# Supervised Learning

$$\underset{\text{(input)}}{X} \longrightarrow \underset{\text{(output)}}{Y}$$

Learn Hypothesis: $h(x) \approx Y$

**Terminology**

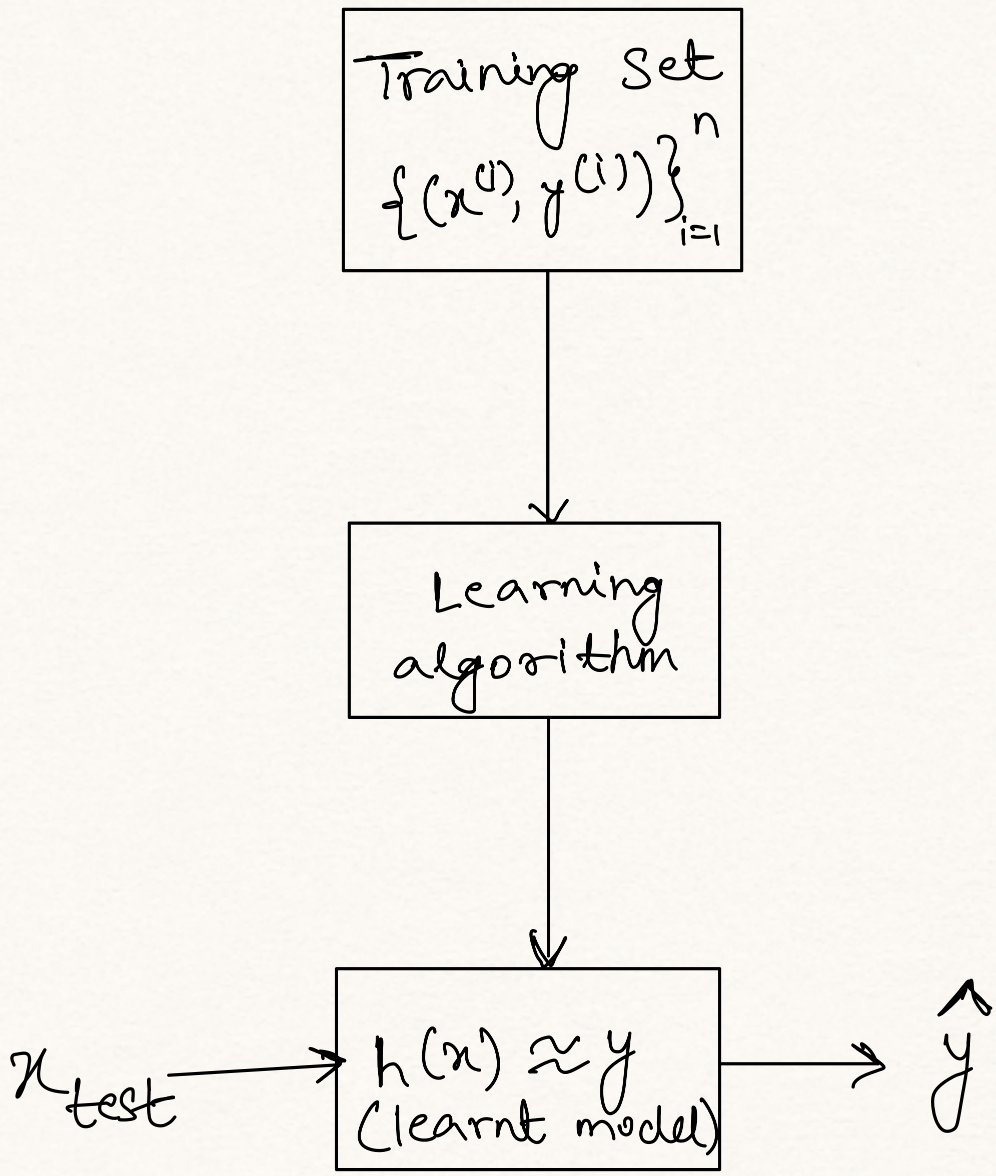
$n$: number of $(x, y)$ examples in training set.

$d$: $x \in \mathbb{R}^d \rightarrow$ dimension of input

$x^{(i)}$: $i^{th}$ example input

$y^{(i)}$: $i^{th}$ example output/label/Ground truth

$(x^{(i)}, y^{(i)})$: $i^{th}$ example



## Linear Regression

$x \in \mathbb{R}^d$, $y \in \mathbb{R}$, $n$ such examples.

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \dots + \theta_d x_d$$

$$\theta \in \mathbb{R}^{d+1}$$

$$h_\theta(x) = \left(\sum_{i=1}^{d} \theta_i x_i\right) + \theta_0$$

$\boxed{x_0 = 1}$ – intercept term

$$h_\theta(x) = \sum_{i=0}^{d} \theta_i x_i = \theta^T x$$

## Cost Function / Loss function

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

squared cost error func$^n$.

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

# Gradient Descent

$\theta^{(0)} :=$ Initialization

$$\theta_j^{(1)} = \theta_j^{(0)} - \alpha \frac{\partial}{\partial \theta_j} J(\theta^{(0)})$$

$\alpha$: learning rate

Repeat till convergence:

$$\boxed{\theta^{(1)} = \theta^{(0)} - \alpha \nabla_\theta J(\theta^{(0)})}$$ In vector form

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_\theta J(\theta^{(t)})$$

- To check convergence we pre-define an $\epsilon < 10^{-5}$ & check either $\|\nabla_\theta(J(\theta^{(t)})\|$ or $\|\theta^{(t)} - \theta^{(t-1)}\|$ or $\|J(\theta^{(t)}) - J(\theta^{(t-1)})\|$ becomes $< \epsilon$.

---

# Gradient descent on Linear Regression

$\theta^{(0)} =$ initialisation

Repeat until convergence:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \nabla_\theta J(\theta^{(t)})$$

$$= \theta^{(t)} - \alpha \nabla_\theta \left[ \frac{1}{2} \sum_{i=1}^{n} \left( h_\theta(x^{(i)}) - y^{(i)} \right)^2 \right]$$

$$= \theta^{(t)} - \alpha \nabla_\theta \left[ \frac{1}{2} \sum_{i=1}^{n} \left( \theta^T x^{(i)} - y^{(i)} \right)^2 \right]$$

$$= \theta^{(t)} - \alpha \left[ \frac{1}{\cancel{2}} \sum_{i=1}^{n} \cancel{2} \left( \theta^T x^{(i)} - y^{(i)} \right) x^{(i)} \right]$$

$$= \underbrace{\theta^{(t)}}_{\text{vector of dim. } d+1} - \alpha \left[ \sum_{i=1}^{n} \underbrace{\left( \theta^T x^{(i)} - y^{(i)} \right)}_{\text{scalar}} \underbrace{x^{(i)}}_{\text{vector of dim. } d+1} \right]$$

## Stochastic Gradient Descent (SGD)

$$\theta^{(t+1)} = \theta^{(t)} - \alpha . \nabla_\theta \tilde{J}(\theta)$$

$$\tilde{J}(\theta) = \frac{1}{2} \left( \theta^T x^{(k)} - y^{(k)} \right)^2$$

$k$: uniformly at random sampled from training set.

(Random updates)

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

Design matrix

$$X = \begin{array}{c} \xleftrightarrow{d+1} \\ \begin{bmatrix} - x^{(1)} - \\ \vdots \\ - x^{(i)} - \\ \vdots \\ - x^{(n)} - \end{bmatrix} \end{array} \Big\updownarrow n \qquad \vec{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_d \end{bmatrix} \qquad \theta_0 x_0 + \theta_1 x_1 + \dots + \theta_d x_d$$

$$X\theta - Y = \begin{bmatrix} x^{(1)T}\theta - y^{(1)} \\ \vdots \\ x^{(i)T}\theta - y^{(i)} \\ \vdots \\ x^{(n)T}\theta - y^{(n)} \end{bmatrix}$$

$$\mathbb{R}^{n \times d+1} \cdot \mathbb{R}^{d+1} - \mathbb{R}^n$$

$$\mathbb{R}^n - \mathbb{R}^n$$

$$\mathbb{R}^n$$

$$J(\theta) = \frac{1}{2}(X\theta - Y)^T (X\theta - Y)$$

$$\nabla_\theta J(\theta) = 0$$

$$\nabla_\theta \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$$

$$\nabla_\theta \frac{1}{2}\left[(X\theta)^T(X\theta) - (X\theta)^T Y - Y^T(X\theta) + Y^T Y\right]$$

$$\nabla_\theta \frac{1}{2}\left[\theta^T(X^T X)\theta - 2\theta^T(X^T Y) + Y^T Y\right]$$

$$= \frac{1}{2}\left[2(X^T X)\theta - 2X^T Y\right] = 0$$

$$\Rightarrow (X^T X)\theta = X^T Y \rightarrow \text{Normal Eq}^n$$

$$\boxed{\theta = (X^T X)^{-1} X^T Y}$$

Given $X^T X$ is invertible.