Linear Regression:
Some Derivations

Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-335

2/23/2026

Special Thanks: Sargur N. Srihari, Andrew Ng, Carlos Guestrin

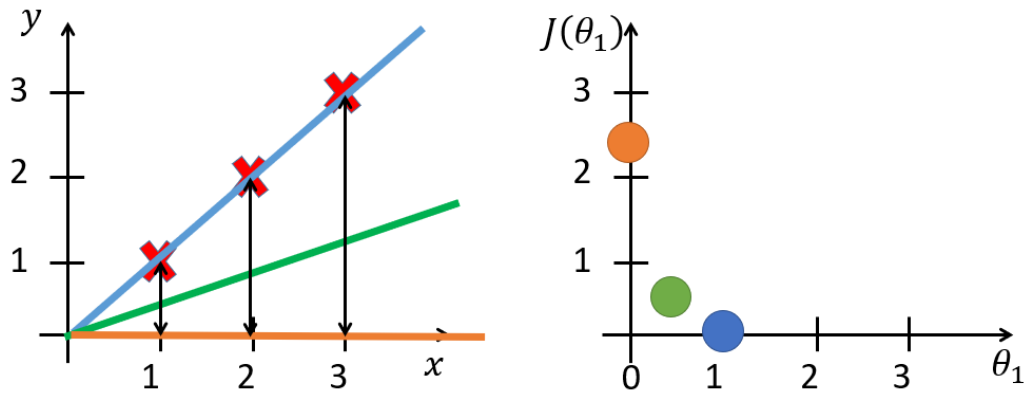
- **Hypothesis:** $h_\theta(x) = \theta_0 + \theta_1 x$

  (Representation)

- **Parameters:** θ_0, θ_1

  (Representation)

$h_\theta(x)$, function of x $J(\theta_1)$, function of θ_1



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

  (Evaluation)

- **Goal**: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

  (Optimization)

Linear Regression

- Model representation
- Cost function
- **Gradient descent**
- Multivariate regression

Gradient Descent

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

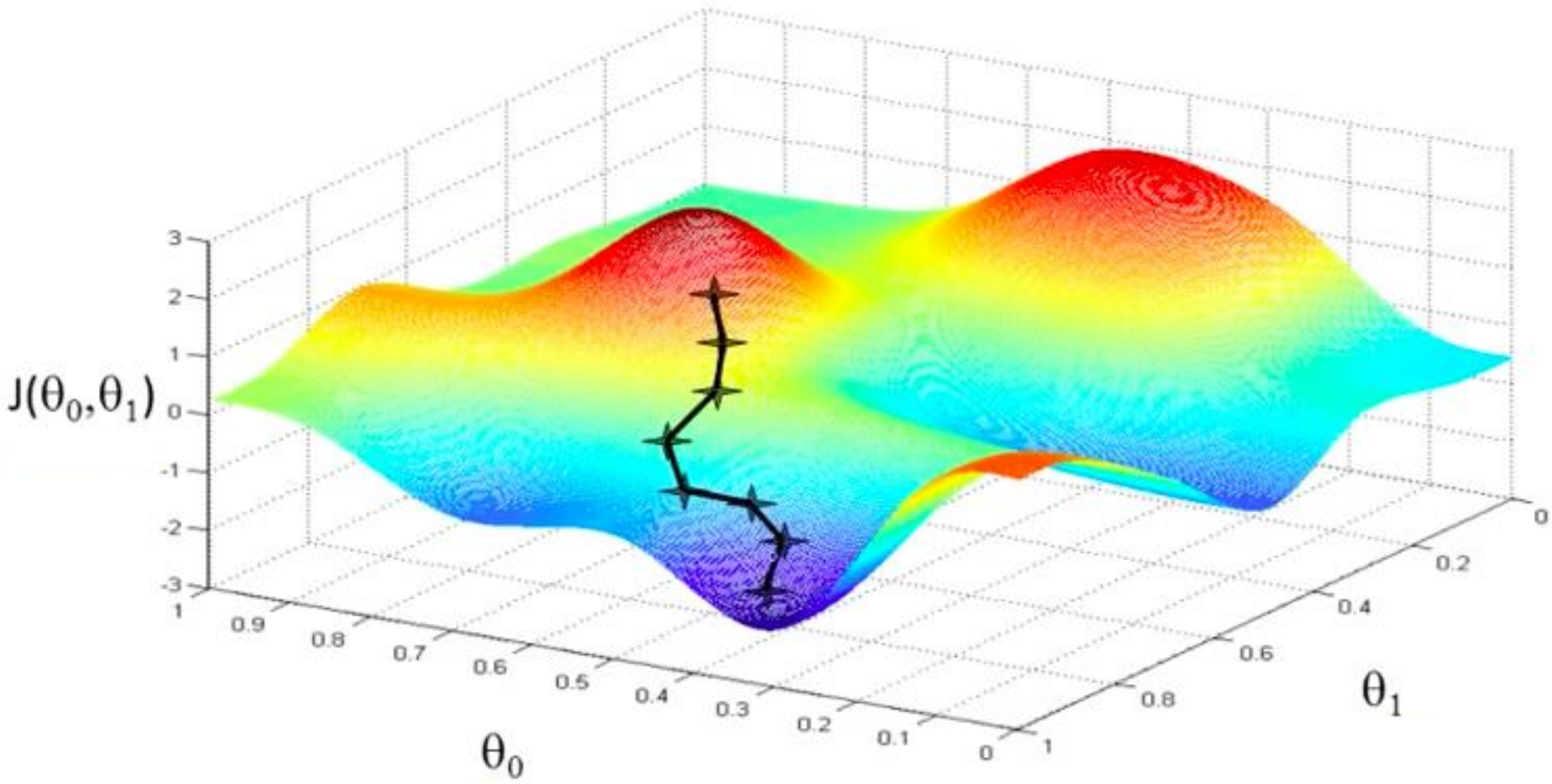
m = 47

Have some function $J(\theta_0, \theta_1)$

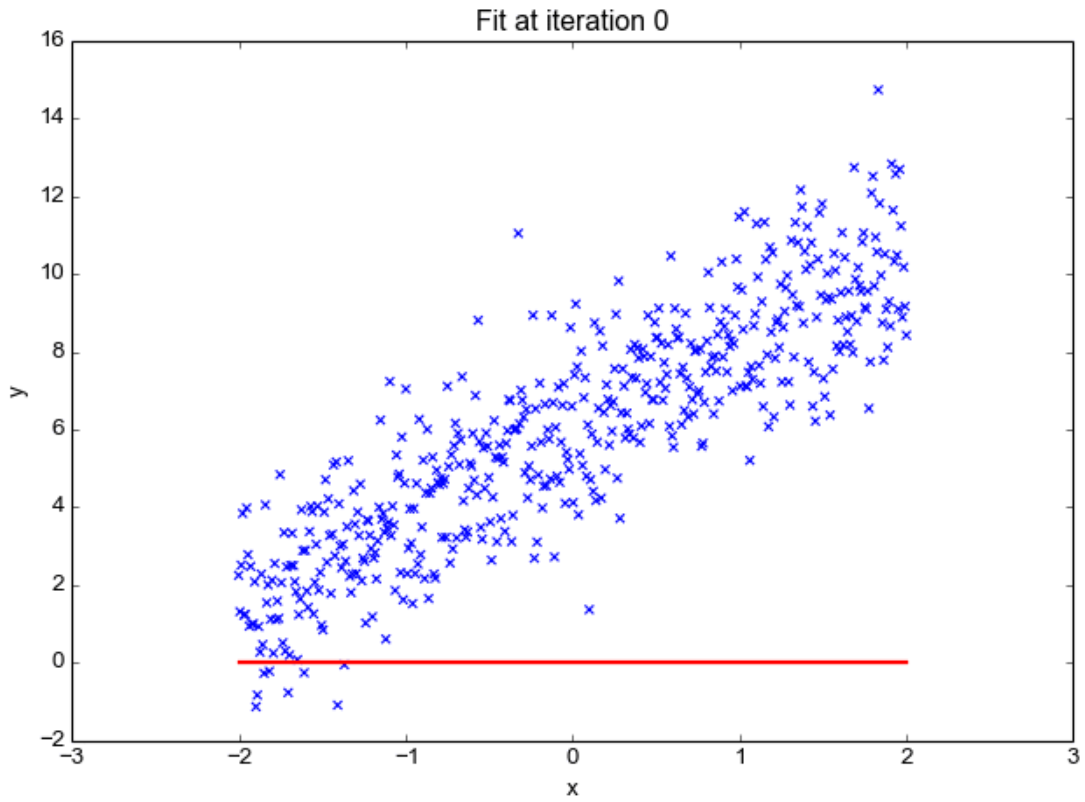
Want: $\underset{\theta_0, \theta_1}{\text{argmin}} \; J(\theta_0, \theta_1)$

Outline:

- Start with some θ_0, θ_1
- Keep changing θ_0, θ_1 to reduce $J(\theta_0, \theta_1)$ until we hopefully end up at minimum


J(θ0,θ1)
3
2
1
0
-1
-2
-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
θ0
0
0.2
0.4
0.6
0.8
1
θ1


WHITE BOARD TIME!


Fit at iteration 0
16
14
12
10
8
6
4
2
0
−2
y
−3
−2
−1
0
1
2
3
x

Gradient Descent

Repeat until convergence{

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ (for $j = 0$ and $j = 1$)

}

α: Learning rate (step size)

$\frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$: derivative (rate of change)

Gradient Descent

***Correct*: simultaneous update**

$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

$$\theta_0 := \text{temp0}$$

$$\theta_1 := \text{temp1}$$

***Incorrect*:**

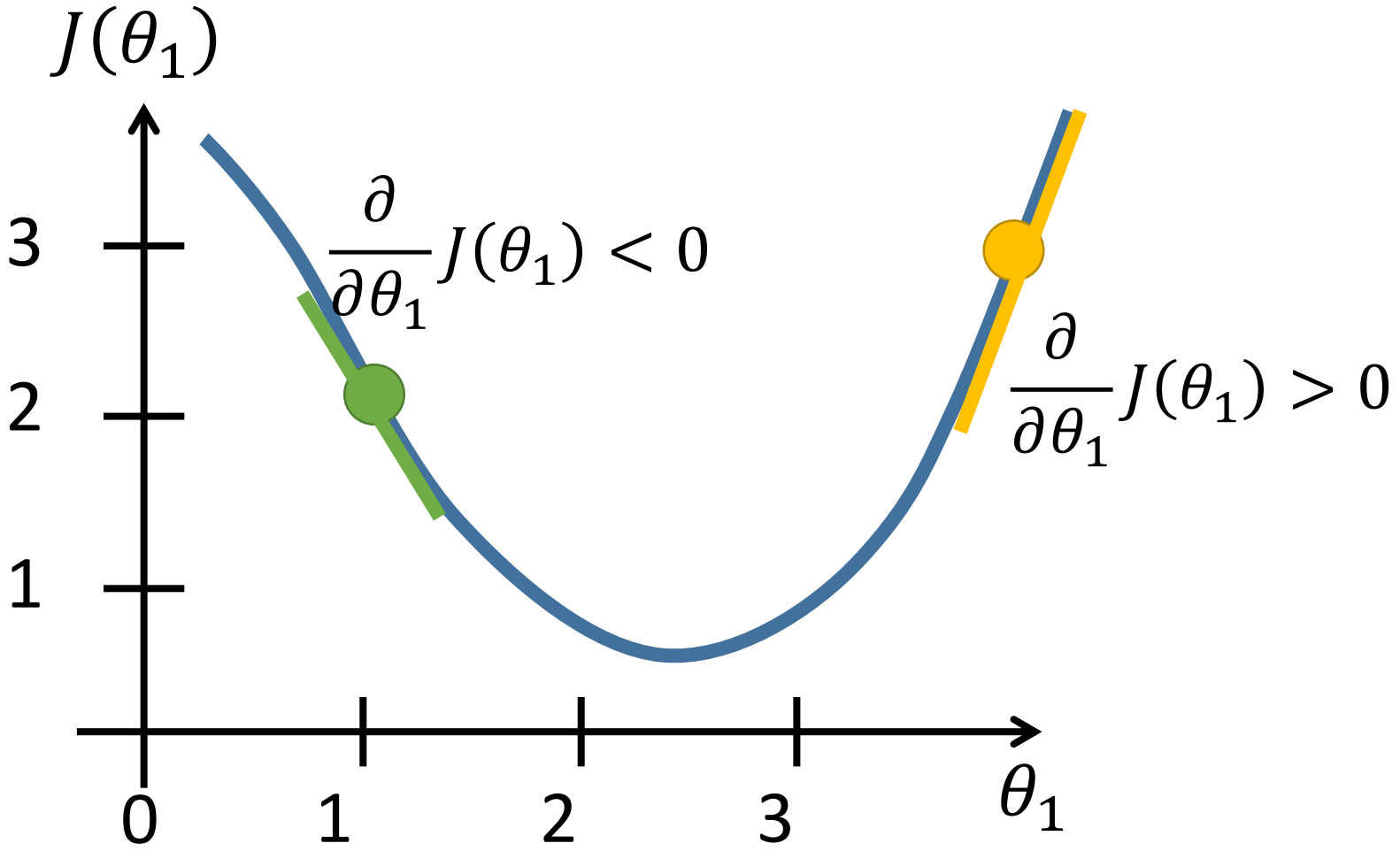
$$\text{temp0} := \theta_0 - \alpha \frac{\partial}{\partial \theta_0} J(\theta_0, \theta_1)$$

$$\theta_0 := \text{temp0}$$

$$\text{temp1} := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_0, \theta_1)$$

$$\theta_1 := \text{temp1}$$

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



Gradient Descent for Linear Regression

Repeat until convergence{

$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)$ (for $j = 0$ and $j = 1$)

}

- Linear regression model

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

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

We derived this!

Computing Partial Derivative(s)

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

$$= \frac{\partial}{\partial \theta_j} \frac{1}{2m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)} \right)^2$$

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

Gradient Descent for Linear Regression

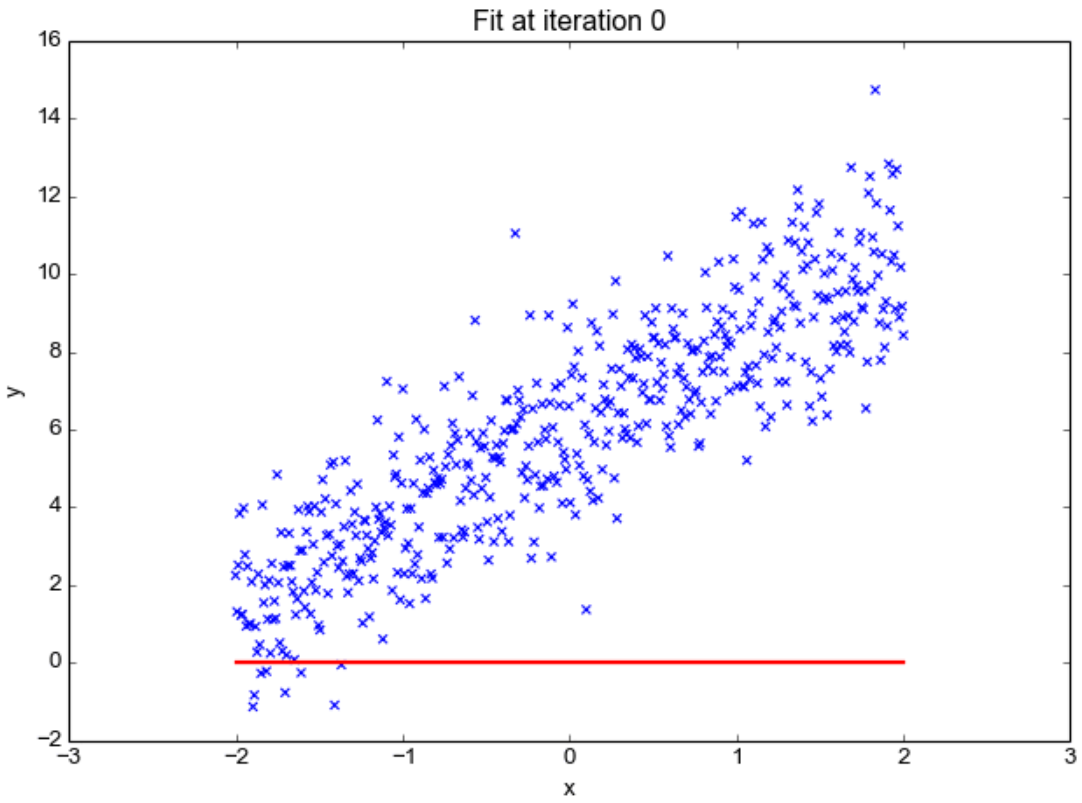
Repeat until convergence{

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x^{(i)}$$

}

Update θ_0 and θ_1 ***simultaneously***



Batch Gradient Descent

- "*Batch*": Each step of gradient descent uses **all** the training examples

Repeat until convergence{

m: Number of training examples

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x^{(i)}$$

}

Batching Up Gradients

- *Another trade-off*: compute/simulation time and gradient error
  - *N* patterns in a dataset, *M* used patterns at any time (in a “mini-batch” subset)
  - Batch gradient descent (GD)
    - $m = M = N$
  - Mini-Batch GD
    - $m = M$; $M < N$, generally $M << N$
  - Stochastic GD (SGD)
    - $m = M = 1$

Involve sampling randomness!

Computational Bottleneck

- A recurring problem in machine learning:
  - Large training sets are necessary for good generalization
  - *BUT* large training sets are also computationally expensive to use
- Stochastic gradient descent (SGD) is an extension of gradient descent (GD) that offers a solution
  - Moreover, it is a vehicle for generalization beyond training set
  - Expectation may be approximated using small set of samples (we will also later refer to these sets as “mini-batches” → mini-batch GD)

Linear Regression

- Model representation
- Cost function
- Gradient descent
- **Multivariate regression**

Training Dataset

What if instead of this univariate dataset…

Size in feet^2 (x)	Price ($) in 1000's (y)
2104	460
1416	232
1534	315
852	178
…	…

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

Multiple Features (Input Variables)

…we had this multivariate dataset instead?

Size in feet^2 (x_1)	Number of bedrooms (x_2)	Number of floors (x_3)	Age of home (years) (x_4)	Price ($) in 1000's (y)
2104	5	1	45	460
1416	3	2	40	232
1534	3	2	30	315
852	2	1	36	178
…				…

Notation:

$D = n$ = Number of features

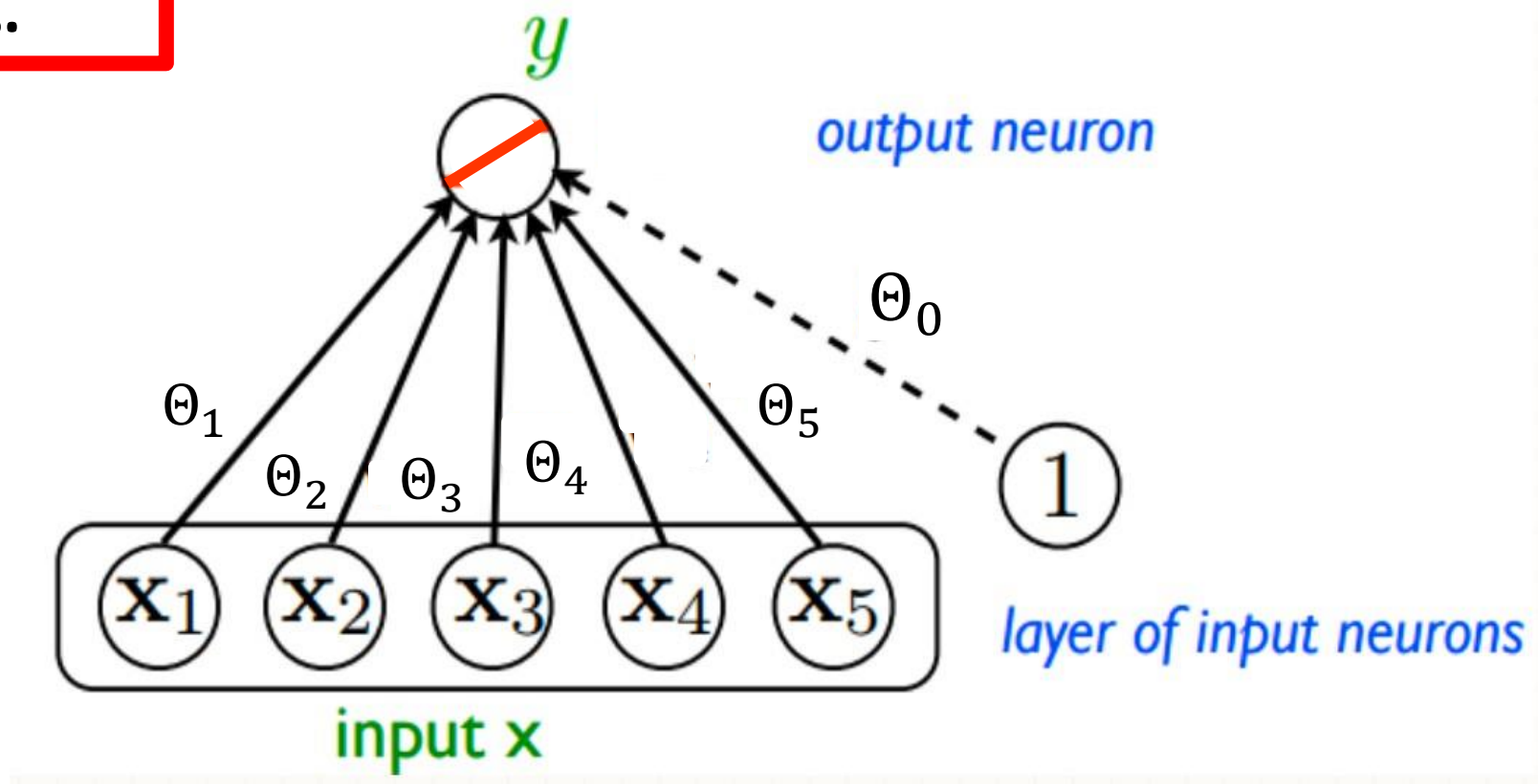
$x^{(i)}$= Input features of i^{th} training example

$x_j^{(i)}$= Value of feature j in i^{th} training example

$x_3^{(2)} = ?$

$x_3^{(4)} = ?$

Looks like we're going to need to alter our regressor just a bit…



Hypothesis: Multivariate Form

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

Now: $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

- For convenience of notation, define $x_0 = 1$ ($x_0^{(i)} = 1$ for *all* examples)

- $\boldsymbol{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \in R^{n+1}$ $\quad \boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \in R^{n+1}$

This yields the "bias"

This is that "padding feature" we talked about last time!

- $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$
  $= \boldsymbol{\theta}^\top \boldsymbol{x}$

Notice that this becomes a simple dot-product!

Gradient Descent (Multivariate)

- Previously ($n = 1$)

Repeat until convergence{

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right)$$

$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x^{(i)}$$

}

- New algorithm ($n \geq 1$)

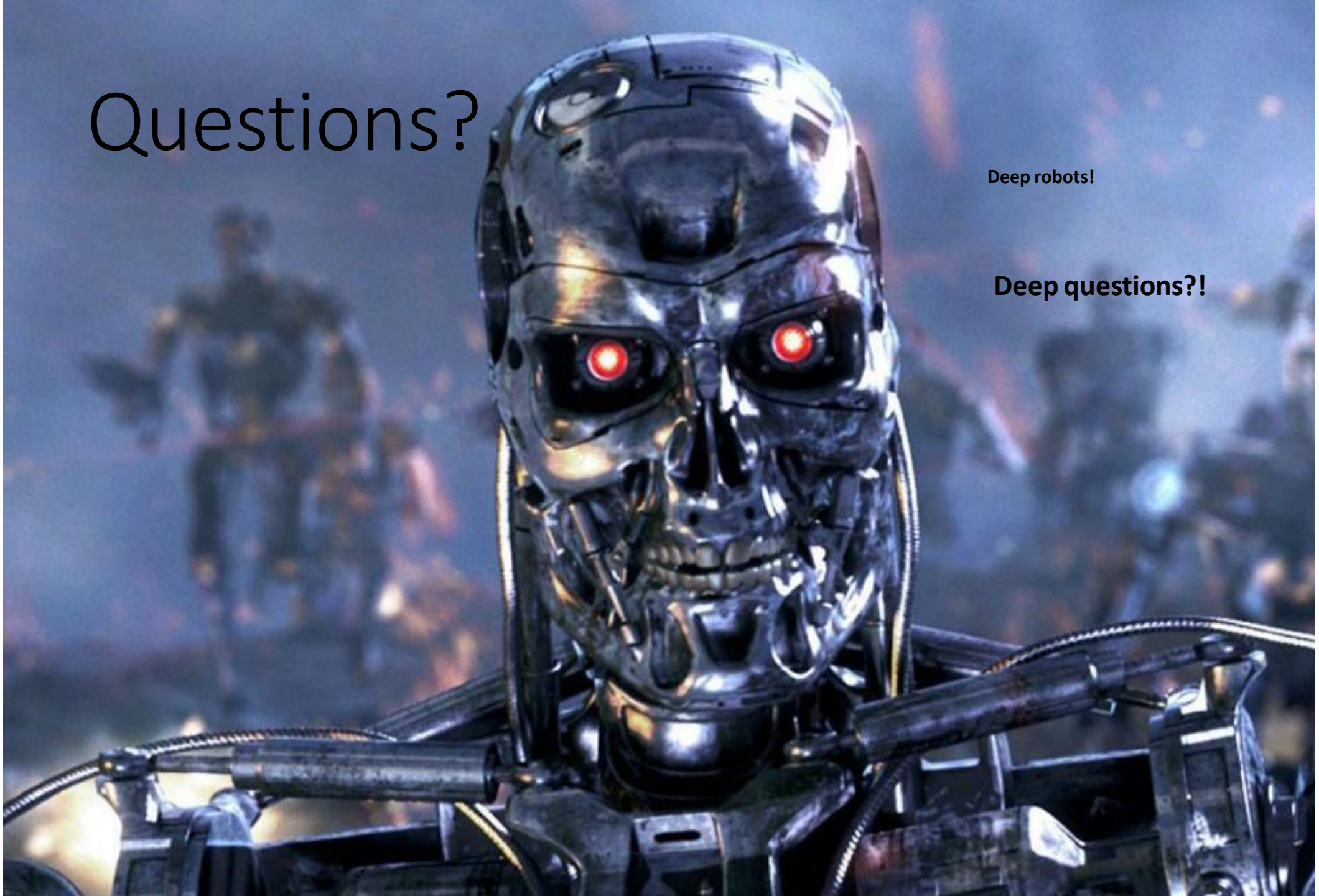
Repeat until convergence{

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

}

Simultaneously update
$\theta_j, \text{for } j = 0, 1, \cdots, n$

Questions?



Deep robots!

Deep questions?!