From Empirical Laws to Linear Regression

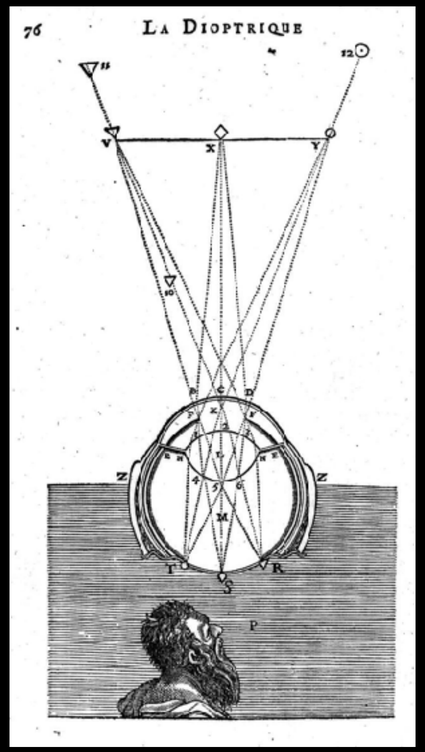
Lecture 3

Prepared by: Joseph Bakarji

Physics from data

Galileo

Kepler



Descartes

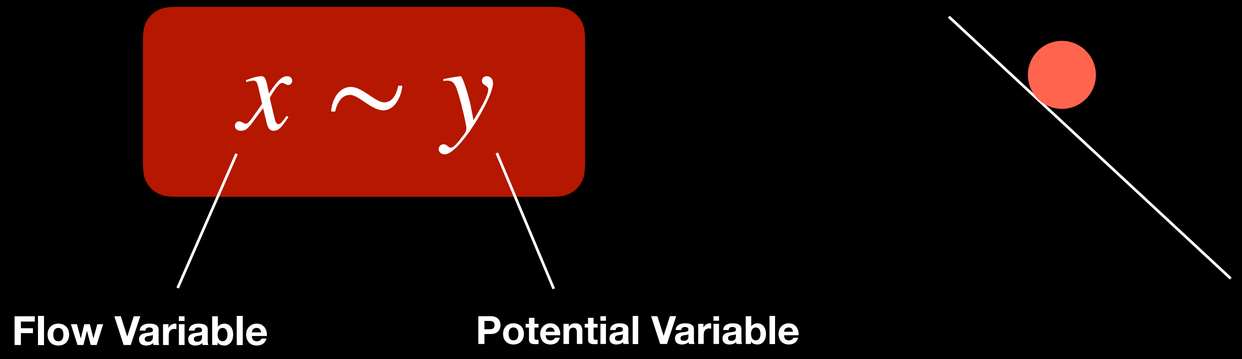
Finding ***constants*** of nature that ***generalize*** in space and time

Galileo with a Computer

Homework Assignment for Next Week

The discovery of empirical laws

- The standard model was



- The art of scientific discovery was in finding **which variables** x and y follow that law and setting up an experiment for it
- The craft involved: designing an experimental setup that
  - **Quantifies** very specific properties of the world — a controlled environment
  - The resulting relationship between a controllable input and a changing output is **linear**
  -

Laws are linear

Pascal's law (1653)

$$\Delta p = \rho g \Delta h$$

Hooke's law (1678)

$$F = -kx$$

Newton's law of viscosity (1701)

$$\tau = \mu \frac{du}{dy}$$

Ohm's law (1781)

$$I = V/R$$

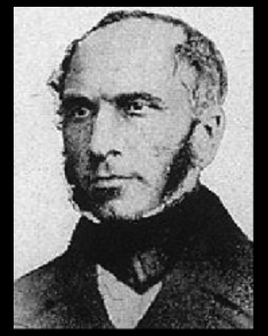
Fourier's law (1822)

$$q = -k \frac{dT}{dx}$$

Fick's law (1855)

$$J = -D \frac{dC}{dx}$$

Darcy's law (1856)

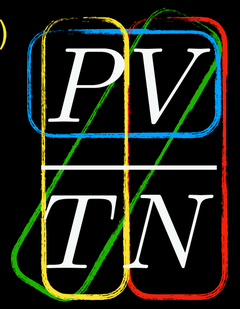


$$Q = \frac{kA}{\mu L} \Delta p$$

Ideal gas law (1834)

Amonton's law (1808)

Boyle's law (1662)



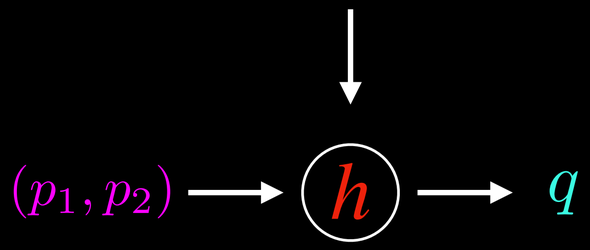
$$\frac{PV}{TN} = k_B$$

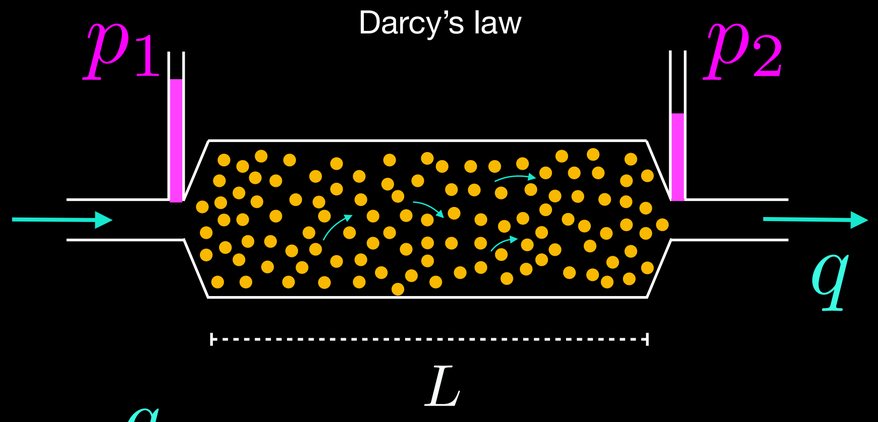
Charles's law (1787)

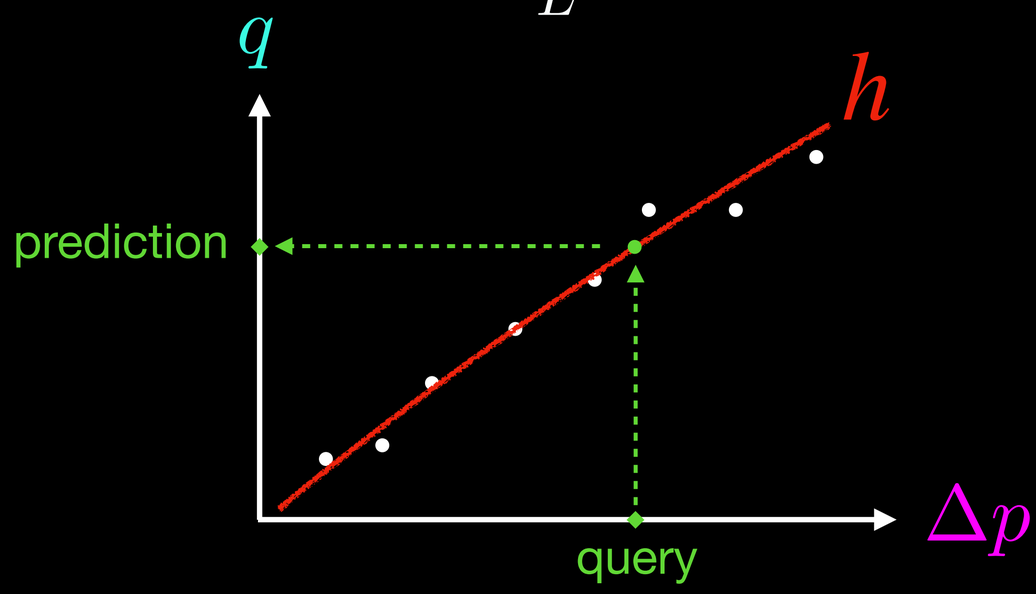
Avogadro's law (1811)

From experiment to Law

p_1	p_2	q
1.3	1.0	22
1.6	1.5	23
3.4	2.4	46
4.8	3.5	67
6.7	4.5	83
…		
2.3	**1.4**	**?**







Empirical Law to Machine Learning

Training data: $\mathcal{D}_{\text{train}}$

	p_1	p_2	q
Example →	1.3	1.0	22
	1.6	1.5	23
	3.4	2.4	46
	4.8	3.5	67
	6.7	4.5	83
	…		

Training

Feature extractor

$(p_1, p_2) \rightarrow \phi \rightarrow \Delta p \rightarrow h \rightarrow q$

Predictor *Labels*

Empirical Law

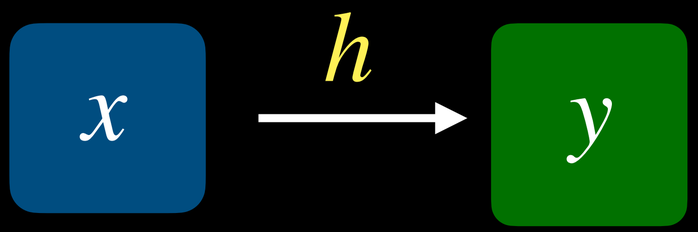
- **Hypothesis:** linear relationship $\Delta p \sim q$
- **Objective Function:** How linear is the relationship?
- **Optimization:** What's the proportionality constant?

Machine Learning

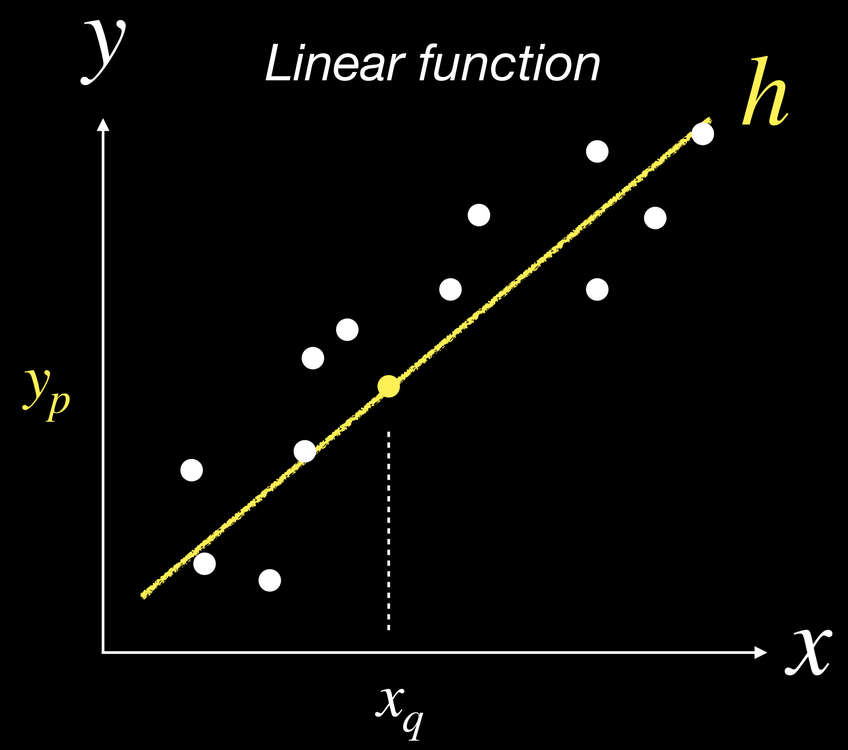
- **Hypothesis:** Which predictors f are possible?
- **Objective Function:** How good is the predictor?
- **Optimization:** How can we find the best predictor?

Given new input, what's the output?

Assume a linear hypothesis



$$h(x) = ax + b$$

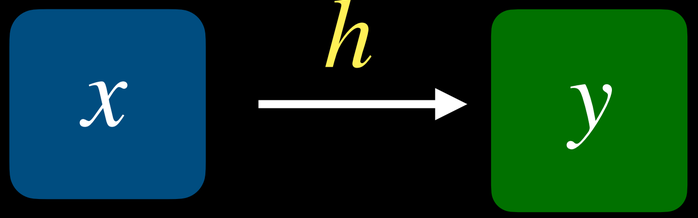


What are the best a and b that fit the data?

a, b are fitting parameters

Given new input, what's the output?

Assume a linear hypothesis

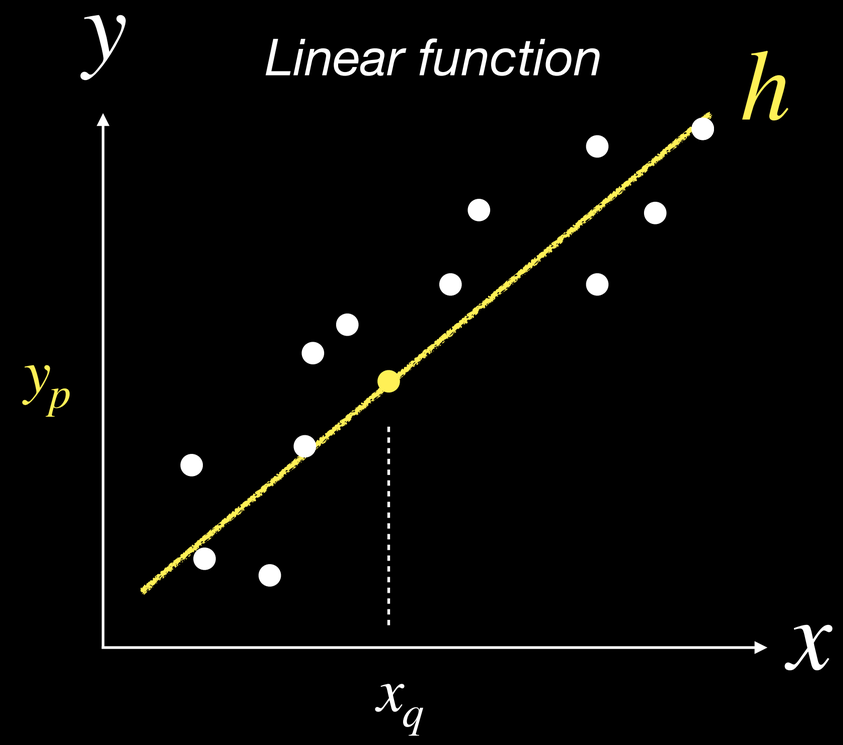


$$h_\theta(x) = \theta_0 + \theta_1 x$$

$$h_\theta(x) = [\theta_0, \theta_1] \cdot [1, x]$$

Unknown parameters **Input features**

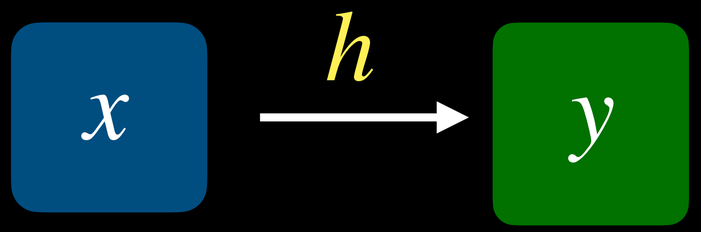
$$\theta \cdot \mathbf{x}$$



What's the best $\theta = [\theta_0, \theta_1]$?

What happens if we have more inputs?

Assume a linear hypothesis



$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots$$

$$h_\theta(x) = \underbrace{[\theta_0, \theta_1, \theta_2, \theta_3, \dots]}_{\text{weights } \theta} \cdot \underbrace{[1, x_1, x_2, x_3, \dots]}_{\text{features } x}$$

$$h_\theta(x) = \theta \cdot x = \theta^\top x$$

Inputs		Output
x_1	x_2	y
$x_1^{(1)}$	$x_2^{(1)}$	$y^{(1)}$
$x_1^{(2)}$	$x_2^{(2)}$	$y^{(2)}$
$x_1^{(3)}$	$x_2^{(3)}$	$y^{(3)}$
$x_1^{(4)}$	$x_2^{(4)}$	$y^{(4)}$
⋮	⋮	⋮

How do you find

The Best Model

that fits the data?

Titius–Bode law

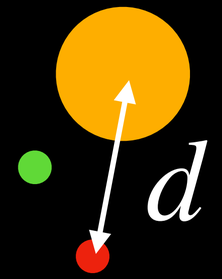
- Johan Bode (Law): “Supposing the distance of the Earth from the Sun to be divided into ten equal Parts, of these the distance of Mercury will be about four, of Venus seven, of Mars fifteen, of Jupiter fifty two, and that of Saturn ninety five.”
- **In 1766,** Johann Titius noticed a mathematical pattern in planetary distances

$$d = (3 * 2^n + 4)/10 \quad \text{AU}$$

- The formula predicted planets at

0.4, 0.7, 1.0, 1.6, 2.8, 5.2, 10.0 AU

- All known planets fit, except there was a gap at **2.8** AU between Mars and Jupiter!
- Astronomers believed a "missing planet" existed there



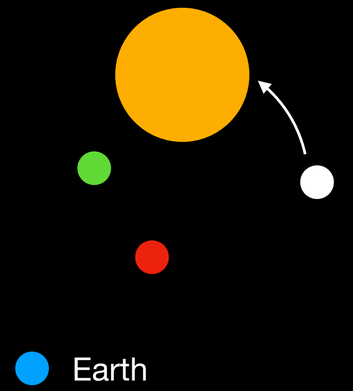
Earth

m	k	T–B rule distance (AU)	Planet	Semimajor axis (AU)	Deviation from prediction[1]
−∞	0	0.4	Mercury	0.39	−3.23%
0	1	0.7	Venus	0.72	+3.33%
1	2	1.0	Earth	1.00	0.00%
2	4	1.6	Mars	1.52	−4.77%
3	8	2.8	Ceres[2]	2.77	−1.16%
4	16	5.2	Jupiter	5.20	+0.05%
5	32	10.0	Saturn	9.58	−4.42%
6	64	19.6	Uranus	19.22	−1.95%
–	–	–	Neptune	30.07	–
7	128	38.8	Pluto[2]	39.48	+1.02%

https://en.wikipedia.org/wiki/Titius%E2%80%93Bode_law

The Unsolvable Problem

- Giuseppe Piazzi, observing from Sicily, spotted a faint object moving against the stars
- He tracked it for 41 days through only 9 degrees of arc, then it disappeared into the Sun's glare
- **Problem: How do you find it again with less than 1% of its orbit observed?**
- **(Noisy) Data:** 22 observations over 41 days (timestamp + 2 angles each)
- Unknown parameters: 6 orbital parameters (shape, size, orientation of ellipse)
- Laplace declared it **unsolvable!**

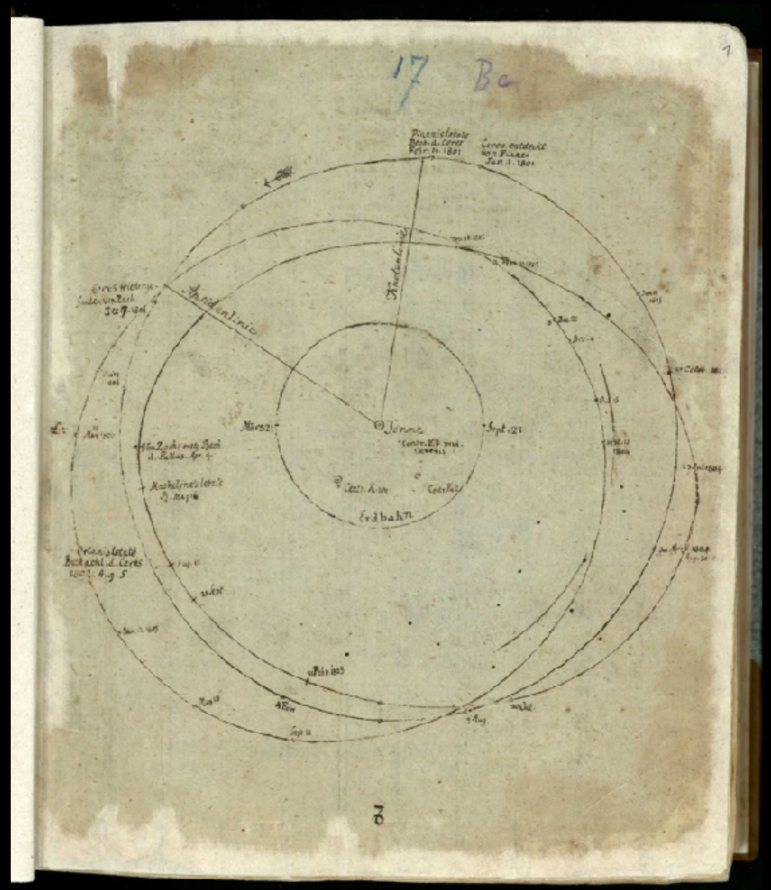


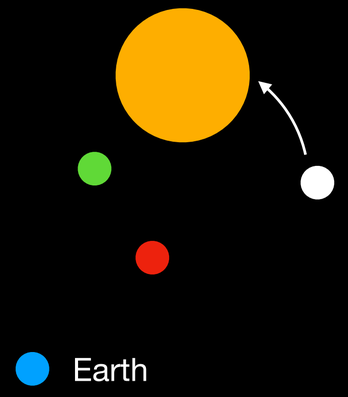
https://en.wikipedia.org/wiki/Titius%E2%80%93Bode_law

Gauss's Solution

- Gauss selected 3 well-spaced observations (Jan 1, Jan 21, Feb 11).
- Made **no assumptions** about eccentricity (unlike others who assumed circular or parabolic)
- Iteratively refined using all 22 observations!
- Strategy: find parameters that minimize the sum of squared errors across ALL observations.
- The innovation is that instead of fitting exactly through any single point, he found a curve that minimizes:

$$\sum_{i=1}^{22} (\text{observation - prediction})^2$$





https://www.actuaries.asn.au/research-analysis/gauss-least-squares-and-the-missing-planet

Gauss’s Solution

- On December 1801, Gauss published his prediction, and his position was 6° away from other astronomers' guesses
- Franz von Zach found Ceres on Dec 31, 1801, within 0.5° of Gauss's prediction, and was quote saying:

  “Ceres can never again be lost, since the ellipse of Dr. Gauss agrees so exactly with its location.”

- What equation did Gauss fit? Ellipses described by Kepler’s laws (who solved a hard ‘fitting’ problem!)

```
00001         A1801 01 01.82630 03 38 23.07 +16 17 25.5              MC004535
00001         A1801 01 02.82337 03 38 05.84 +16 20 51.5              MC004535
00001         A1801 01 03.82045 03 37 50.6  +16 24 21.2              MC004535
00001         A1801 01 04.81755 03 37 35.52 +16 27 50.7              MC004535
00001         A1801 01 10.80058 03 36 45.9  +16 50 36.9              MC004535
00001         A1801 01 11.79783 03 36 43.82 +16 55                   MC004535
00001         A1801 01 13.79236 03 36 44.91 +17 02 54.7              MC004535
00001         A1801 01 14.78965 03 36 46.62 +17 07 10.5              MC004535
00001         A1801 01 18.77899 03 37 11    +17 25                   MC004535
00001         A1801 01 19.77641 03 37 24.74 +17 29 12.2              MC004535
00001         A1801 01 21.77126 03 37 51.61 +17 38 25.9              MC004535
00001         A1801 01 22.76871 03 38 07.15 +17 43 05.0              MC004535
00001         A1801 01 23.76618 03 38 25.02 +17 47 47.9              MC004535
00001         A1801 01 28.75376 03 40 14.78 +18 12 11.1              MC004535
00001         A1801 01 30.74893 03 41 09.30 +18 22 15.3              MC004535
00001         A1801 01 31.74652 03 41 38.88 +18 27 18.9              MC004535
00001         A1801 02 01.74414 03 42 09.3  +18 32 26.8              MC004535
00001         A1801 02 02.74178 03 42 41.69 +18 37 40.2              MC004535
00001         A1801 02 05.73479 03 44 26.88 +18 53 19.2              MC004535
00001         A1801 02 08.72793 03 46 24.27 +19 09 13.7              MC004535
00001         A1801 02 11.72121 03 48 33.97 +19 25 18.3              MC004535
00001         A1802 01 26.17022 12 43 22.43 +10 51 17.1              APO01500
00001         A1802 01 27.16767 12 43 38.14 +10 55 33.5              APO01500
00001         A1802 02 04.14664 12 44 47.02 +11 34 23.0              APO01500
00001         A1802 02 11.12723 12 44 21.07 +12 15 23.6              APO01500
00001         A1802 02 27.07946 12 38 26.46 +14 03 35.0              APO01500
00001         A1802 02 28.07632 12 37 51.71 +14 10 36.2              APO01500
00001         A1802 03 06.05722 12 33 57.12 +14 51 40.0              APO01500
00001         A1802 03 07.05399 12 33 13.85 +14 58 15.1              APO01500
00001         A1802 03 10.04425 12 30 59.16 +15 17 31.1              APO01500
00001         A1802 03 11.04097 12 30 12.39 +15 23 42.7              APO01500
00001         A1802 03 15.02781 12 26 59.26 +15 47 13.1              APO01500
00001         A1802 03 16.02450 12 26 09.77 +15 52 48.2              APO01500
00001         A1802 03 18.01788 12 24 29.19 +16 03 14.6              APO01500
00001         A1802 03 19.01456 12 23 38.18 +16 08 18.5              APO01500
00001         A1802 03 20.01124 12 22 47.29 +16 13 09.0              APO01500
```

https://github.com/aerospaceresearch/orbitdeterminator/blob/master/orbitdeterminator/example_data/mpc_ceres_data.txt

Gauss and Ceres

- Giuseppe Piazzi, observing from Sicily, spotted a faint object moving against the stars
- He tracked it for 41 days through only 9 degrees of arc, then it disappeared into the Sun's glare
- Problem: How do you find it again with less than 1% of its orbit observed?

m	k	T–B rule distance (AU)	Planet	Semimajor axis (AU)	Deviation from prediction[1]
−∞	0	0.4	Mercury	0.39	−3.23%
0	1	0.7	Venus	0.72	+3.33%
1	2	1.0	Earth	1.00	0.00%
2	4	1.6	Mars	1.52	−4.77%
3	8	2.8	Ceres[2]	2.77	−1.16%
4	16	5.2	Jupiter	5.20	+0.05%
5	32	10.0	Saturn	9.58	−4.42%
6	64	19.6	Uranus	19.22	−1.95%
–	–	–	Neptune	30.07	–
7	128	38.8	Pluto[2]	39.48	+1.02%

https://en.wikipedia.org/wiki/Titius%E2%80%93Bode_law

How do we pick the best parameters θ ?

Hypothesis

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

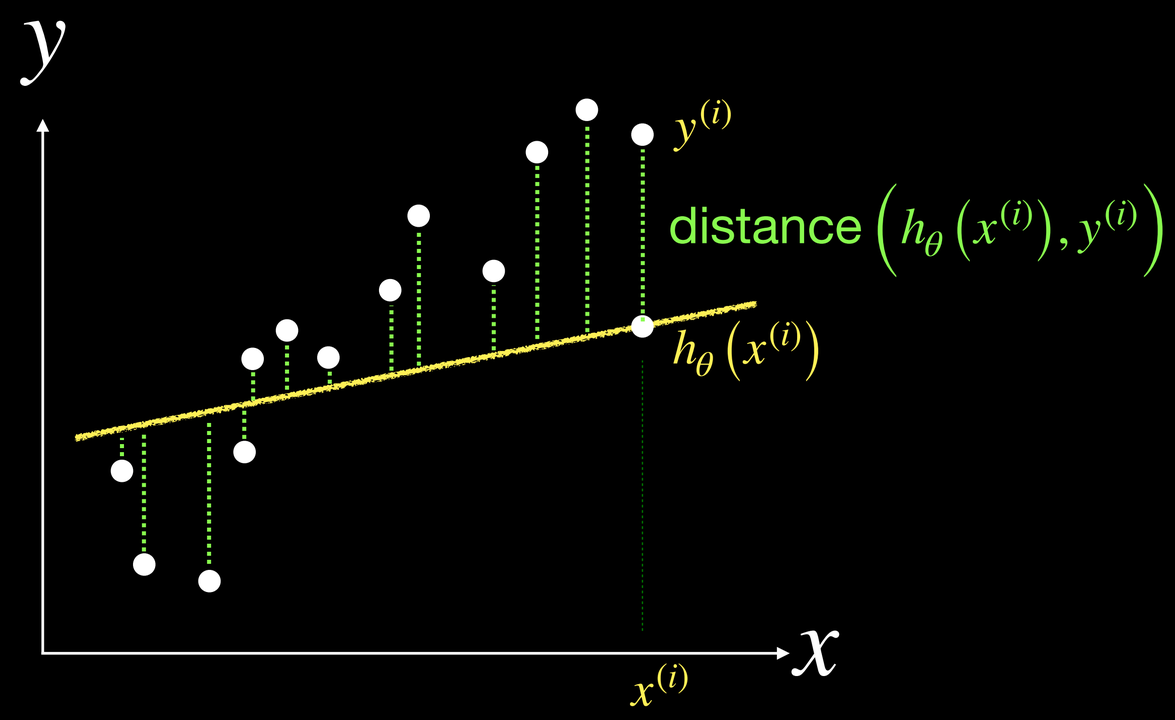
$h_\theta\left(x^{(i)}\right) - y^{(i)}$	Residuals
$\left\lvert h_\theta\left(x^{(i)}\right) - y^{(i)} \right\rvert$	Absolute loss
$\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$	Square loss

Cost function

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

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

Ordinary least squares



Choose θ to minimize the cost $J(\theta)$

Cost function

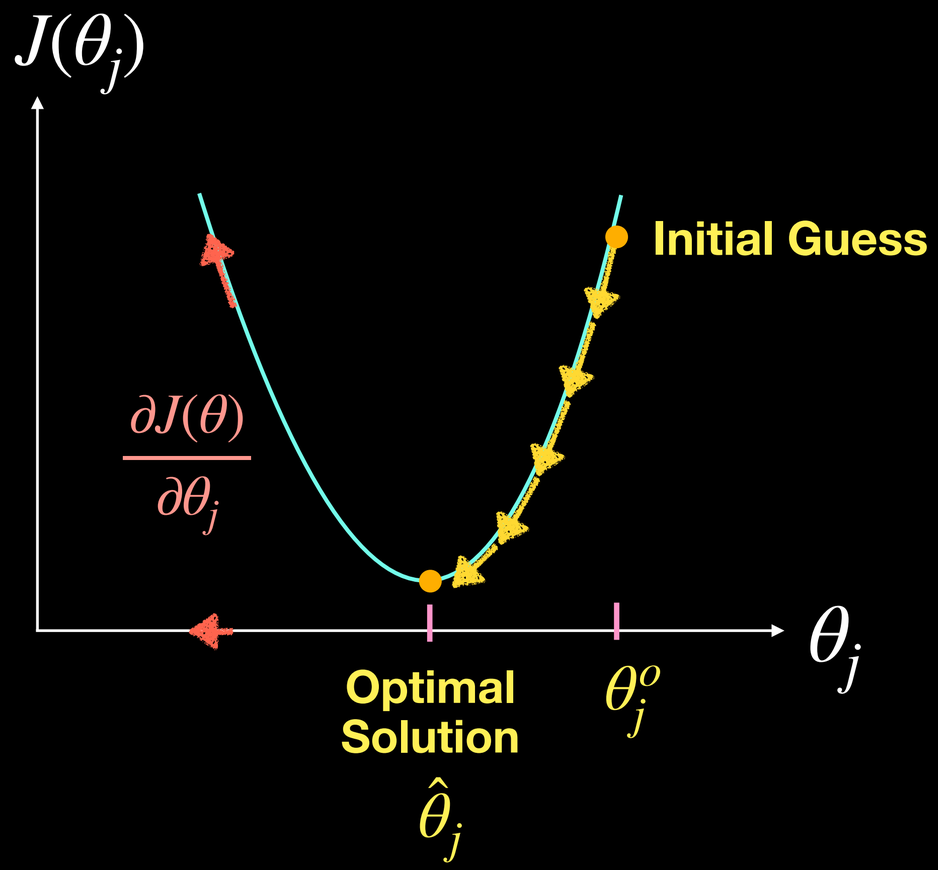
$$J(\theta) = \frac{1}{2} \sum_{i=1}^{d} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right)^2$$

Gradient Descent Update

while not converged:

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

Learning Rate



The gradient can be computed explicitly

Gradient Descent Update

while not converged:

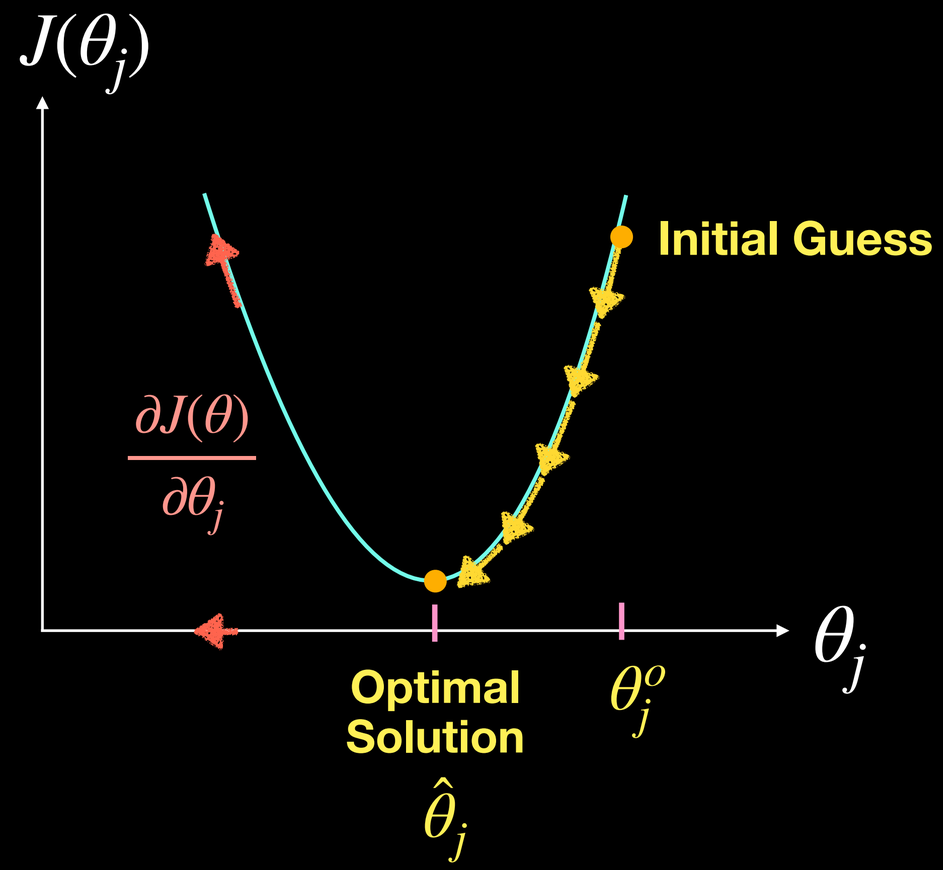
$$\theta_j := \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j}$$

Learning Rate

Derive $\frac{\partial J(\theta)}{\partial \theta_j}$ explicitly

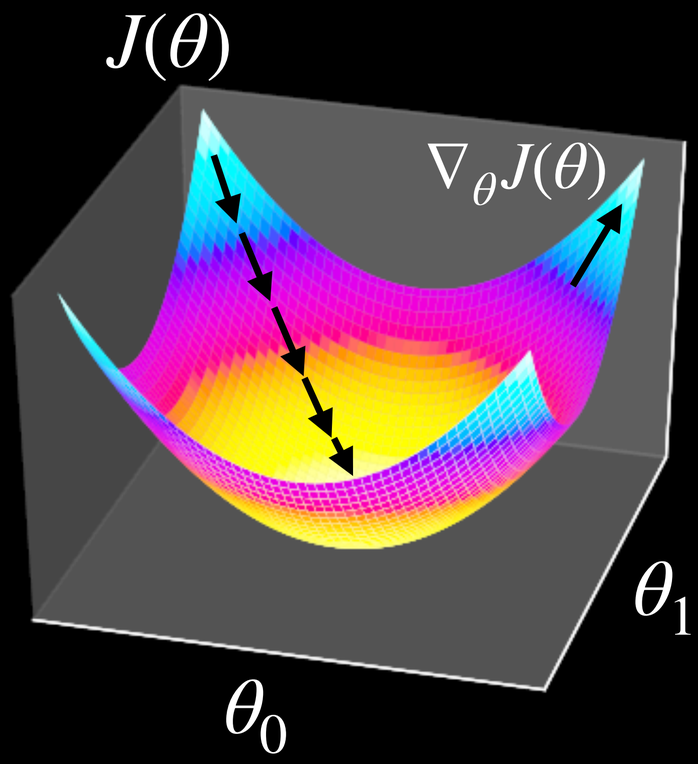
For a single (x, y) pair

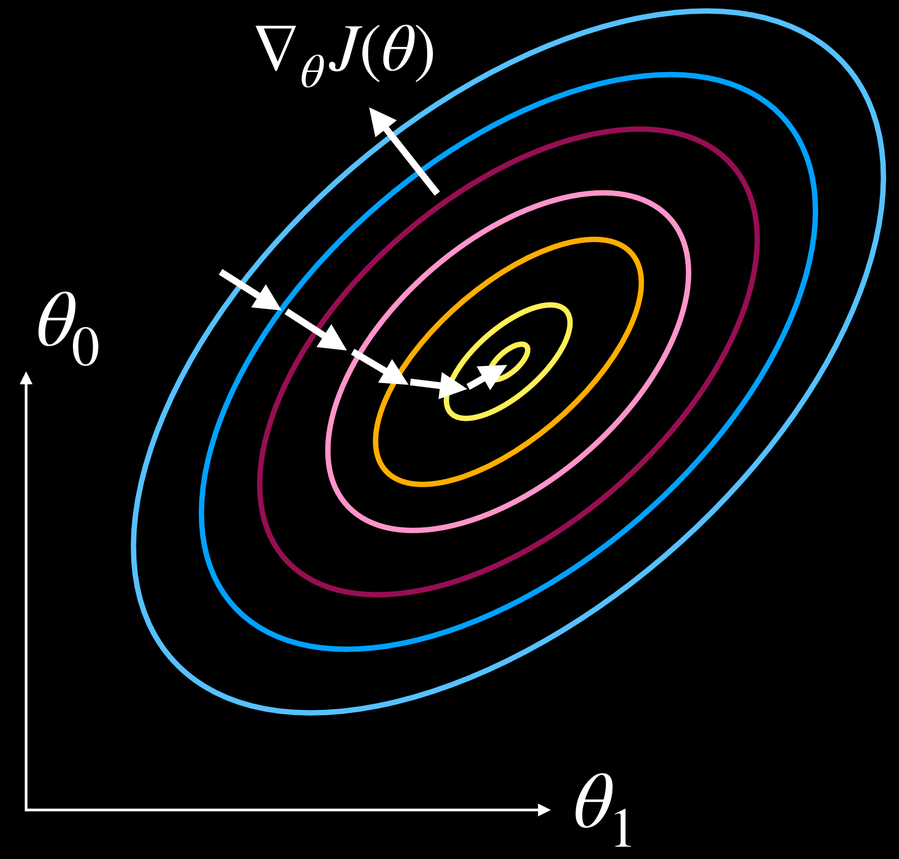
Assume: $y = \theta_0 + \theta_1 x$



Gradient Descent Update in 2D

$$\vec{\theta} := \vec{\theta} - \alpha \nabla_\theta J(\theta)$$





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

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

Stack and vectorize

$$\vec{\theta} := \vec{\theta} - \alpha \vec{\nabla}_\theta J(\theta)$$

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

Batch Gradient Descent

for t = 1...T: (Epochs)

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

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

too expensive for large n

For a single training example $\left(x^{(i)}, y^{(i)} \right)$:

$$\vec{\theta} := \vec{\theta} - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \vec{x}^{(i)}$$

Stochastic Gradient Descent

for t = 1...T: (Epochs)

for all examples i:

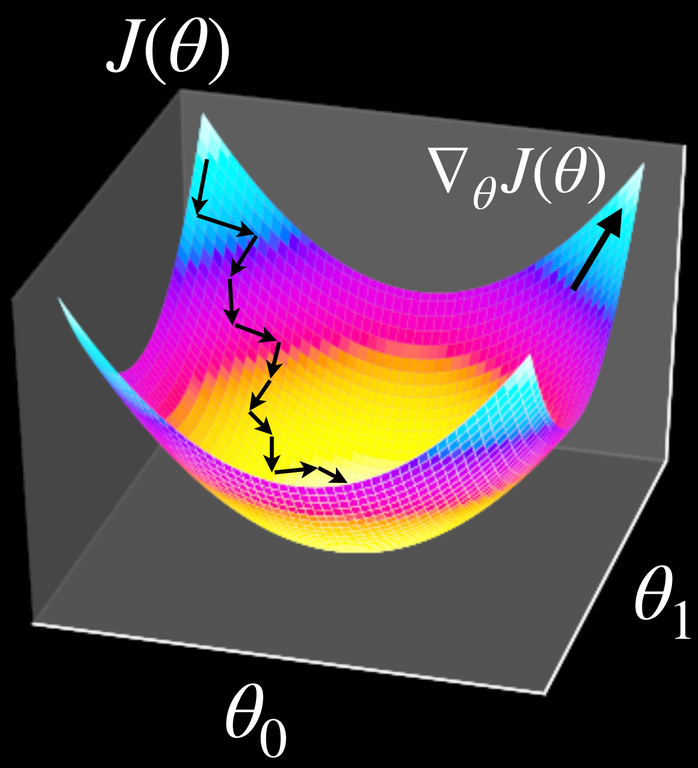
$$\vec{\theta} := \vec{\theta} - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \vec{x}^{(i)}$$

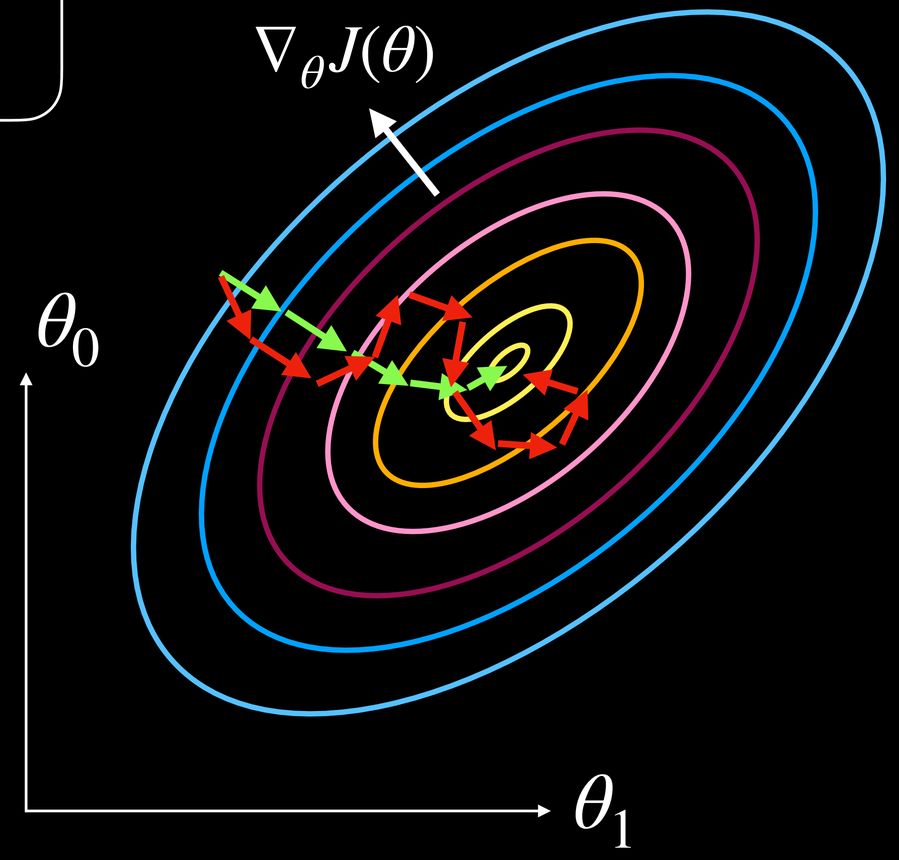
Stochastic Gradient Descent

for t = 1…T: (Epochs)

for all examples i:

$$\vec{\theta} := \vec{\theta} - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \vec{x}^{(i)}$$





Batch Gradient Descent

for t = 1...T:

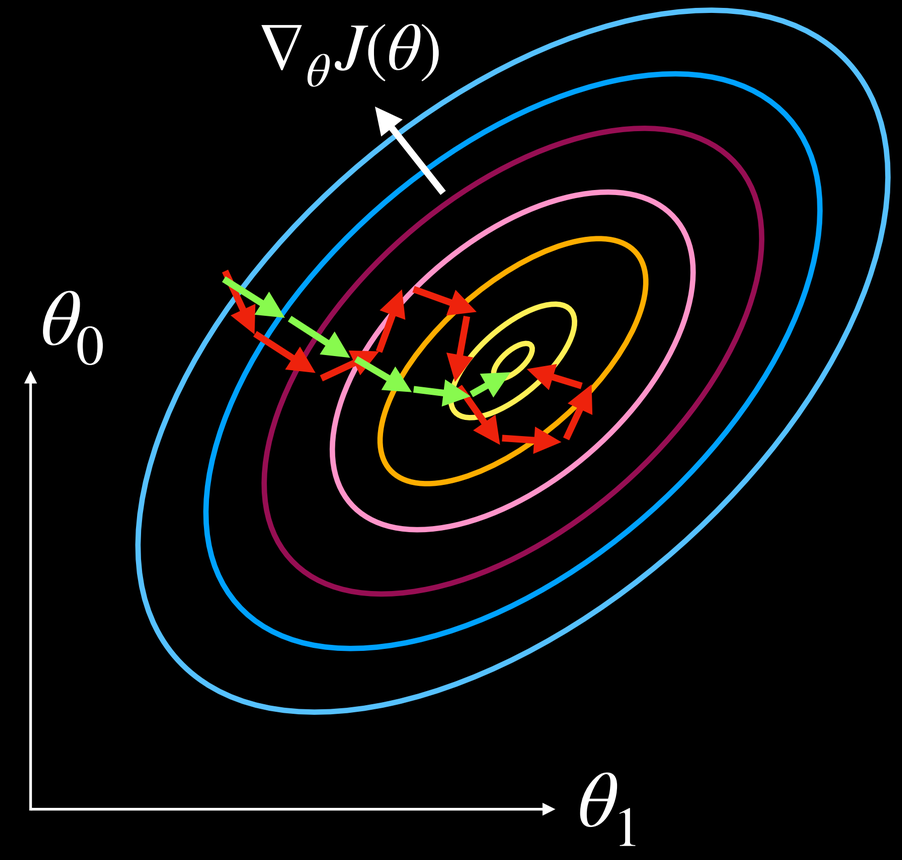
$$\theta := \theta - \alpha \sum_{i=1}^{n} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x^{(i)}$$

Stochastic Gradient Descent

for t = 1...T:

  for all examples i:

$$\vec{\theta} := \vec{\theta} - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \vec{x}^{(i)}$$



1. Assume a linear hypothesis

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

2. Cost function

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

3. Minimize

$$\hat{\theta} = \arg\min_\theta J(\theta)$$

SGD

for t = 1...T:

  for all examples i:

$$\vec{\theta} := \vec{\theta} - \alpha \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) \vec{x}^{(i)}$$

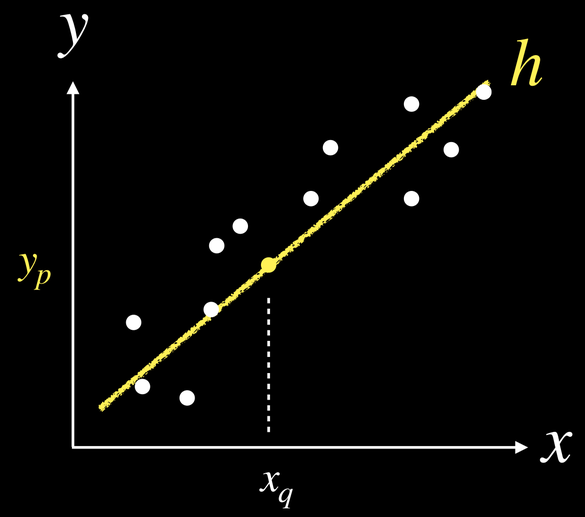
4. Optimal predictor

$$y = h_{\hat{\theta}}(x)$$

$\hat{\theta}$

5. Predict unseen data

$$y_{pred} = h_{\hat{\theta}}(x_{new})$$



Hooke’s Law

Homework assignment for next week - with leaderboard