香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 2

Math Basics and Supervised Learning: Regression

Junxian He
Feb 10, 2026

Announcement

Lecture videos till next week will be released considering the Lunar New Year

Example Distributions

Distribution	PDF or PMF	Mean	Variance
$Bernoulli(p)$	$\begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0. \end{cases}$	p	$p(1-p)$
$Binomial(n, p)$	$\binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$	np	$np(1-p)$
$Geometric(p)$	$p(1-p)^{k-1}$ for $k = 1, 2, \ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
$Poisson(\lambda)$	$\frac{e^{-\lambda}\lambda^k}{k!}$ for $k = 0, 1, \ldots$	λ	λ
$Uniform(a, b)$	$\frac{1}{b-a}$ for all $x \in (a, b)$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$Gaussian(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ for all $x \in (-\infty, \infty)$	μ	σ^2
$Exponential(\lambda)$	$\lambda e^{-\lambda x}$ for all $x \geq 0, \lambda \geq 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Joint and Marginal Distributions

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

$$p_{XY}(x, y) = P(X = x, Y = y)$$

Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x, y) = 1$

Joint and Marginal Distributions

- **Joint PMF** for discrete RV's X, Y:

$$p_{XY}(x,y) = P(X = x, Y = y)$$

Note that $\sum_{x \in Val(X)} \sum_{y \in Val(Y)} p_{XY}(x,y) = 1$

marginalizing out y

- **Marginal PMF** of X, given joint PMF of X, Y:

$$p_X(x) = \sum_y p_{XY}(x,y)$$

$P(x, y)$ $P(x)$ $P(y)$

Joint and Marginal Distributions

- **Joint PDF** for continuous RV's $X_1, ..., X_n$:

$$f(x_1, ..., x_n) = \frac{\delta^n F(x_1, ...x_n)}{\delta x_1 \delta x_2 ... \delta x_n}$$

  Note that $\int_{x_1} \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_1 ... dx_n = 1$

- **Marginal PDF** of X_1, given joint PDF of $X_1, ..., X_n$:

$$f_{X_1}(x_1) = \int_{x_2} ... \int_{x_n} f(x_1, ..., x_n) dx_2 ... dx_n$$

Expectation for multiple random variables

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$ of X, Y,

- for discrete X, Y:

$$\mathbb{E}[g(X,Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x,y) p_{XY}(x,y)$$

Expectation for multiple random variables

Given two RV's X, Y and a function $g : \mathbb{R}^2 \to \mathbb{R}$ of X, Y,

- for discrete X, Y:

$$\mathbb{E}[g(X, Y)] := \sum_{x \in Val(x)} \sum_{y \in Val(y)} g(x, y) p_{XY}(x, y)$$

- for continuous X, Y:

$$\mathbb{E}[g(X, Y)] := \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

Covariance

$E[(X - E(x))^2]$

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$
\begin{aligned}
Cov[X, Y] &:= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

Covariance

Intuitively: measures how much one RV's value tends to move with another RV's value. For RV's X, Y:

$$Cov[X, Y] := \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$
$$= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

- If $Cov[X, Y] < 0$, then X and Y are negatively correlated
- If $Cov[X, Y] > 0$, then X and Y are positively correlated
- If $Cov[X, Y] = 0$, then X and Y are uncorrelated

Variance of two variables

$$Var[X + Y] = Var[X] + Var[Y] + 2Cov[X, Y]$$

Conditional distributions for RVs

Works the same way with RV's as with events:

joint

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{p_X(x)}$$

marginal

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)}$$

- In general, for continuous $X_1, \ldots, X_n$:

$$f_{X_1|X_2,\ldots,X_n}(x_1|x_2,\ldots,x_n) = \frac{f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n)}{f_{X_2,\ldots,X_n}(x_2,\ldots,x_n)}$$

Bayes' Rule for RVs

$P(y) \quad P(x|y)$

Also works the same way for RV's as with events:

$P(y|x) = \frac{P(x,y)}{P(x)} = \frac{P(y)P(x|y)}{}$

- For discrete X, Y:

$$p_{Y|X}(y|x) = \frac{p_{X|Y}(x|y)p_Y(y)}{\sum_{y' \in Val(Y)} p_{X|Y}(x|y')p_Y(y')}$$

$P(x) = \sum_y P(x,y)$

$= \sum_y P(y)P(x|y)$

- For continuous X, Y:

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{-\infty}^{\infty} f_{X|Y}(x|y')f_Y(y')dy'}$$

$P(z)$

$z \to x$

$P(x|z)$

$P(z|x)$

Random Vectors

Random Vectors

Given n RV's $X_1, \ldots, X_n$, we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Random Vectors

Given n RV's $X_1, \ldots, X_n$, we can define a random vector X s.t.

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix}$$

Note: all the notions of joint PDF/CDF will apply to X.
Given $g : \mathbb{R}^n \to \mathbb{R}^m$, we have:

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \vdots \\ g_m(x) \end{bmatrix}, \mathbb{E}[g(X)] = \begin{bmatrix} \mathbb{E}[g_1(X)] \\ \mathbb{E}[g_2(X)] \\ \vdots \\ \mathbb{E}[g_m(X)] \end{bmatrix}$$

Covariance Matrices

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

Covariance Matrices

n

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix} \rightarrow n \times n$$

Covariance Matrices

$X \to \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ \vdots \end{pmatrix}$

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

high-dim

Gaussian

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix}$$

$XX^T \to n \times n$

$X^TX \to$ scalar

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

outer product

$x - E(x)$ $x - E(x)$

n n

Covariance Matrices

For a random vector $X \in \mathbb{R}^n$, we define its **covariance matrix** Σ as the $n \times n$ matrix whose ij-th entry contains the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} Cov[X_1, X_1] & \dots & Cov[X_1, X_n] \\ \vdots & \ddots & \vdots \\ Cov[X_n, X_1] & \dots & Cov[X_n, X_n] \end{bmatrix}$$

applying linearity of expectation and the fact that $Cov[X_i, X_j] = \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])]$, we obtain

$$\Sigma = \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T]$$

A is

$z^T A z \geq 0 \longrightarrow$ PSD

Properties:

- Σ is symmetric and PSD
- If $X_i \perp X_j$ for all i, j, then $\Sigma = diag(Var[X_1], ..., Var[X_n])$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{\det(\Sigma)^{\frac{1}{2}} (2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Variance $= \sigma^2$

$$\frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$

Multivariate Gaussian

The multivariate Gaussian $X \sim \mathcal{N}(\mu, \Sigma)$, $X \in \mathbb{R}^n$:

$$p(x; \mu, \Sigma) = \frac{1}{det(\Sigma)^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

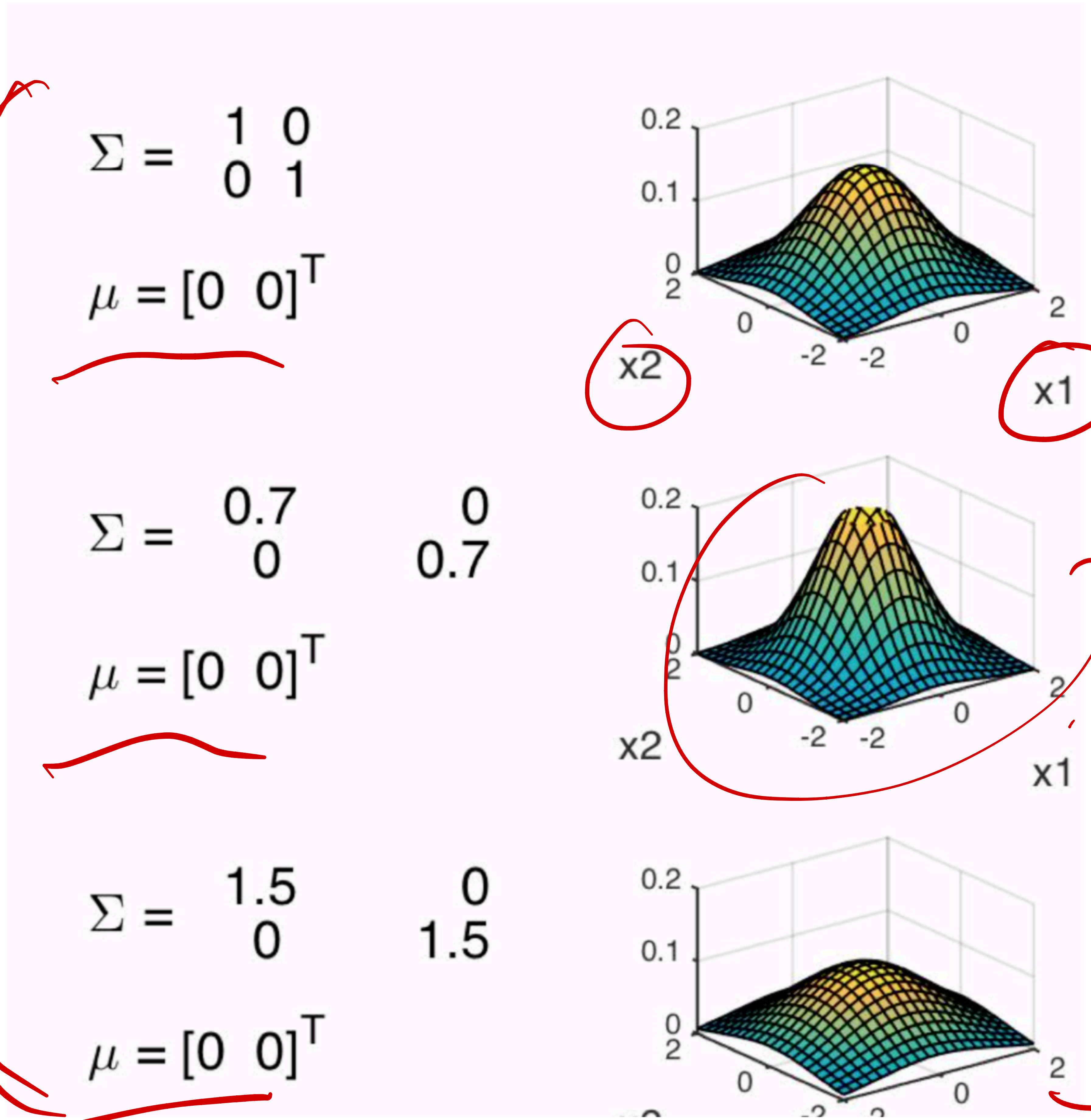
Gaussian when $n = 1$.

$$p(x; \mu, \sigma^2) = \frac{1}{\sigma(2\pi)^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

Notice that if $\Sigma \in \mathbb{R}^{1\times 1}$, then $\Sigma = Var[X_1] = \sigma^2$, and so $\Sigma^{-1} = \frac{1}{\sigma^2}$ and $det(\Sigma)^{\frac{1}{2}} = \sigma$

MV Gaussian Visualization

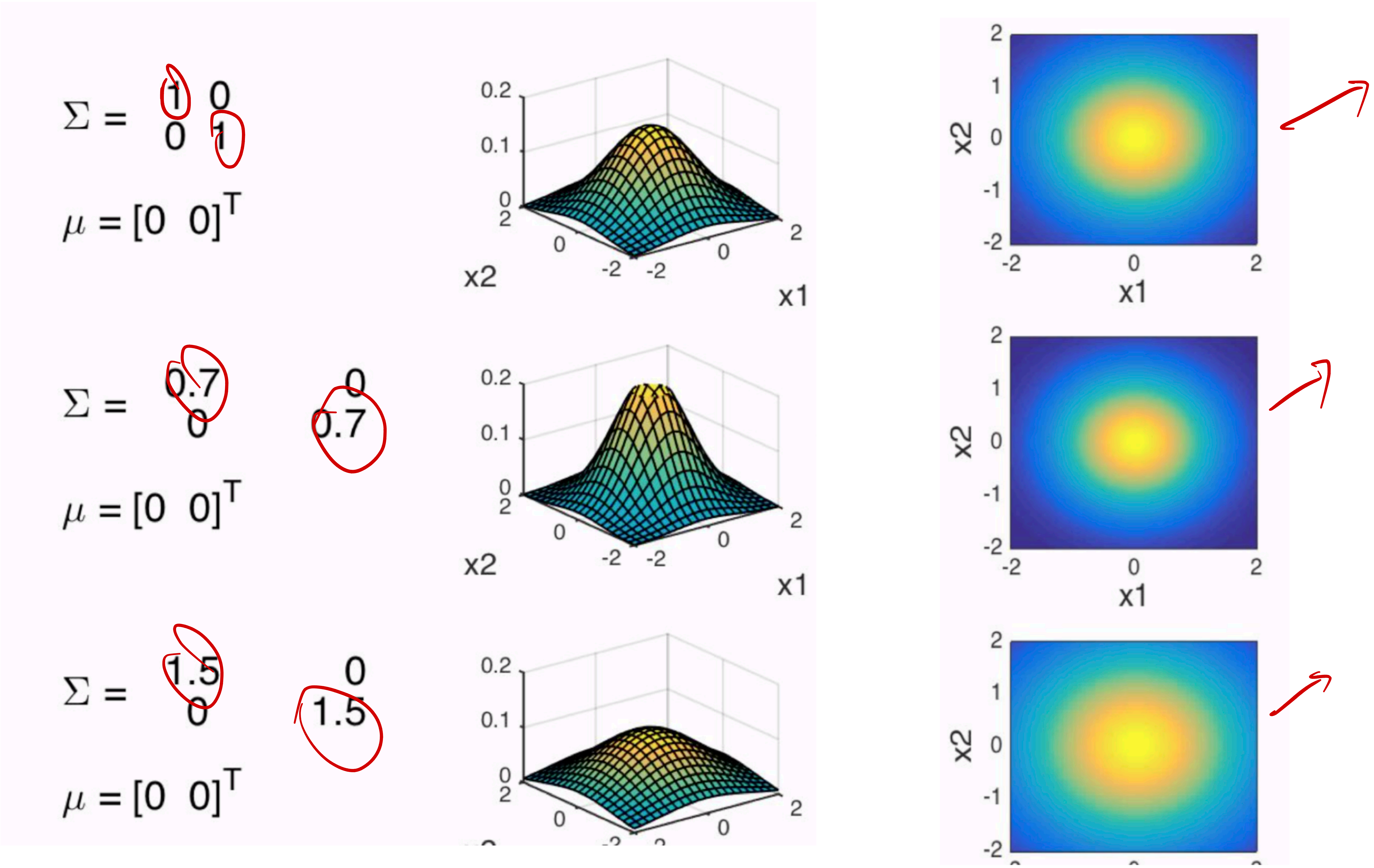
MV Gaussian Visualization



sharper

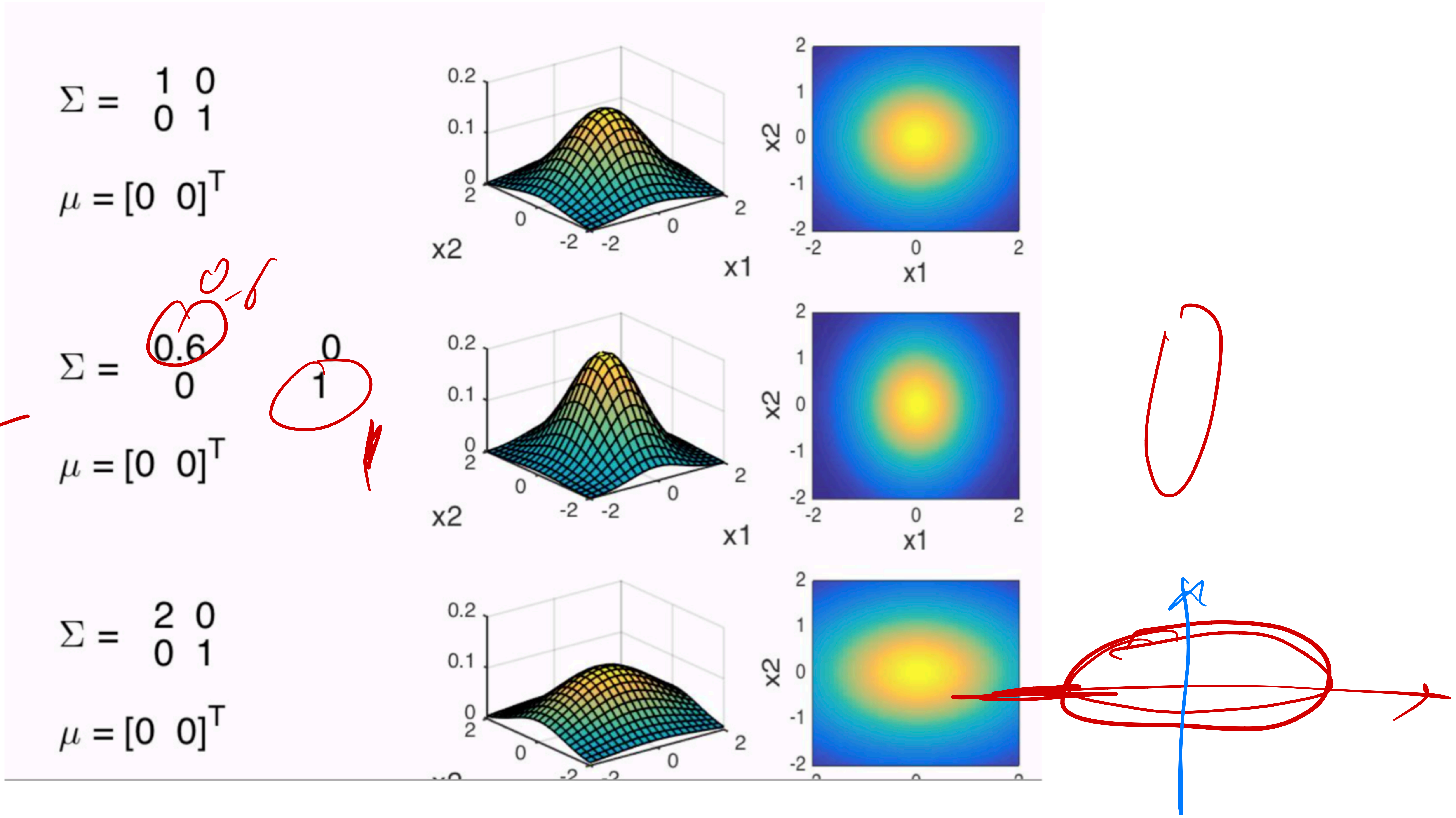
Effect of changing variance

MV Gaussian Visualization



Effect of changing variance

MV Gaussian Visualization

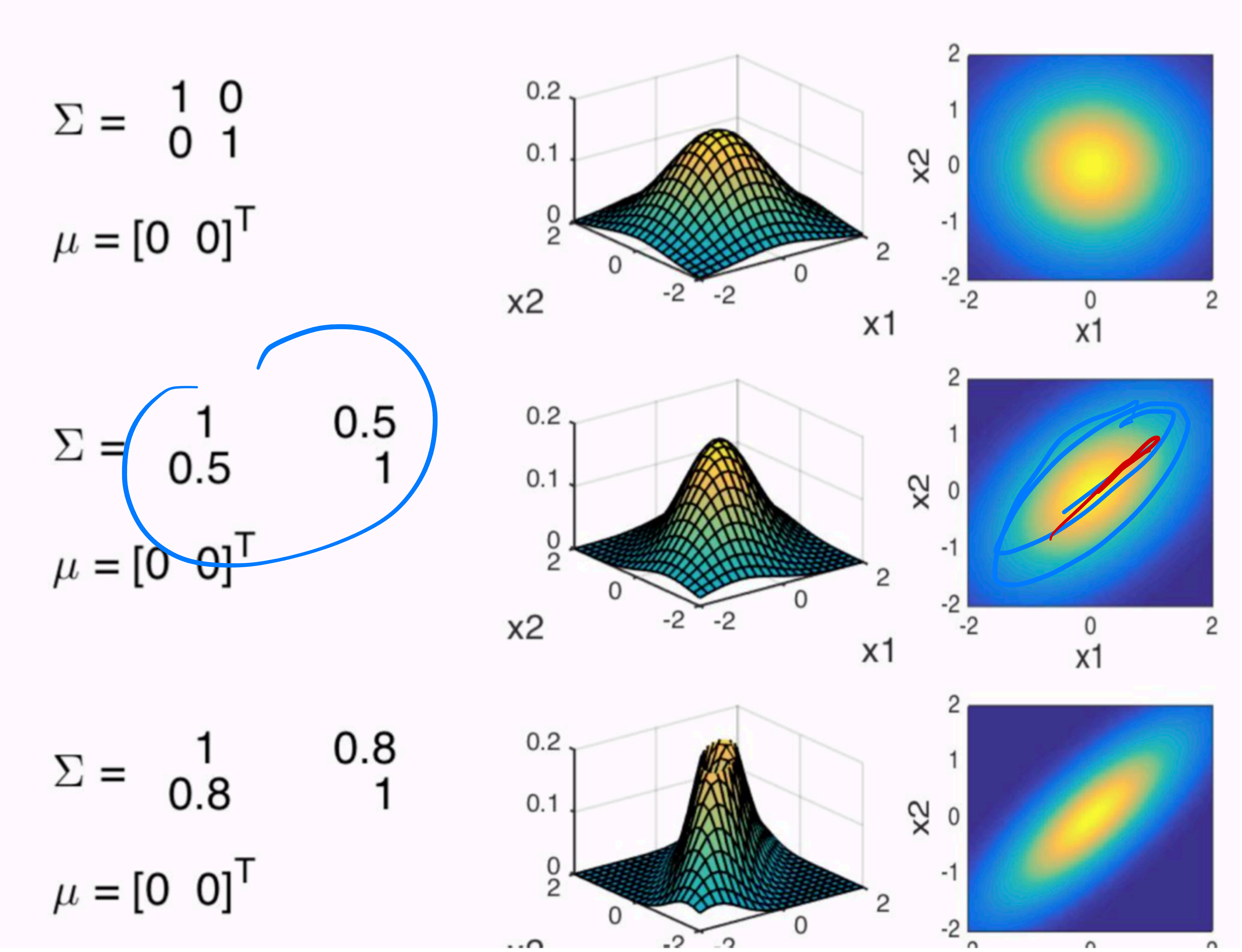


If $Var[X_1] \neq Var[X_2]$:

MV Gaussian Visualization

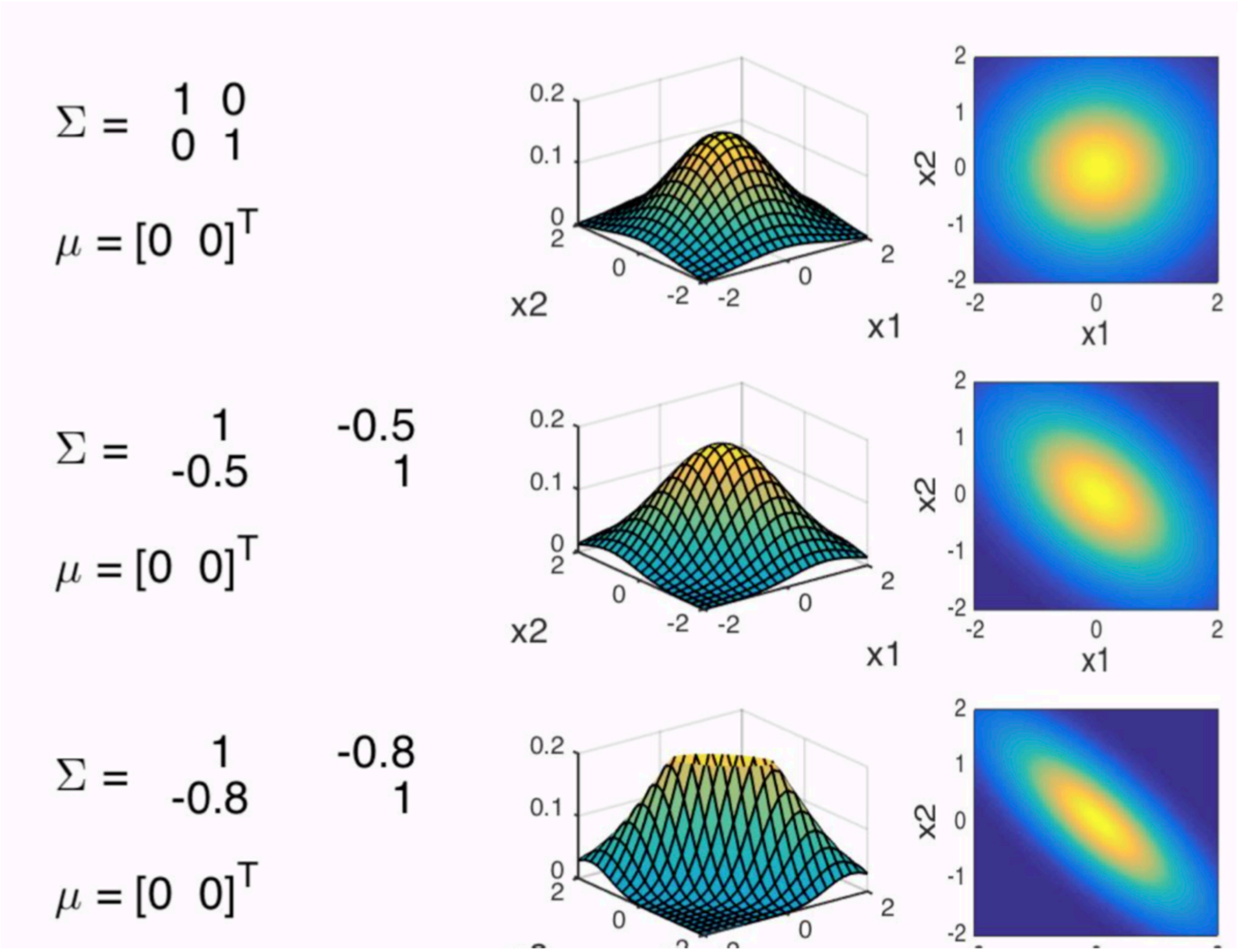
$x_1 \uparrow \quad x_2 \uparrow$

If X_1 and X_2 are positively correlated:



MV Gaussian Visualization

If X_1 and X_2 are negatively correlated:



The purpose of computation is insight, not numbers.

- Richard Hamming

matrix

The purpose of computation is insight, not numbers.
- Richard Hamming

vocabulary embedding

RoPE

https://www.youtube.com/@3blue1brown/courses

1 Vectors | Chapter 1, Essence of linear algebra

2 Linear combinations, span, and basis vectors | Chapter 2, Essence of linear algebra

3 Linear transformations and matrices | Chapter 3, Essence of linear algebra

4 Matrix multiplication as composition | Chapter 4, Essence of linear algebra

5 Three-dimensional linear transformations | Chapter 5, Essence of linear algebra

6 The determinant | Chapter 6, Essence of linear algebra

Inverse matrices, column space and null space | Chapter 7, Essence of linear algebra

3Blue1Brown

@3blue1brown · 5.88M subscribers · 172 videos

My name is Grant Sanderson. Videos here cover a variety of topics in math, or adjacent fiel...

3blue1brown.com and 7 more links

Subscribed

香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

Supervised Learning: Regression

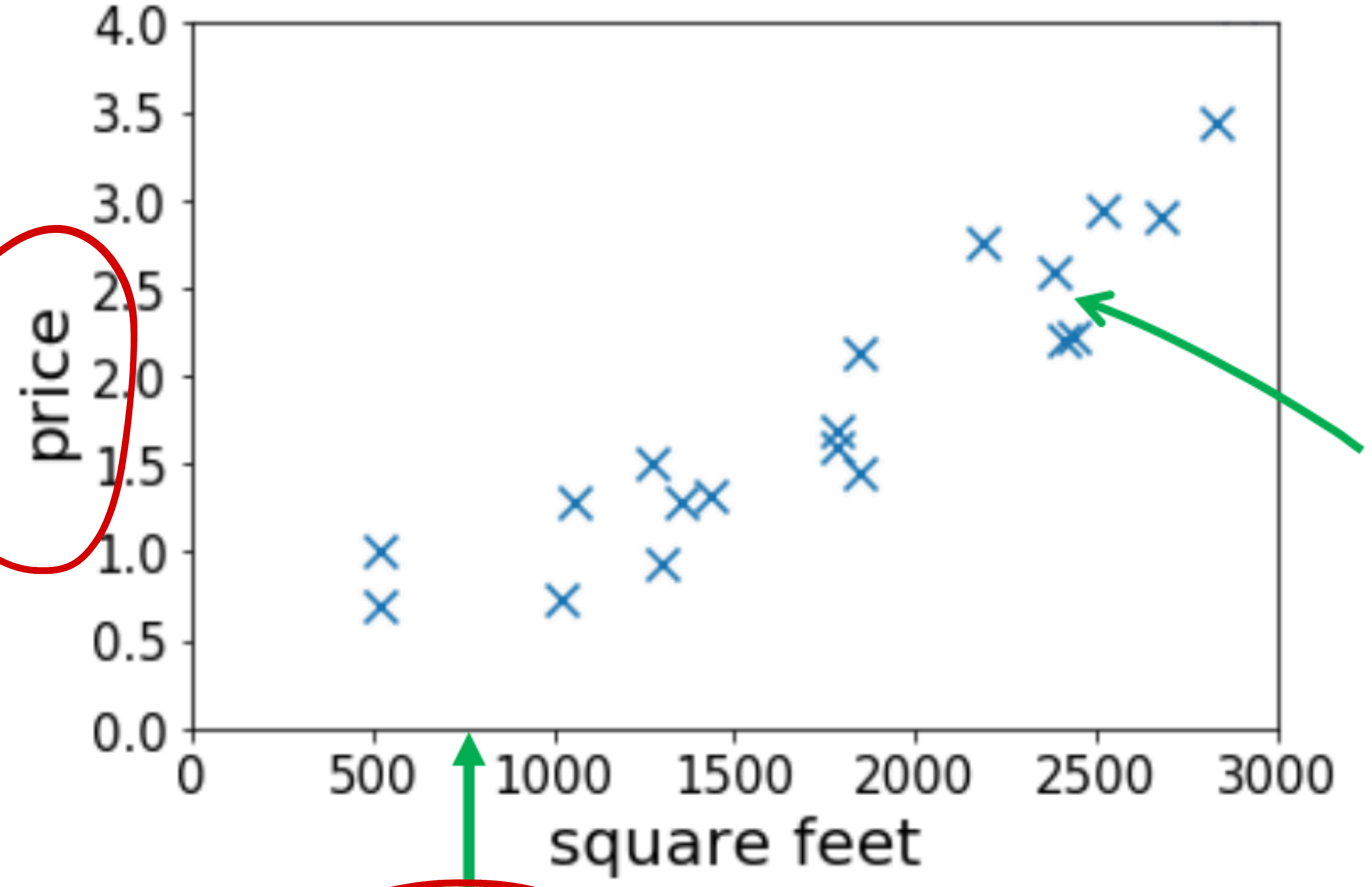
Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \to \mathcal{Y}$

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$

$h(800)$

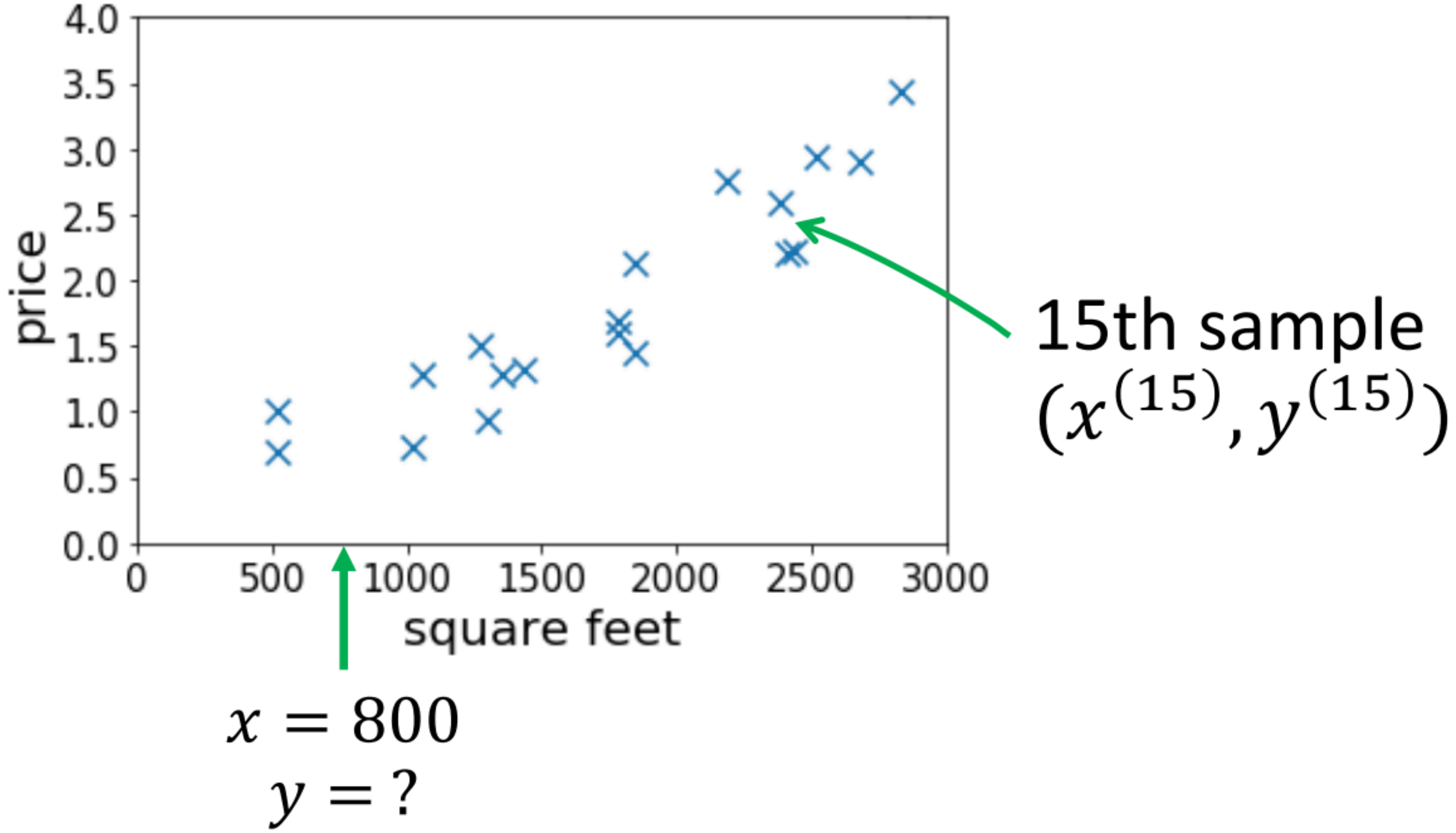


15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

Supervised Learning

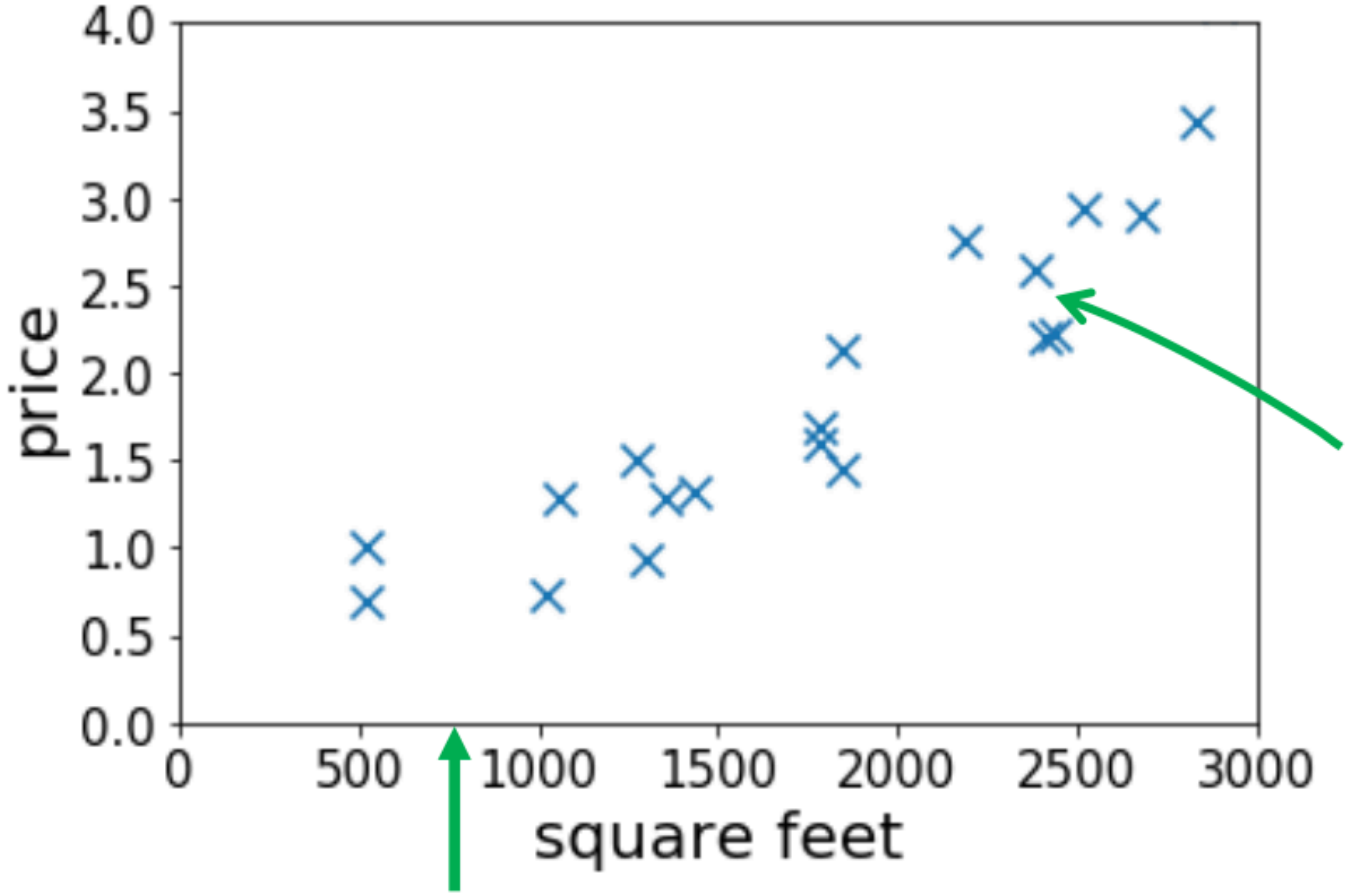
- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$



CAT

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \to \mathcal{Y}$



15th sample
$(x^{(15)}, y^{(15)})$

$x = 800$
$y = ?$

X

CAT Y

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \rightarrow \mathcal{Y}$

- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$
  s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \ldots, n$.

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \to \mathcal{Y}$

- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$

  s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \ldots, n$.

- Given a training set our goal is to produce a good prediction function h

Supervised Learning

- A hypothesis or a prediction function is function $h : \mathcal{X} \to \mathcal{Y}$

- A training set is set of pairs $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})\}$
  s.t. $x^{(i)} \in \mathcal{X}$ and $y^{(i)} \in \mathcal{Y}$ for $i = 1, \ldots, n$.

- Given a training set our goal is to produce a good prediction function h

- If $\mathcal{Y}$ is continuous, then called a regression problem
- If $\mathcal{Y}$ is discrete, then called a classification problem → GPT

Supervised Learning

- How to define "good" for a prediction function?
  - Metrics / performance
  - Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \cdots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Supervised Learning

- How to define "good" for a prediction function?
  - Metrics / performance
  - Good on unseen data

h

test

unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \cdots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

develop

Does not overlap with training dataset

Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \cdots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

Does not overlap with training and validation dataset

Supervised Learning

- How to define "good" for a prediction function?
  - Metrics / performance
  - Good on unseen data

validation

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \cdots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

development

Does not overlap with training dataset

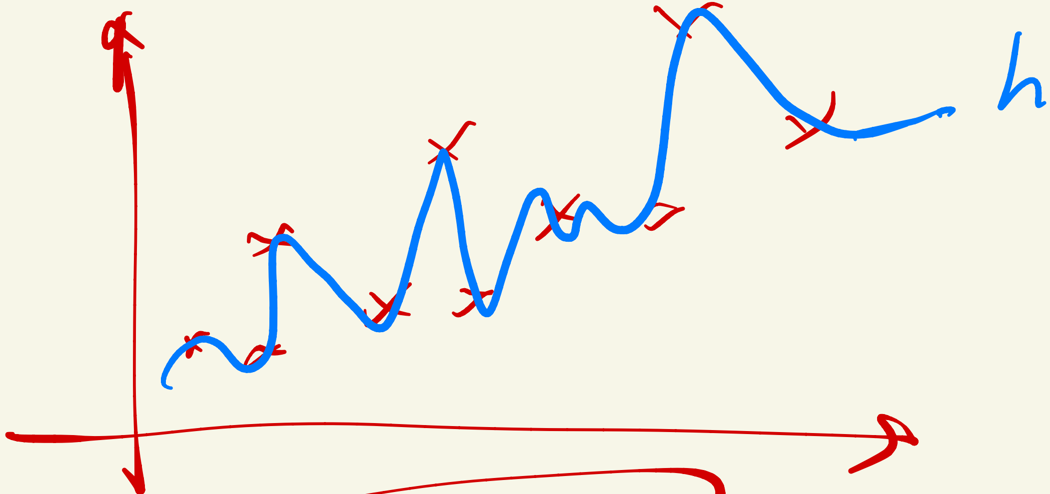
Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \cdots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

real user data

Does not overlap with training and validation dataset

Completely unseen before deployment

Realistic setting



$$\boxed{y = h(x)}$$

linear ?

polynomial x^2, x^3, x^4 what order

Supervised Learning

- How to define "good" for a prediction function?
  - Metrics / performance
  - Good on unseen data

Validation dataset is another set of pairs $\{(\hat{x}^{(1)}, \hat{y}^{(1)}), \cdots, (\hat{x}^{(m)}, \hat{y}^{(m)})\}$

Does not overlap with training dataset

Test dataset is another set of pairs $\{(\tilde{x}^{(1)}, \tilde{y}^{(1)}), \cdots, (\tilde{x}^{(L)}, \tilde{y}^{(L)})\}$

Does not overlap with training and validation dataset

Completely unseen before deployment

parameter

Realistic setting

Hyperparameter tuning is a form of training

test loss

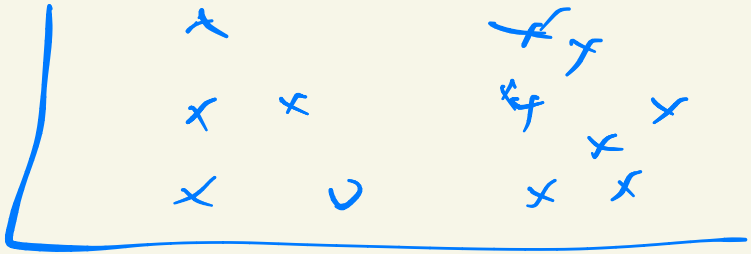
optimizer

human

brain space

1. unsupervised training

kmeans ?



scientific discovery

2. don't care about unseen data

1. Solve a mathematical hypothesis

2. algorithm optimization } → kmeans

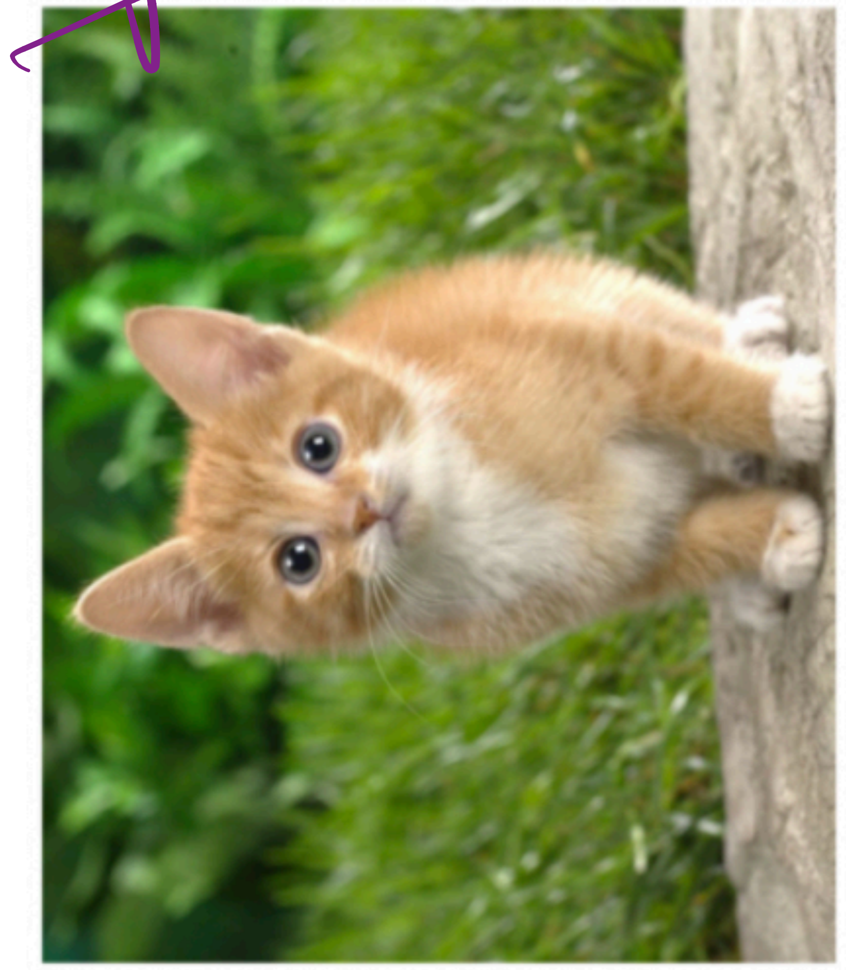
improved

10x

Supervised Training

Generalization

Train



Validation

Test

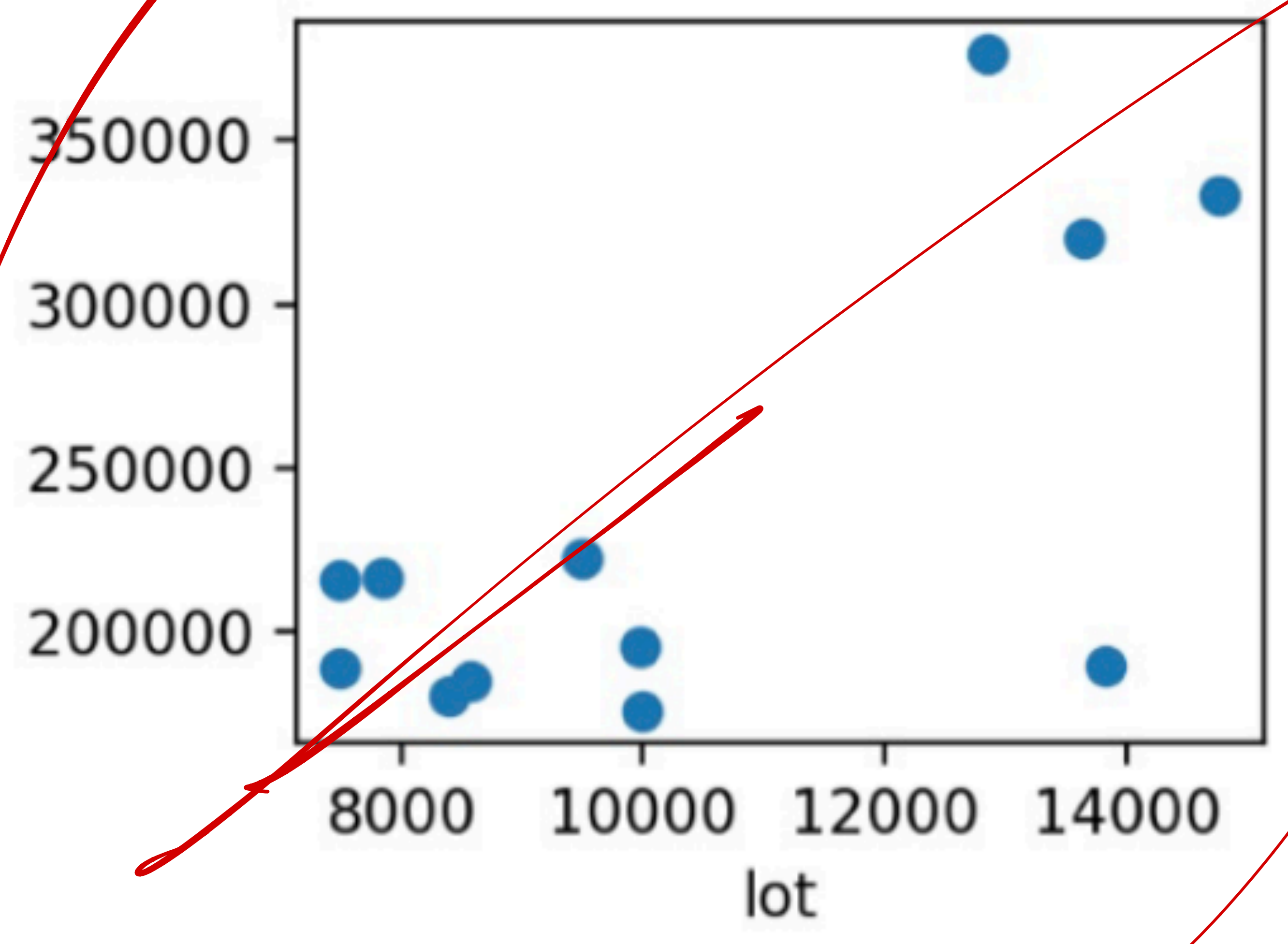
Not only for supervised learning

Example: Regression using Housing Data

Example from Stanford CS229

Example Housing Data

	SalePrice	Lot.Area
4	189900	13830
5	195500	9978
9	189000	7500
10	175900	10000
12	180400	8402
22	216000	7500
36	376162	12858
47	320000	13650
55	216500	7851
56	185088	8577



Represent h as a Linear Function

affine

$h(x) = \theta_0 + \theta_1 x_1$ is an *affine function*

Popular choice

$h(x) = \theta_1 x_1 + \theta_0$

Represent h as a Linear Function

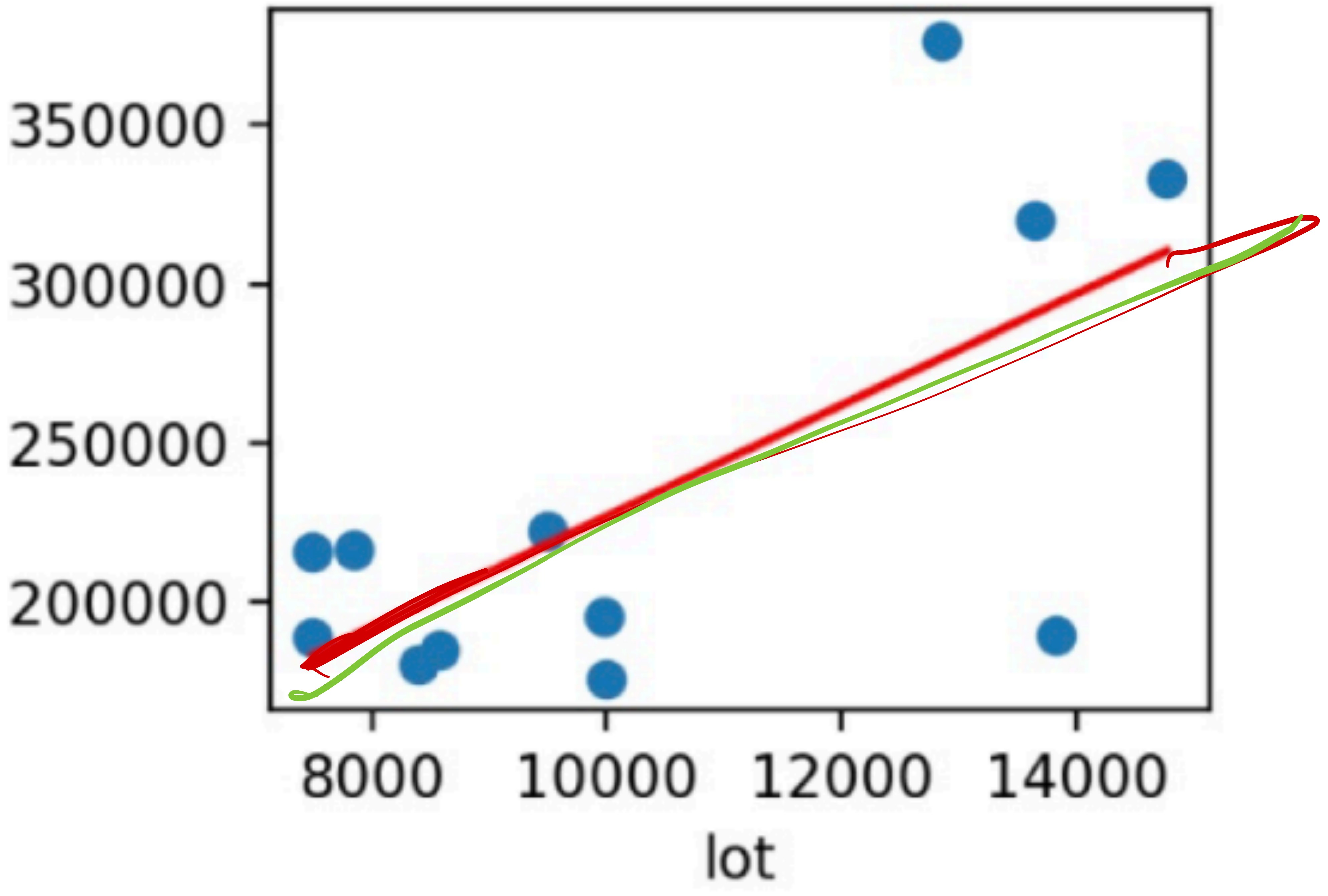
$h(x) = \theta_0 + \theta_1 x_1$ is an *affine function*

Popular choice

learning

The function is defined by **parameters** θ_0 and θ_1, the function space is greatly reduced

Simple Line Fit



More Features

	size	bedrooms	lot size
$x^{(1)}$	2104	4	45k
$x^{(2)}$	2500	3	30k

	Price
$y^{(1)}$	400
$y^{(2)}$	900

More Features

	size	bedrooms	lot size
$x^{(1)}$	2104	4	45k
$x^{(2)}$	2500	3	30k

	Price
$y^{(1)}$	400
$y^{(2)}$	900

What's a prediction here?

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

More Features

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

What's a prediction here?

$$h(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3.$$

With the convention that $x_0 = 1$ we can write:

$$h(x) = \sum_{j=0}^{3} \theta_j x_j$$

$x_0 = 1$

$h(x) = \sum_{j=0}^{3} \theta_j x_j$

Vector Notations

Vector Notations

	size	bedrooms	lot size
$x^{(1)}$	2104	4	45k
$x^{(2)}$	2500	3	30k

	Price
$y^{(1)}$	400
$y^{(2)}$	900

We write the vectors as (important notation)

$$\vec{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } \vec{x}^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

Vector Notations

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

We write the vectors as (important notation)

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

We call θ **parameters**, $x^{(i)}$ is the input or the **features**, and the output or **target** is $y^{(i)}$. To be clear,

(x, y) is a training example and $(x^{(i)}, y^{(i)})$ is the i^{th} example.

Vector Notations

	size	bedrooms	lot size		Price
$x^{(1)}$	2104	4	45k	$y^{(1)}$	400
$x^{(2)}$	2500	3	30k	$y^{(2)}$	900

We write the vectors as (important notation)

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} \text{ and } x^{(1)} = \begin{pmatrix} x_0^{(1)} \\ x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix} = \begin{pmatrix} 1 \\ 2104 \\ 4 \\ 45 \end{pmatrix} \text{ and } y^{(1)} = 400$$

We call θ **parameters**, $x^{(i)}$ is the input or the **features**, and the output or **target** is $y^{(i)}$. To be clear,

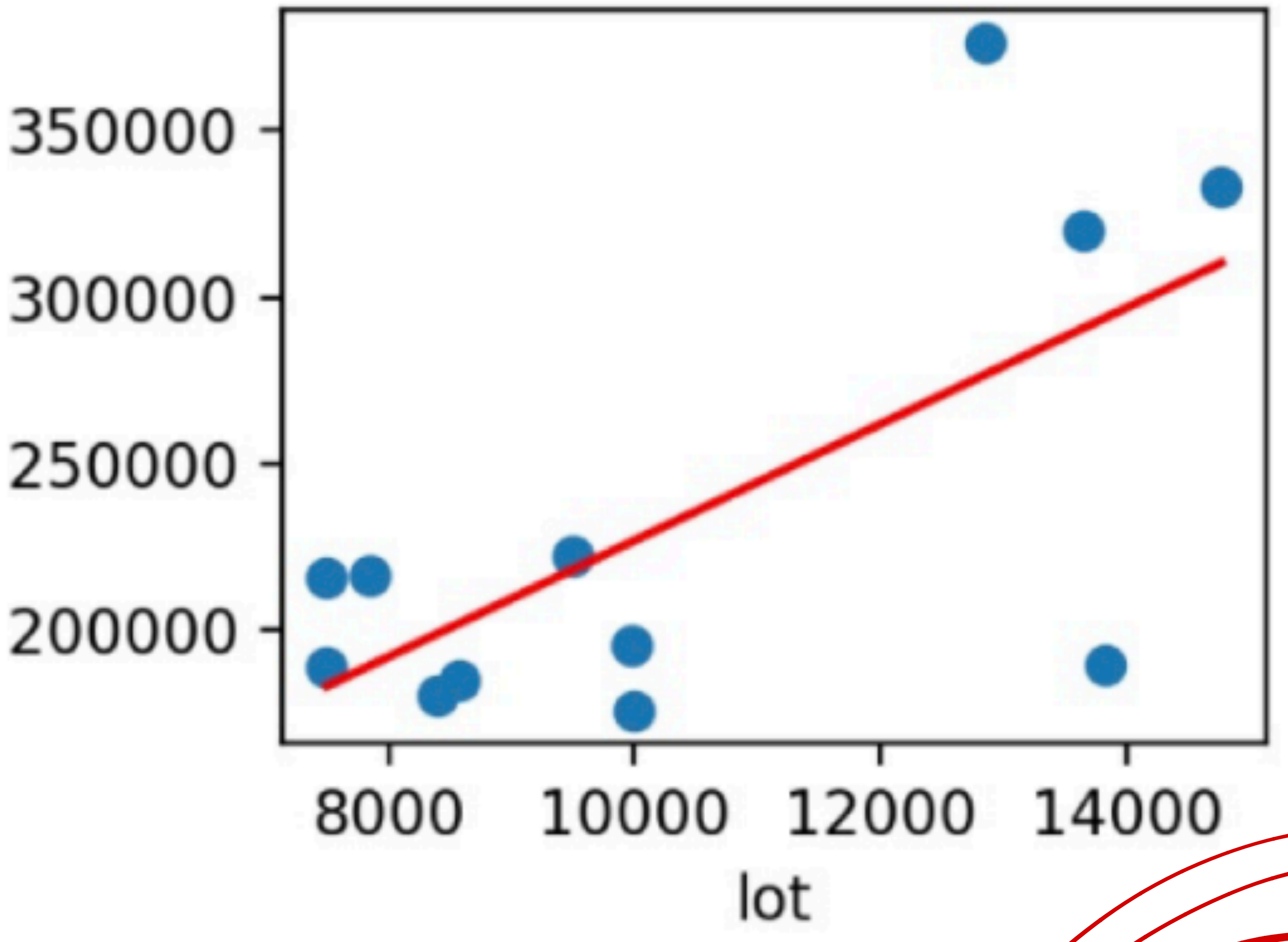
(x, y) is a training example and $(x^{(i)}, y^{(i)})$ is the i^{th} example.

θ_0

$x_0 = 1$

We have n examples. There are d features. $x^{(i)}$ and θ are d+1 dimensional (since $x_0 = 1$)

Vector Notation of Prediction

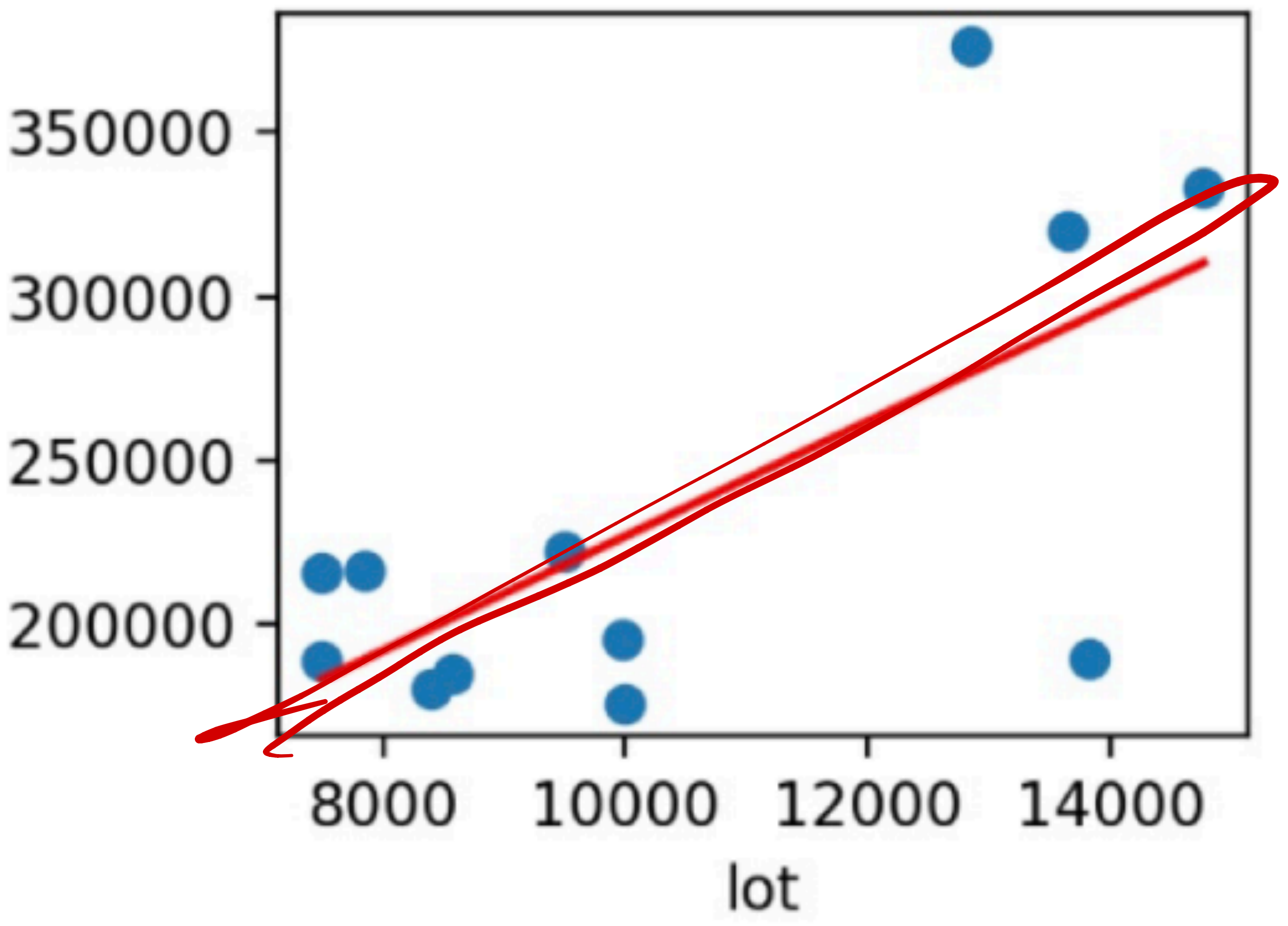


$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

inner product

$x^T\theta$ → Scalar

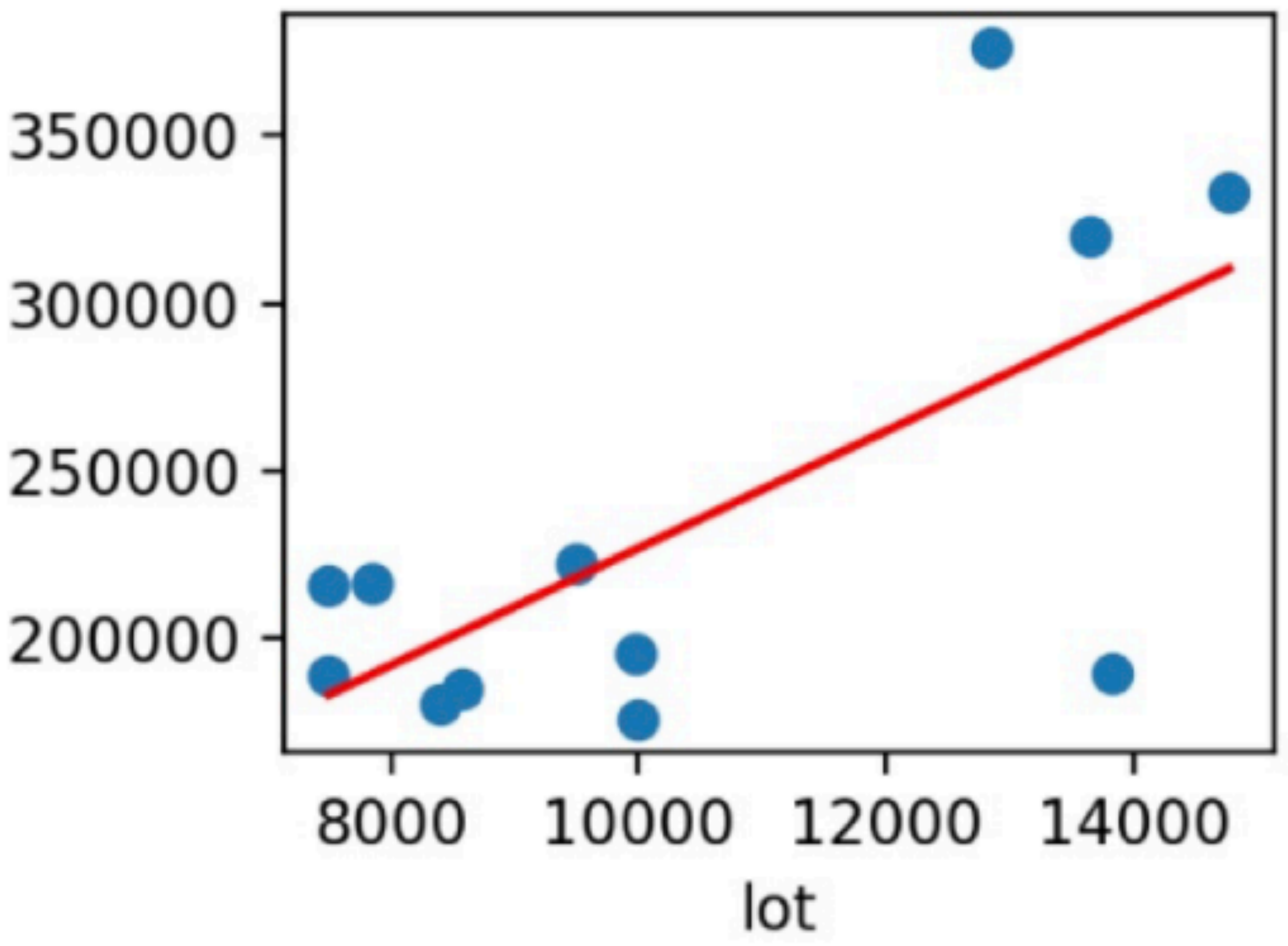
Vector Notation of Prediction



$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T \theta$$

We want to choose θ so that $h_\theta(x) \approx y$

Loss Function



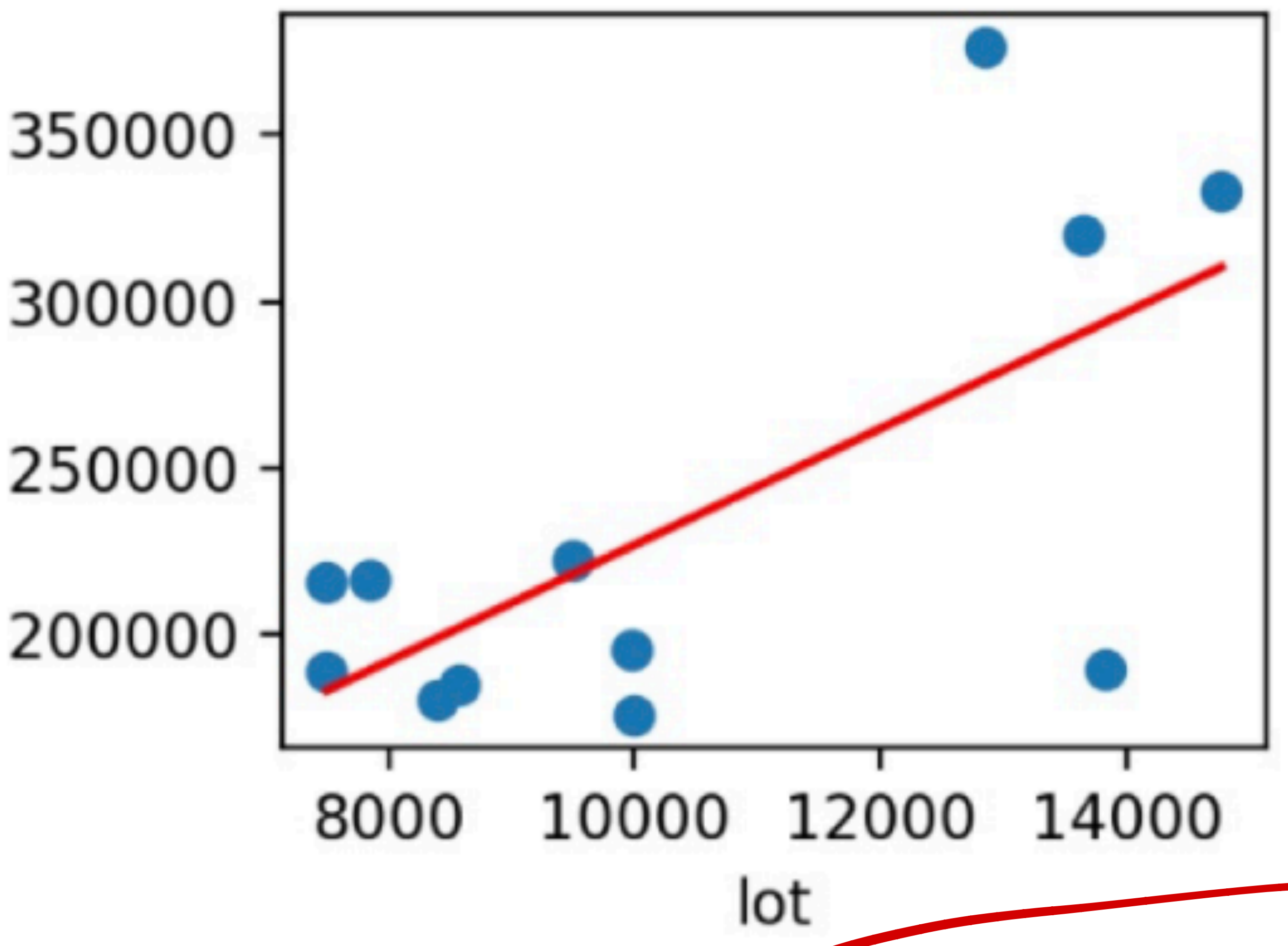
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

We want to choose θ so that $h_\theta(x) \approx y$

How to quantify the deviation of $h_\theta(x)$ from y

metric

Least Squares



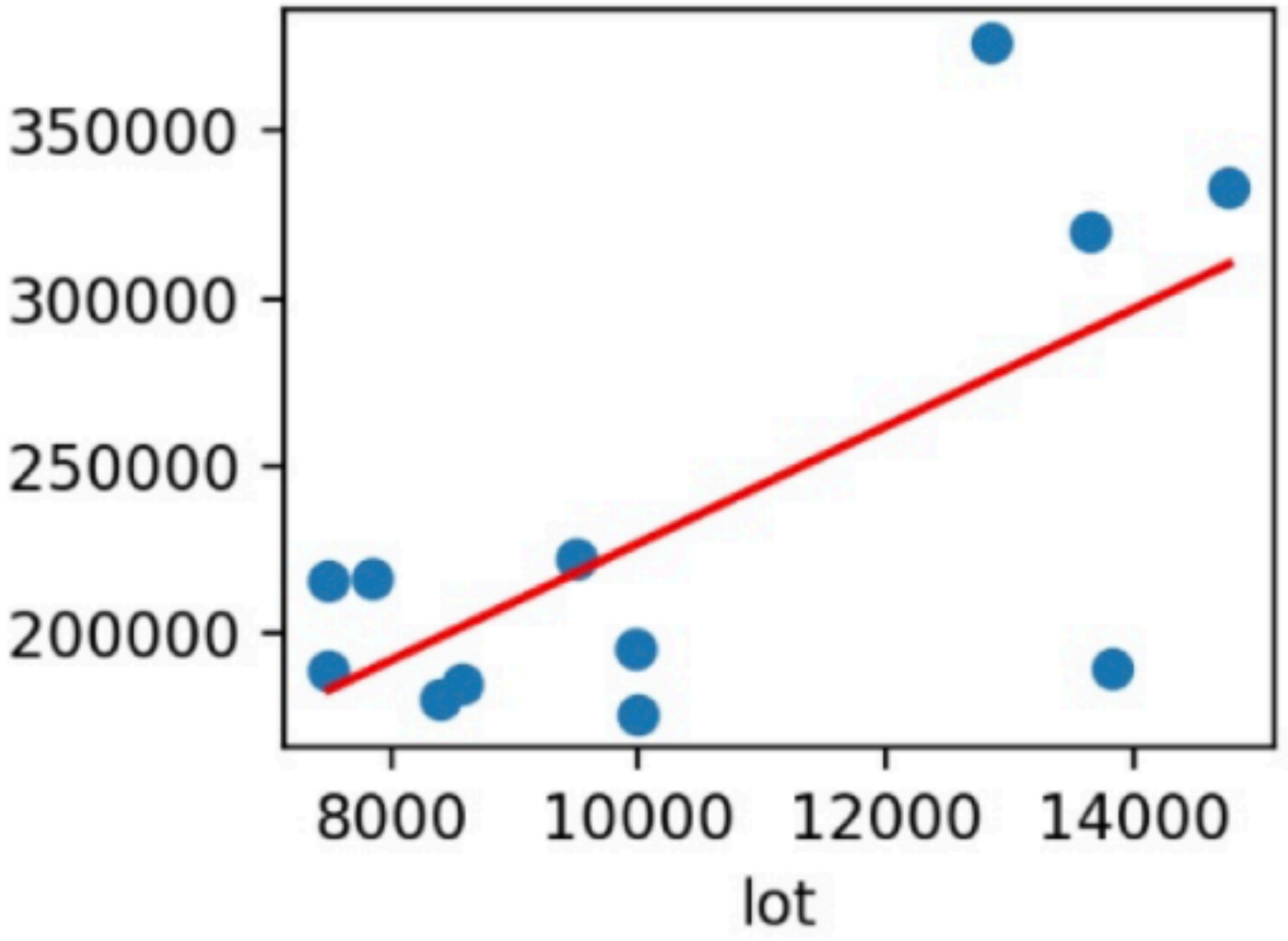
$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

n

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

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

Least Squares



$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

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

Choose

$$\theta = \underset{\theta}{\operatorname{argmin}}\, J(\theta).$$

Solving Least Square Problem

Direct Minimization

$\vec{\theta}$

$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta \qquad J(\theta) = \frac{1}{2}\sum_{i=1}^{n}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

Choose

$$\theta = \underset{\theta}{\operatorname{argmin}}\, J(\theta).$$

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

$\frac{d J_\theta}{d\theta} = 0$

Solving Least Square Problem

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T\vec{y} - \vec{y}^T(X\theta) + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - \vec{y}^T(X\theta) - \vec{y}^T(X\theta)\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - 2(X^T\vec{y})^T\theta\right) \\
&= \frac{1}{2}\left(2X^TX\theta - 2X^T\vec{y}\right) \\
&= X^TX\theta - X^T\vec{y}
\end{aligned}$$

$$X^TX\theta = X^T\vec{y} \qquad \boxed{\theta = (X^TX)^{-1}X^T\vec{y}.}$$

Solving Least Square Problem

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T\vec{y} - \vec{y}^T(X\theta) + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - \vec{y}^T(X\theta) - \vec{y}^T(X\theta)\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - 2(X^T\vec{y})^T\theta\right) \\
&= \frac{1}{2}\left(2X^TX\theta - 2X^T\vec{y}\right) \\
&= X^TX\theta - X^T\vec{y}
\end{aligned}$$

Normal equations $\boxed{X^TX\theta = X^T\vec{y}}$ $\qquad$ $\boxed{\theta = (X^TX)^{-1}X^T\vec{y}.}$

rank $(X^T X) \le \min(\text{rank}(X^T), \text{rank}(X)) = \text{rank}(X)$

$d \times d$

Solving Least Square Problem

neccesry condition:

$\text{rank}(X) \ge d$

$X \in {}^{n \times d}$

$$
\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \\
&= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T\vec{y} - \vec{y}^T(X\theta) + \vec{y}^T\vec{y}\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - \vec{y}^T(X\theta) - \vec{y}^T(X\theta)\right) \\
&= \frac{1}{2}\nabla_\theta \left(\theta^T(X^TX)\theta - 2(X^T\vec{y})^T\theta\right) \\
&= \frac{1}{2}\left(2X^TX\theta - 2X^T\vec{y}\right) \\
&= X^TX\theta - X^T\vec{y}
\end{aligned}
$$

$\text{rank}(X) \le \min(n, d)$

$[n < d]$

n: # samples

d: # features

invertible $\Longleftrightarrow$ full rank

Normal equations $X^TX\theta = X^T\vec{y}$ $\qquad$ $\theta = (X^TX)^{-1}X^T\vec{y}.$

When is X^TX invertible? What if it is not invertible?

$$
n\left\{
\overbrace{
\begin{array}{l}
\theta_1 + \theta_2 x_2 + \cdots - - = 0 \\
\theta_1 + \cdots \\
\theta_1 + \cdots
\end{array}
}^{d+1}
\right.
$$

$n < d+1$ sufficient $\rightarrow$ $\text{rank}(x) < d+1$

$$\begin{array}{ccc} x_1 & x_2 & x_3 \\ \hline 0 & 1 & 1 \\ 2.2 & 3.3 & 3.3 \\ 4 & 5.5 & 5.5 \end{array}$$

$X \in R^{n \times d}$

Why Least-Square Loss Function?

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

$\sum_{i=1}^{n} | h_\theta(x) - y |$

$\sum_{i=1}^{n} | h_\theta(x) - y |^4$

Maximum Likelihood estimation

MLE

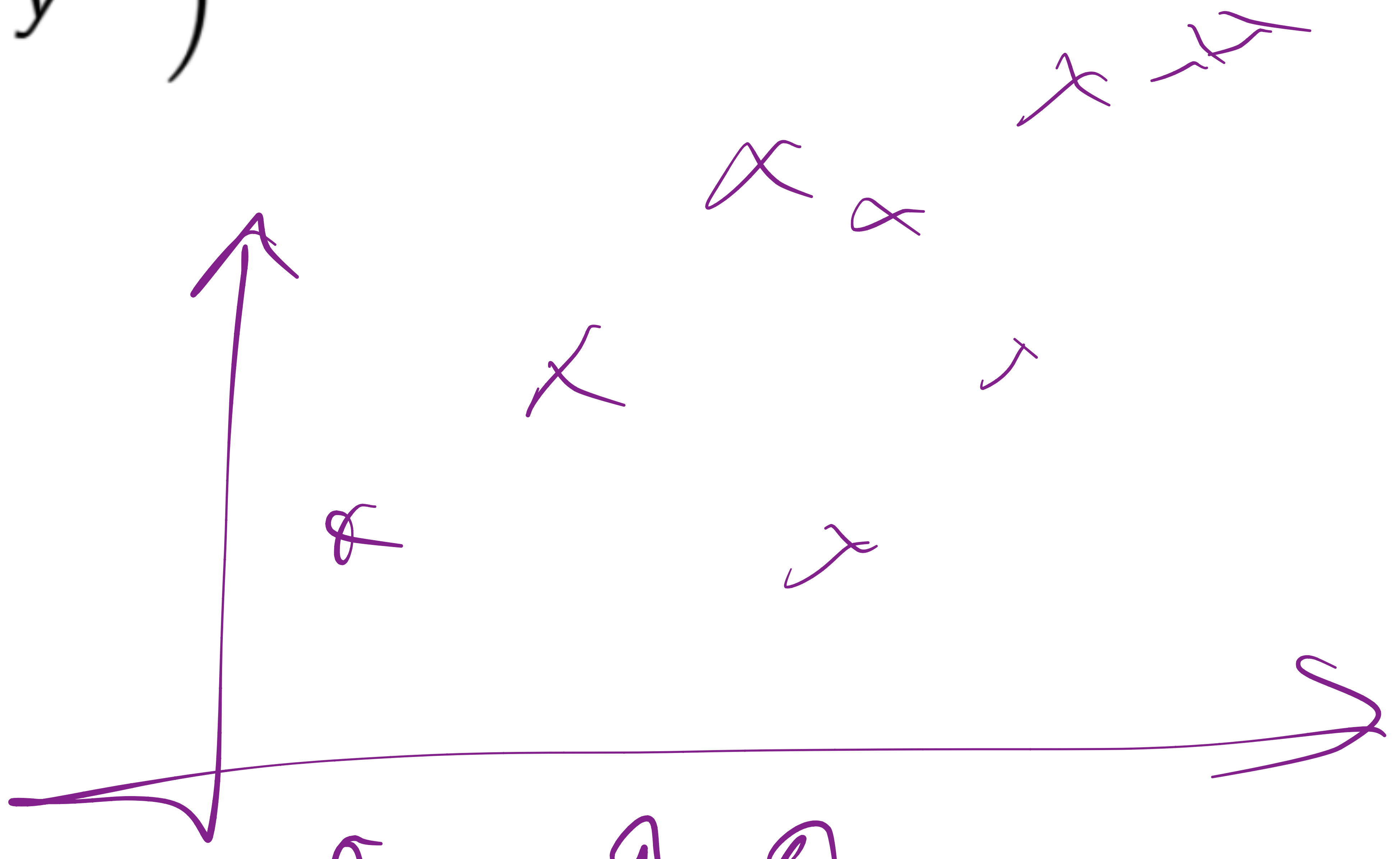
Why Least-Square Loss Function?

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

Assume

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

random noise



$F = ma^2$

Why Least-Square Loss Function?

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

Assume

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

x, y: random variable

Why Least-Square Loss Function?

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

Assume

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

x, y: random variable

ϵ: deviation of prediction from the truth, Gaussian random variable

Why Least-Square Loss Function?

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

Assume

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

x, y: random variable

ϵ: deviation of prediction from the truth, Gaussian random variable

$x^{(i)}, y^{(i)}$: observations, or the data

Y $\quad X \quad$ $\underline{\varepsilon}$

Why Least-Square Loss Function?

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

Assume

$$y^{(i)} = \theta^T x^{(i)} + \epsilon^{(i)}$$

x, y: random variable

ϵ: deviation of prediction from the truth, Gaussian random variable

$x^{(i)}, y^{(i)}$: observations, or the data

$\epsilon^{(i)}$: the actual prediction error of the i_{th} example, sampled from the Gaussian distribution, IID (independently and identically distributed)

Why Least-Square Loss Function?

Why Least-Square Loss Function?

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

Why Least-Square Loss Function?

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Why Least-Square Loss Function?

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$\begin{aligned} p(\vec{y}|X;\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right) \end{aligned}$$

Why Least-Square Loss Function?

$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right)$$

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

$$p(\vec{y}|X;\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)};\theta)$$

Function of θ $= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$

Why Least-Square Loss Function?

Why Least-Square Loss Function?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Why Least-Square Loss Function?

$$L(\theta) = \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)$$

Likelihood Function

Gausissan

Why Least-Square Loss Function?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood Function

What is a reasonable guess of θ?

Why Least-Square Loss Function?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood Function

What is a reasonable guess of θ?

Maximize the probability of Y's happening!

$\underset{\theta}{\operatorname{argmax}}\ L(\theta) = \underset{\theta}{\operatorname{argmax}}\ \log L(\theta)$

Maximum Likelihood Estimation (MLE)

$$
\begin{aligned}
\ell(\theta) &= \log L(\theta) \\
&= \log \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= \sum_{i=1}^{n} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right) \\
&= \underbrace{n \log \frac{1}{\sqrt{2\pi}\sigma}}_{\text{constant}} - \frac{1}{\sigma^2} \cdot \frac{1}{2} \underbrace{\sum_{i=1}^{n} (y^{(i)}-\theta^T x^{(i)})^2}_{\text{least square}}.
\end{aligned}
$$

Why MLE?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood Function

What is a reasonable guess of θ?

Maximize the probability of Y's happening?

Why MLE?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood Function

What is a reasonable guess of θ?

Maximize the probability of Y's happening?

Maximizing likelihood estimation -> $\hat{\theta}$

Why MLE?

$$
\begin{aligned}
L(\theta) &= \prod_{i=1}^{n} p(y^{(i)} \mid x^{(i)}; \theta) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y^{(i)} - \theta^T x^{(i)})^2}{2\sigma^2}\right)
\end{aligned}
$$

Likelihood Function

What is a reasonable guess of θ?

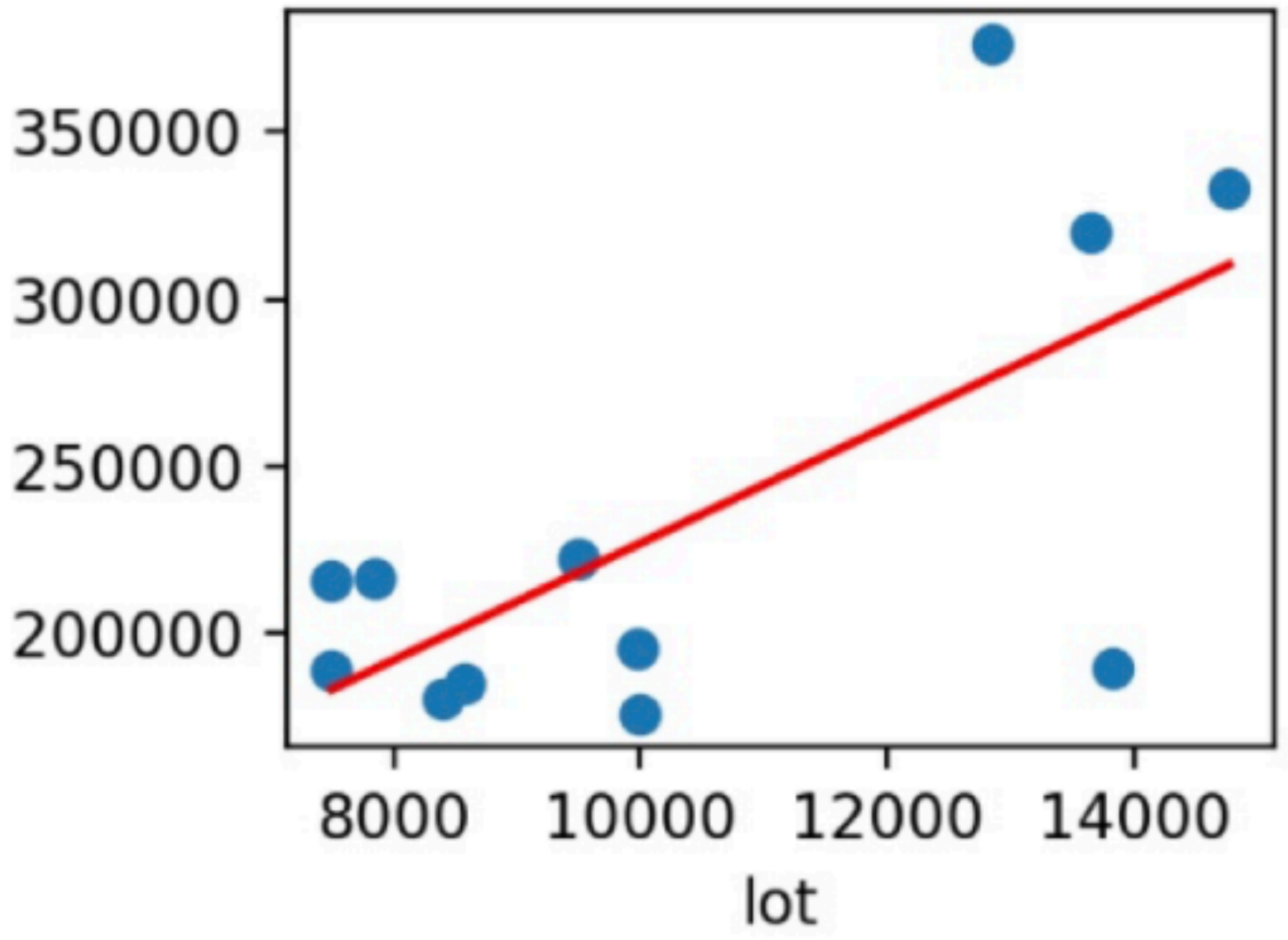
Maximize the probability of Y's happening?

Maximizing likelihood estimation -> $\hat{\theta}$

Ground-truth θ^*

$n \to \infty \quad \hat{\theta} \longrightarrow \theta^*$

Another Solution — Gradient Descent



$$h_\theta(x) = \sum_{j=0}^{d} \theta_j x_j = x^T\theta$$

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

Choose

$$\theta = \underset{\theta}{\operatorname{argmin}}\, J(\theta).$$

Gradient Descent

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

Gradient Descent

step size

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

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

direction

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

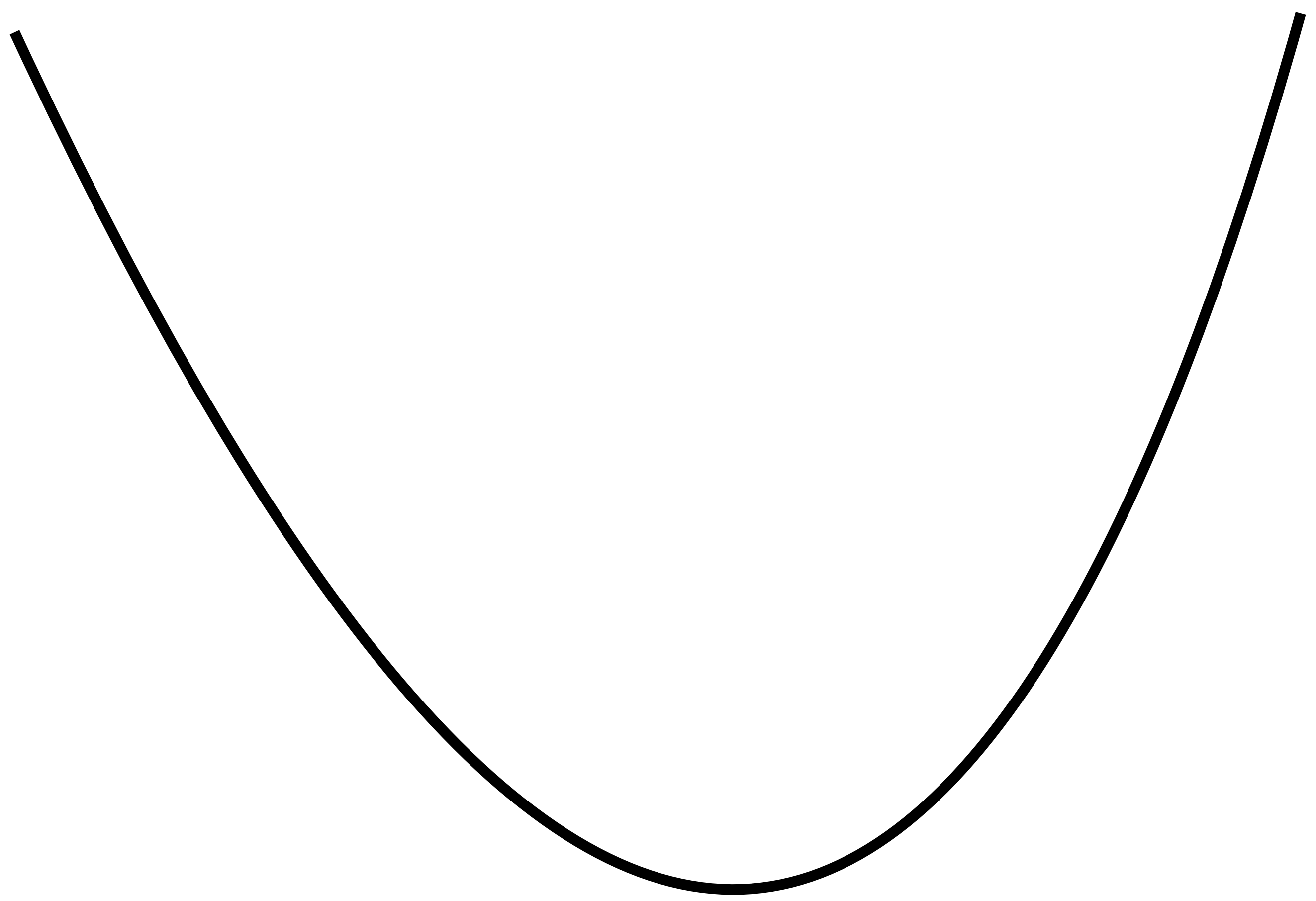
derivative

gradient

Gradient Descent

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

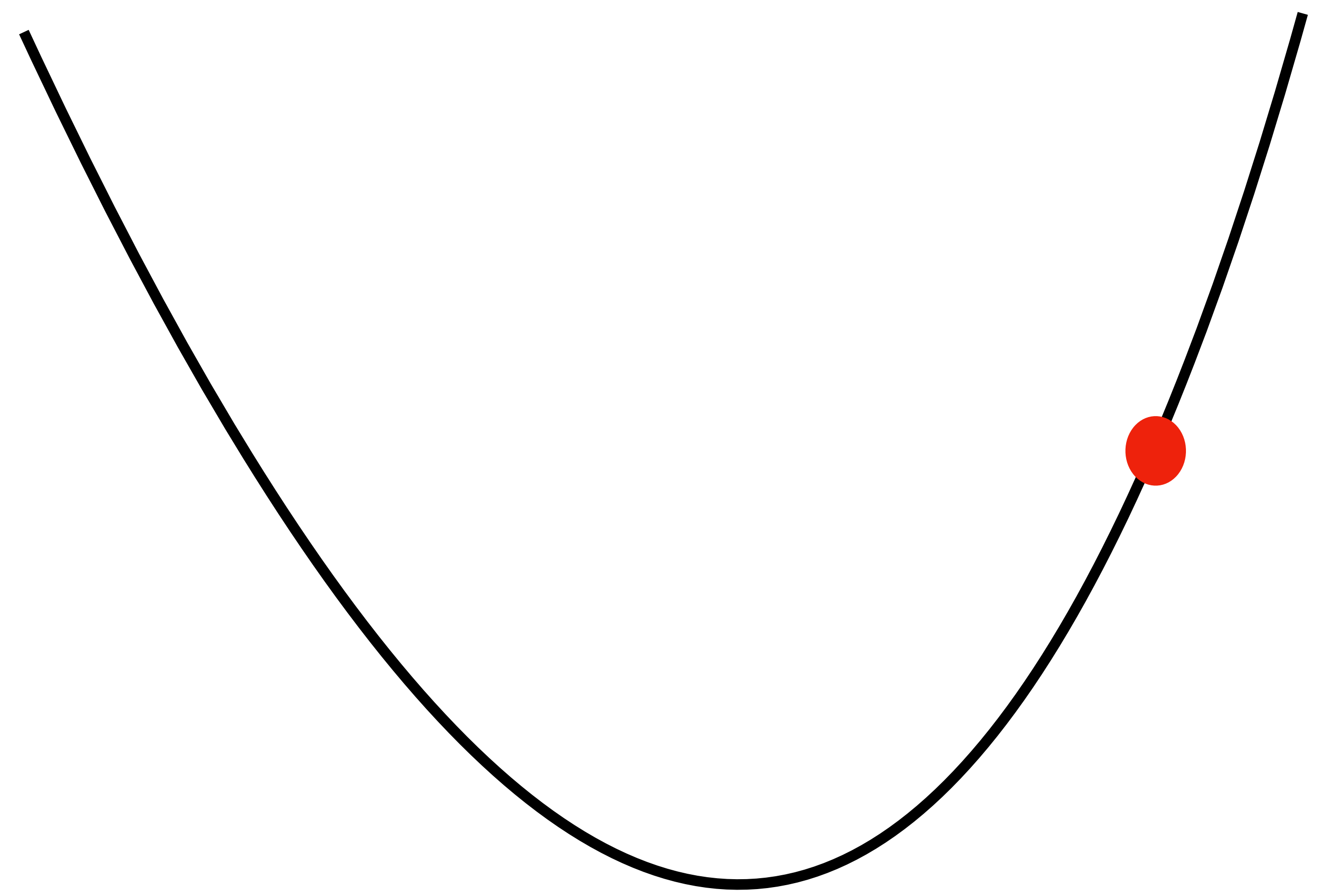
$$\theta_j := \theta_j - \alpha\frac{\partial}{\partial\theta_j}J(\theta)$$



Gradient Descent

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

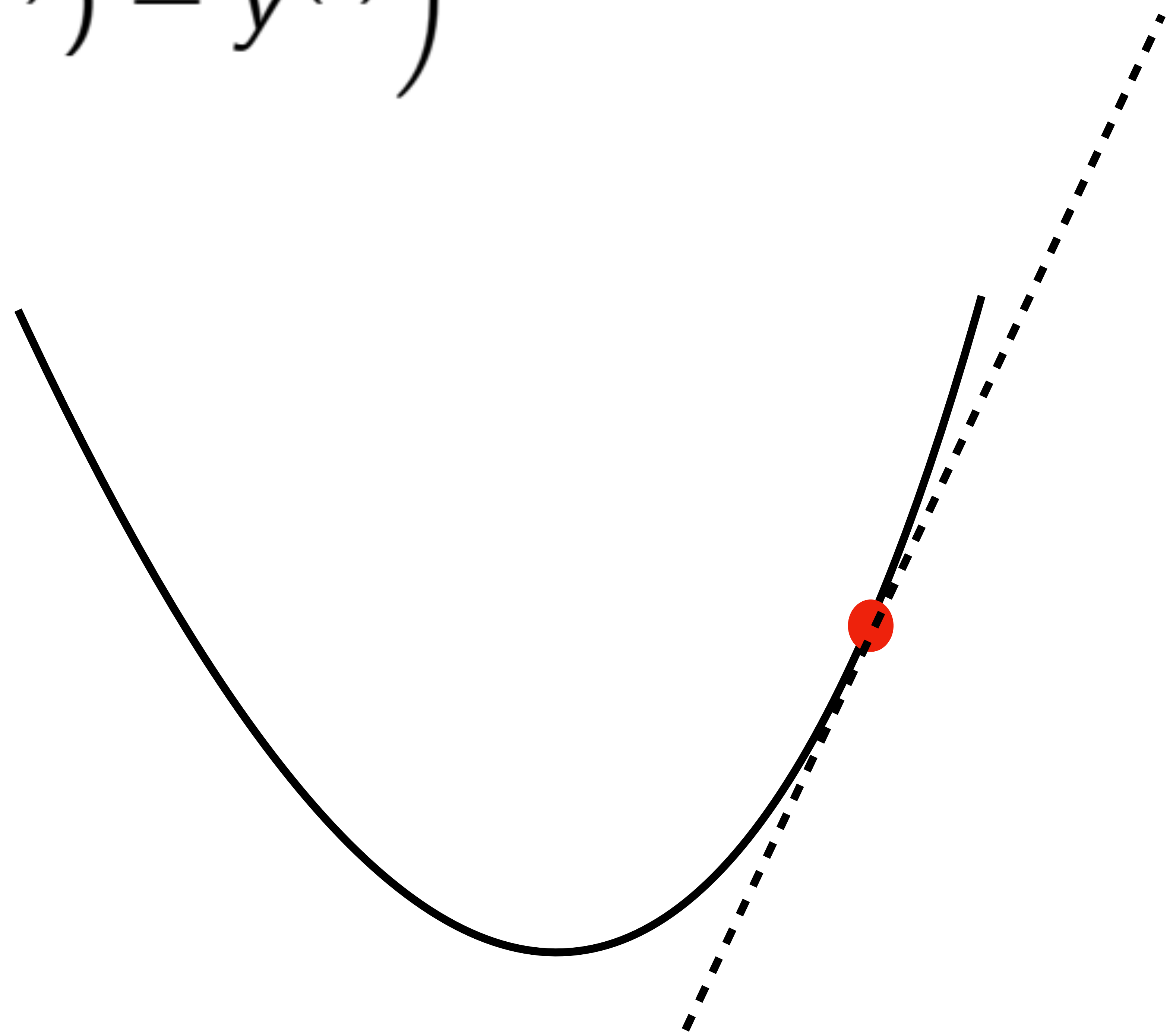
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$



Gradient Descent

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

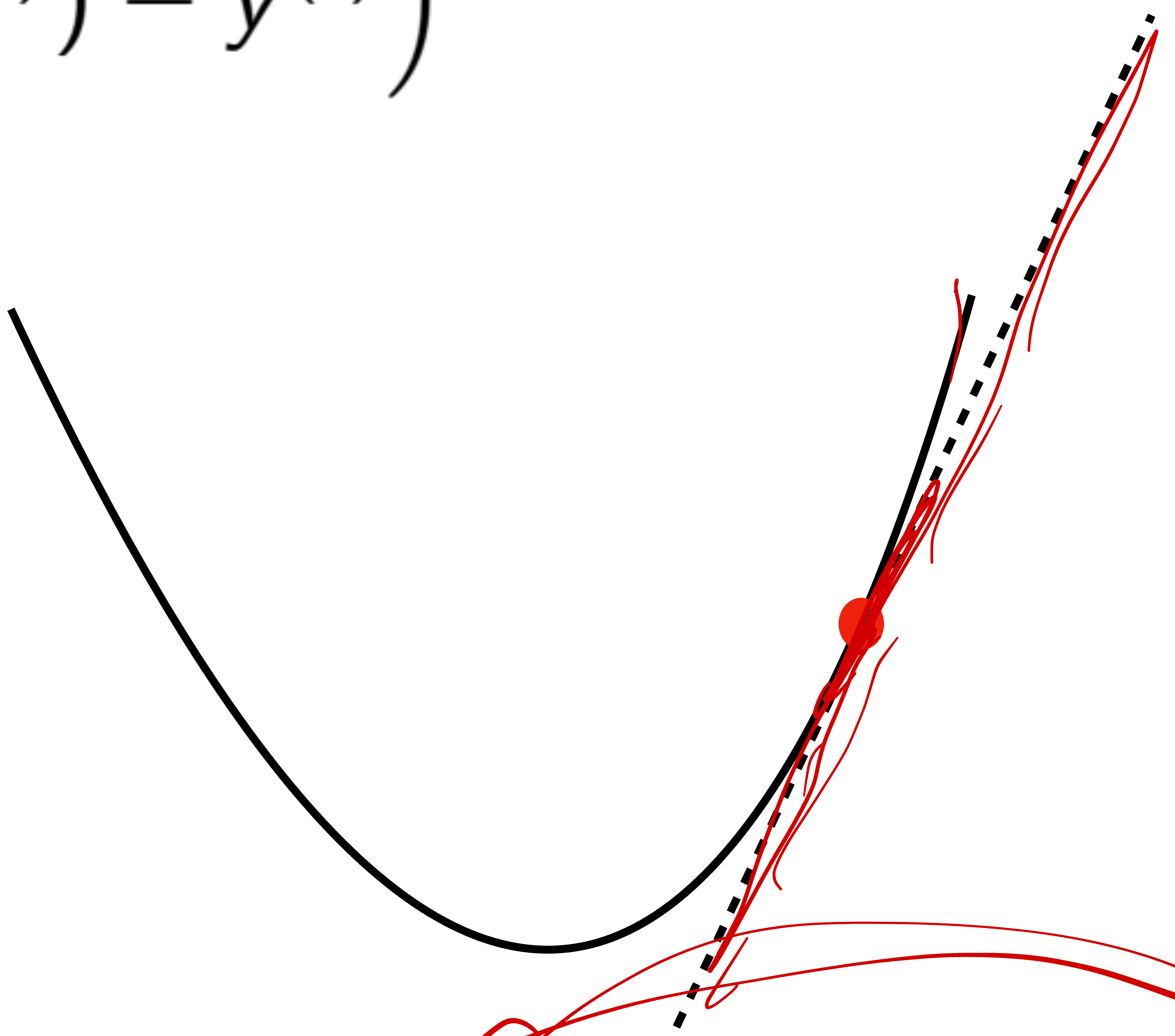
$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$



Gradient Descent

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

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



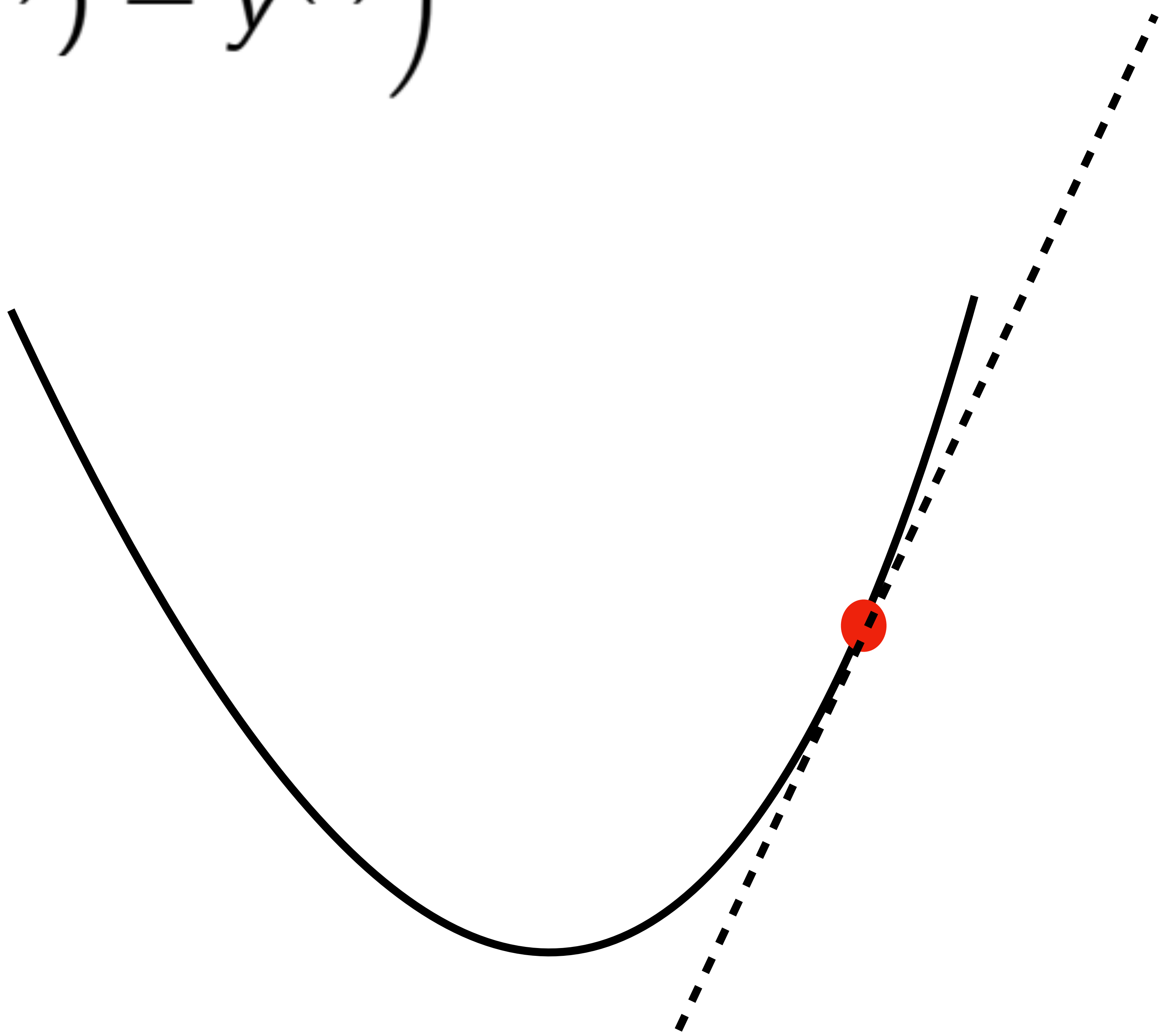
The direction of the steepest decrease of J

Gradient Descent

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

Learning Rate

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



The direction of the steepest decrease of J

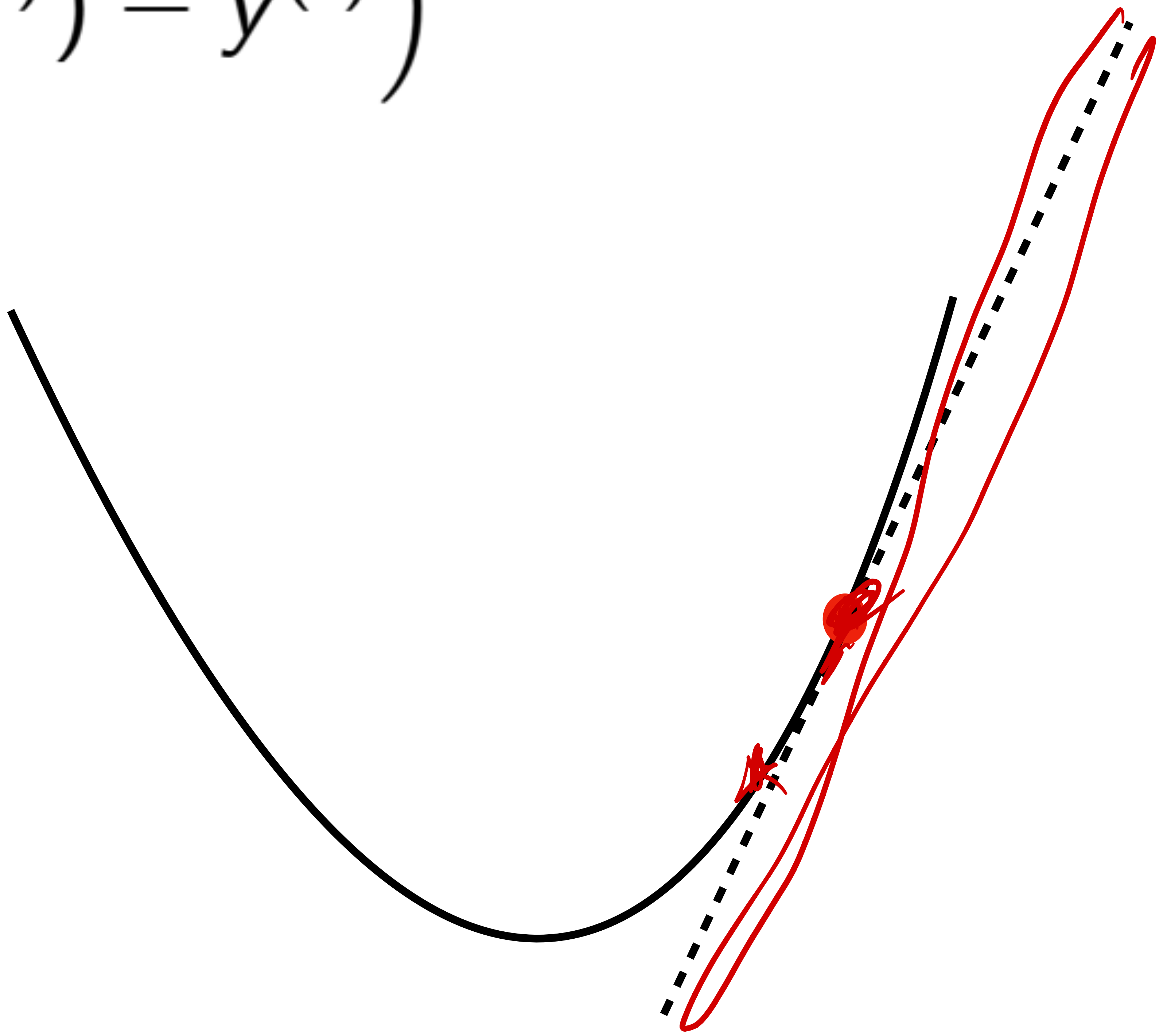
Gradient Descent

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

Learning Rate

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

This update is simultaneously performed for all values of j = 0,...,d.



The direction of the steepest decrease of J

Gradient Descent

For a single training example:

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{d} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}$$

Gradient Descent

For a single training example:

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} J(\theta) &= \frac{\partial}{\partial \theta_j} \frac{1}{2} \left(h_\theta(x) - y\right)^2 \\
&= 2 \cdot \frac{1}{2} \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(h_\theta(x) - y\right) \\
&= \left(h_\theta(x) - y\right) \cdot \frac{\partial}{\partial \theta_j} \left(\sum_{i=0}^{d} \theta_i x_i - y \right) \\
&= \left(h_\theta(x) - y\right) x_j
\end{aligned}$$

LMS (Least Mean Square) Update Rule

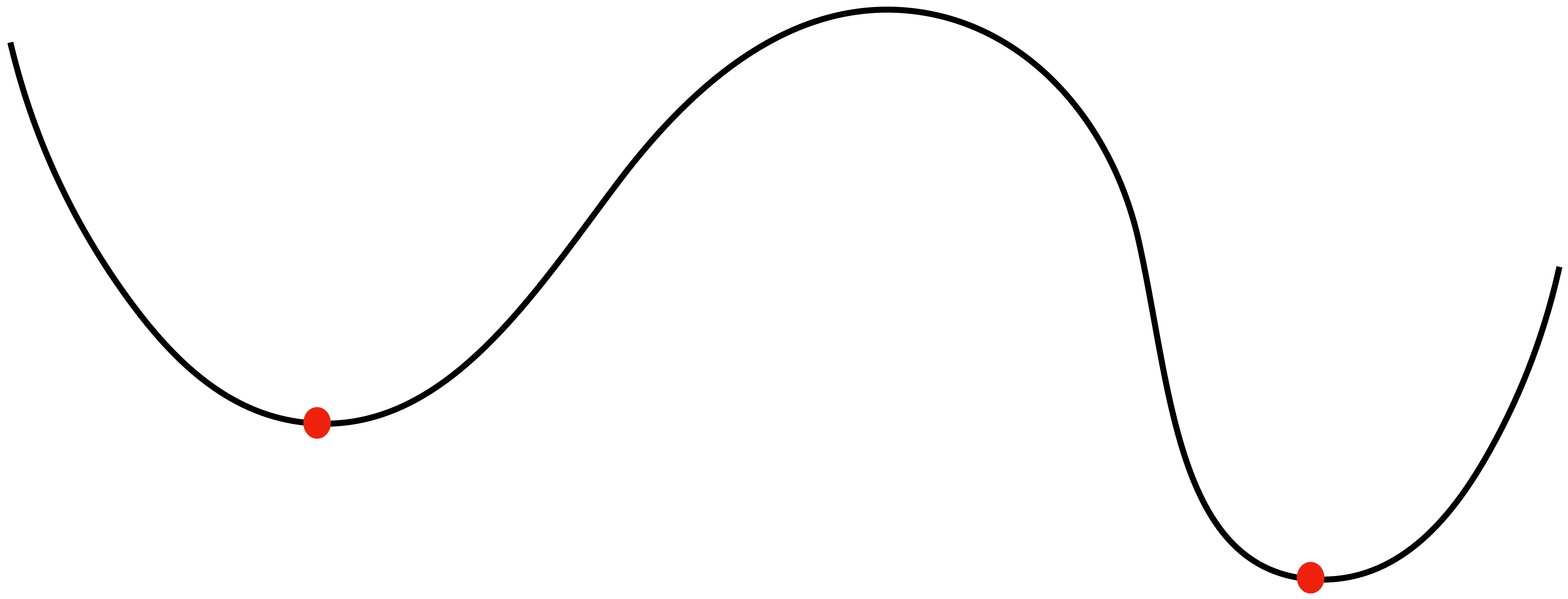
Batch Gradient Descent

For a multiple training examples:

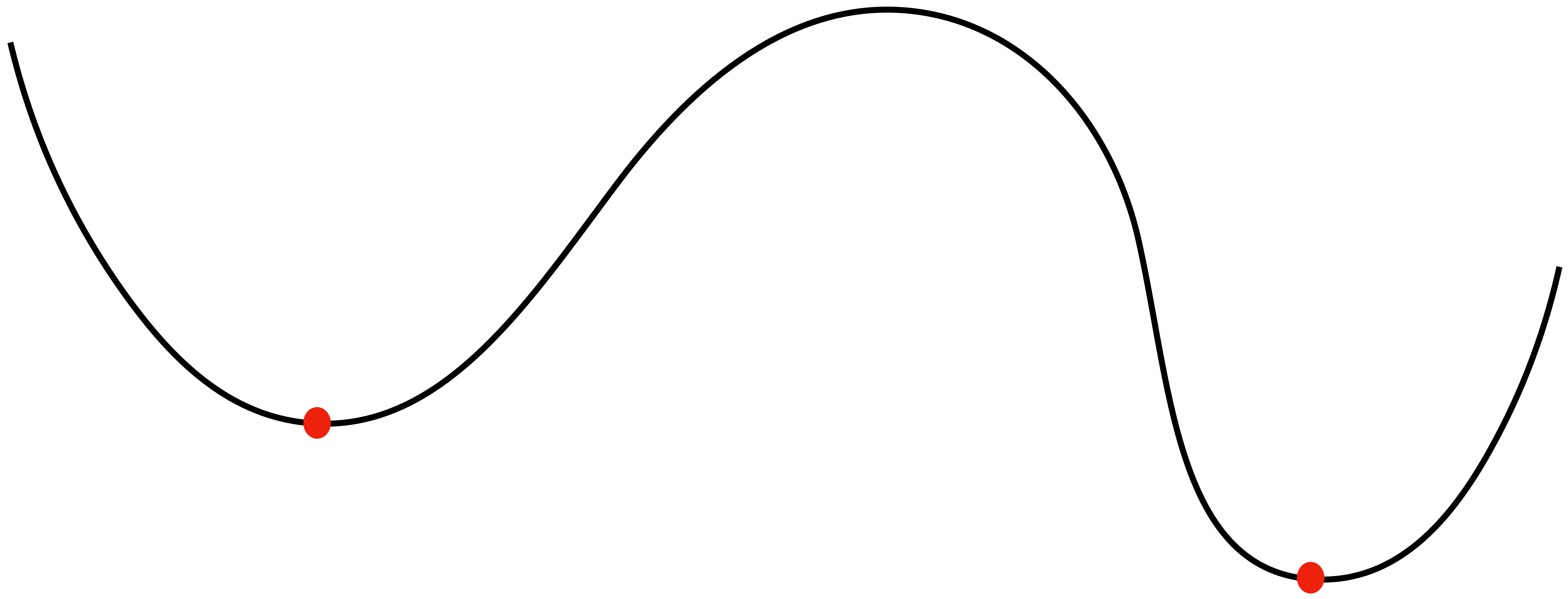
$$\theta_j := \theta_j + \alpha \sum_{i=1}^{n} \left(y^{(i)} - h_\theta(x^{(i)})\right) x_j^{(i)}$$

Repeat until convergence

Local Minimum



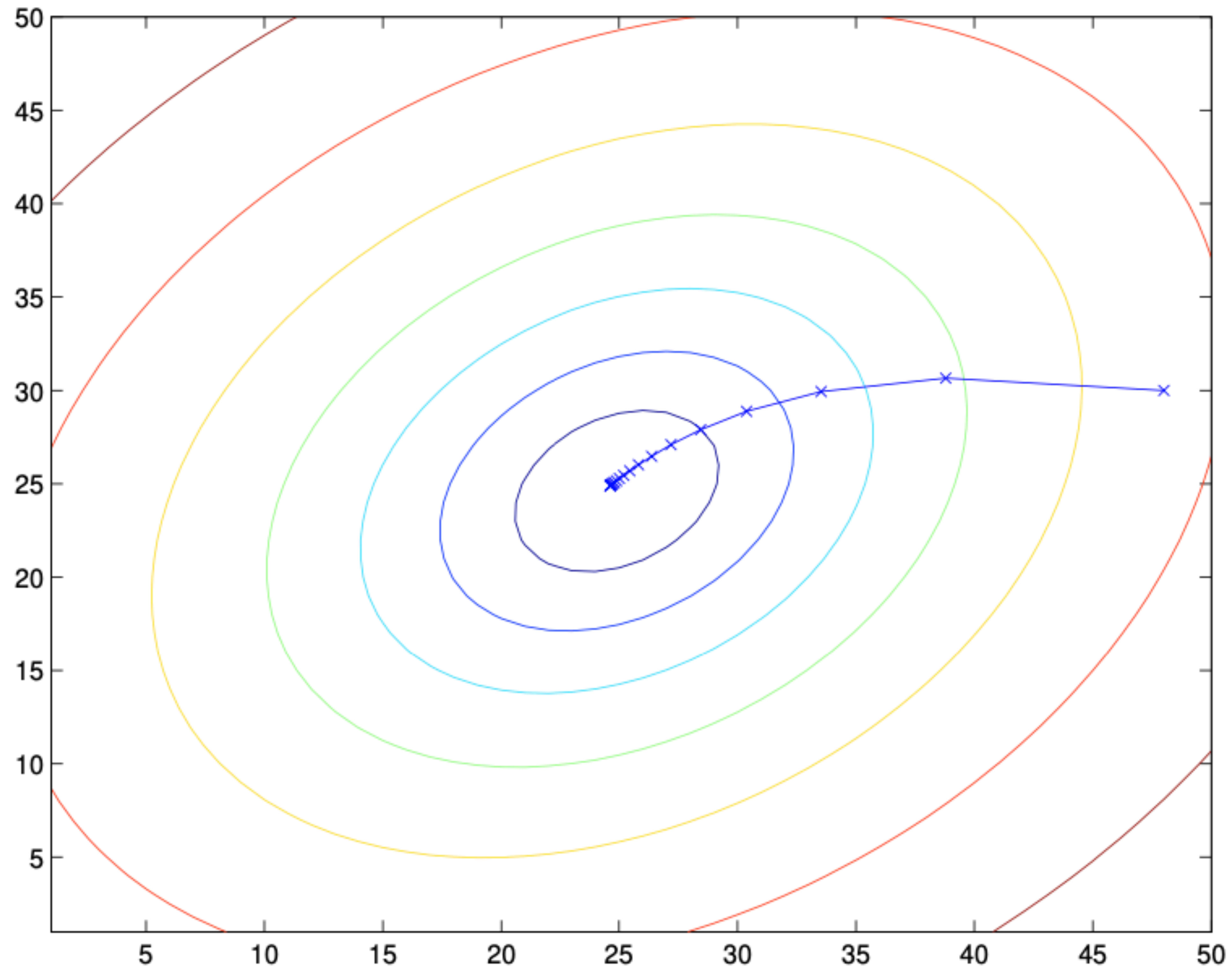
Local Minimum



For least square optimization, are we likely to get local minima rather than the global minima through gradient descent?

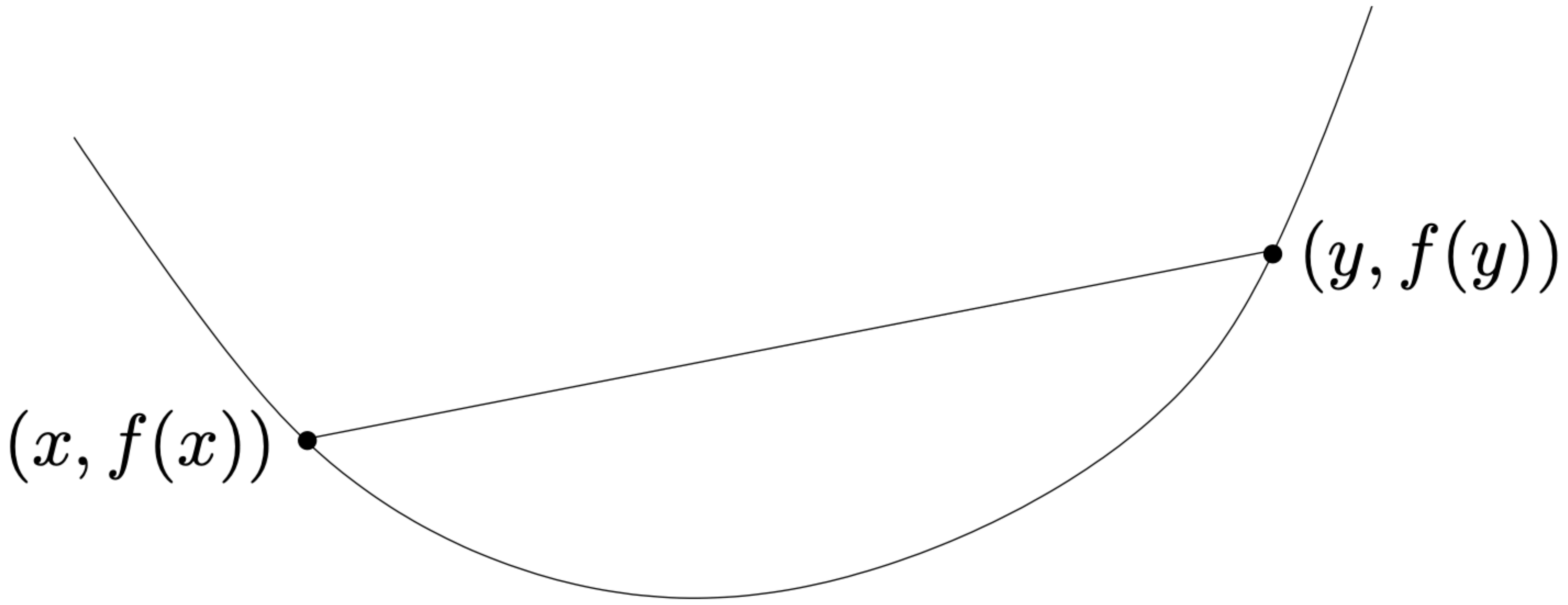
J is a convex quadratic function

There is only one local minima for J



Convex Function

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \quad \text{for } 0 \leq t \leq 1$$



Thank You!
Q & A