller steps. So, no need to decrease $\alpha$ over time.

**Gradient Descent Algorithm** (repeat until convergence);

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J\left(\theta_0, \theta_1\right)$$

(for $j = 0$ and $j = 1$)

**Linear Regression Model**

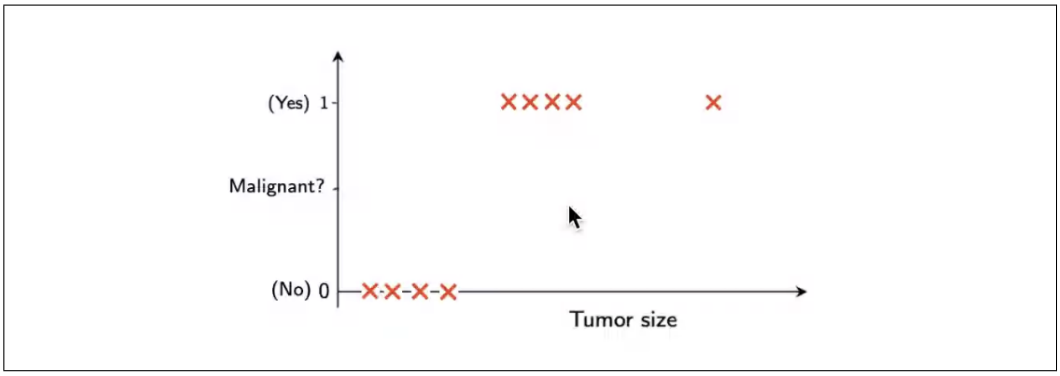
$$h_\theta(x) = \theta_0 + \theta_1 x$$
$$J\left(\theta_0, \theta_1\right) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

### 1.1.1 Two Variable Model

$$\frac{\partial}{\partial \theta_j} J\left(\theta_0, \theta_1\right) = \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$
$$= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)^2$$
$$j = 0 : \frac{\partial}{\partial \theta_0} J\left(\theta_0, \theta_1\right) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$
$$j = 1 : \frac{\partial}{\partial \theta_1} J\left(\theta_0, \theta_1\right) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) \cdot x^{(i)}$$

Classification

- Email: Spam/Not spam?
- Online Transactions: Fraudulent (Yes/No)?
- Tumor: Malignant / Benign?
- $y \in \{0, 1\}$, where
  - 0: “Negative class” (e.g., benign tumor) and
  - 1: “Positive class” (e.g., malignant tumor).




Threshold classifier output $h_\theta(x)$ at 0.5:

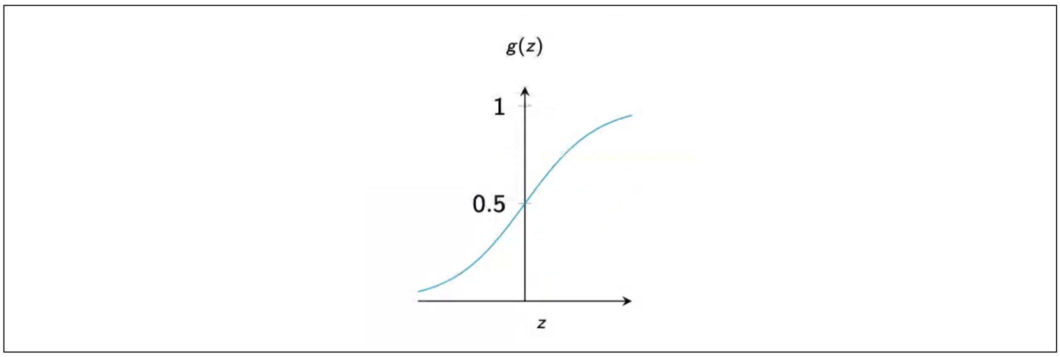
- If $h_\theta(x) \geq 0.5$, predict "$y = 1$"
- If $h_\theta(x) < 0.5$, predict "$y = 0$"

# 2 Logistic Regression Model

We want $0 \leq h_\theta(x) \leq 1$

$$h_\theta(x) = g\left(\theta^T x\right) = \frac{1}{1 + e^{-\theta^T x}}$$

where $g(z) = \frac{1}{1+e^{-z}}$ and it is known as Sigmoid function/Logistic function.




**Interpretation of Hypothesis Output** $h_\theta(x) =$ estimated probability that $y = 1$ on input $x$.

Example: If $x = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumor size} \end{bmatrix}$

$h_\theta(x) = 0.7$: tell patient that 70% chance of tumor being malignant.

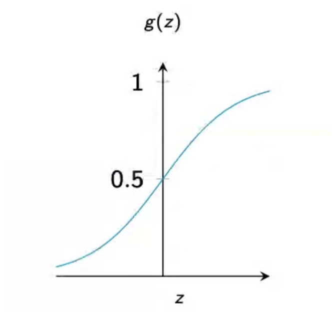
$h_\theta(x)$ gives probability that $y = 1$, given $x$, parameteized by $\theta$.

## 2.1 Logistic Regression: Decision Boundary

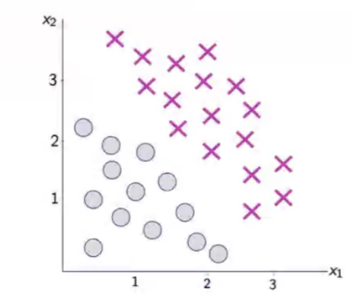
**Logistic Regression**

$$h_\theta(x) = g\left(\theta^T x\right) \quad ; \quad g(z) = \frac{1}{1 + e^{-z}}$$

- Suppose predict "$y = 1$" if $h_\theta(x) \geq 0.5$, which happens when $\theta^T x \geq 0$.
- Suppose predict "$y = 0$" if $h_\theta(x) < 0.5$, which happens when $\theta^T x < 0$.

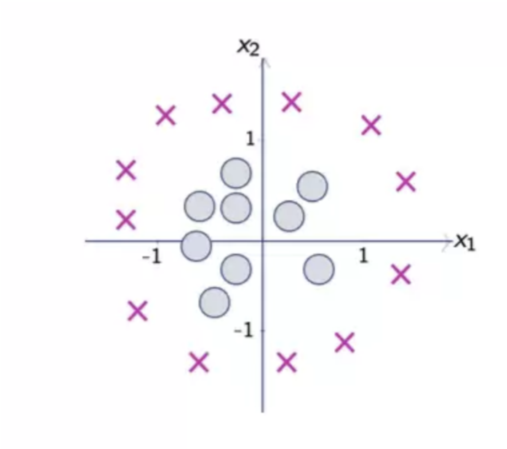



### 2.1.1 Decision Boundary




- $h_\theta(x) = g\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2\right)$
- Predict "$y = 1$" if $-3 + x_1 + x_2 \geq 0$

### 2.1.2 Non-Linear Decision Boundaries




- $h_\theta(x) = g\left(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2\right)$
- Predict "$y = 1$" if $-1 + x_1^2 + x_2^2 \geq 0$

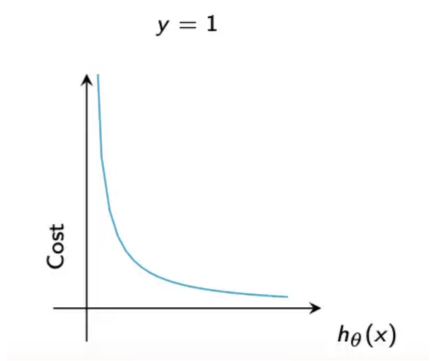
## 2.2 Logistic regression: Cost Function

- Training set: $\left\{\left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \ldots, \left(x^{(m)}, y^{(m)}\right)\right\}$
- m examples
- $x = \begin{bmatrix} x_0 \\ x_1 \\ \cdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}, x_0 = 1, y \in \{0, 1\}$
- $h_\theta(x) = \frac{1}{1+e^{-\theta^\top x}}$
- How to choose parameters $\theta$?

**Logistic regression cost function**

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

$$\operatorname{Cost}\left(h_\theta(x), y\right) = -y \log\left(h_\theta(x)\right) - (1 - y) \log\left(1 - h_\theta(x)\right)$$




- Cost $= 0$ if $y = 1, h_\theta(x) = 1$. But as $h_\theta(x) \to 0$, we have Cost $\to \infty$.
- Captures intuition that if $h_\theta(x) = 0$, (predict $\Pr(y = 1 \mid x; \theta) = 0$), but $y = 1$, we'll penalize learning algorithm by a very large cost.