Linear Regression

Szymon Bobek

Institute of Applied Computer science
AGH University of Science and Technology

Based on Carlos Guestrin adn Emily Fox slides from
Coursera Specialization on Machine Learnign
http://geist.agh.edu.pl

Outline I

Presentation Outline

Roadmap

1. Regression
   - Linear regression
   - Ridge regression
   - Bias, variance tradeoff
2. Classification
   - Logistic regression
   - Support Vector Machines
   - Decision trees

Presentation Outline

Problem

Problem

Problem

Presentation Outline

Model

Data

$(x_1 = 150\ m^2, y_1 = 100\ 000\$)$

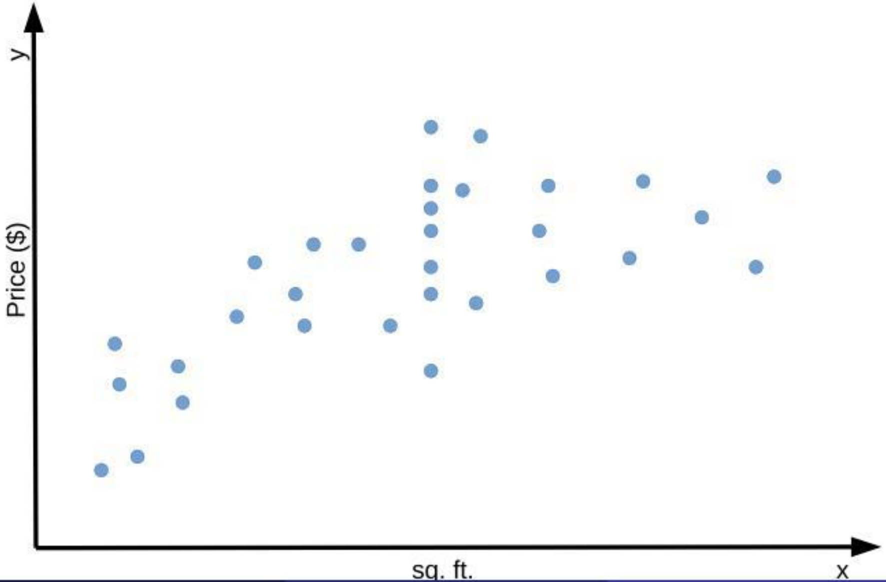
$(x_2 = 150\ m^2, y_2 = 100\ 000\$)$

$(x_3 = 150\ m^2, y_3 = 100\ 000\$)$

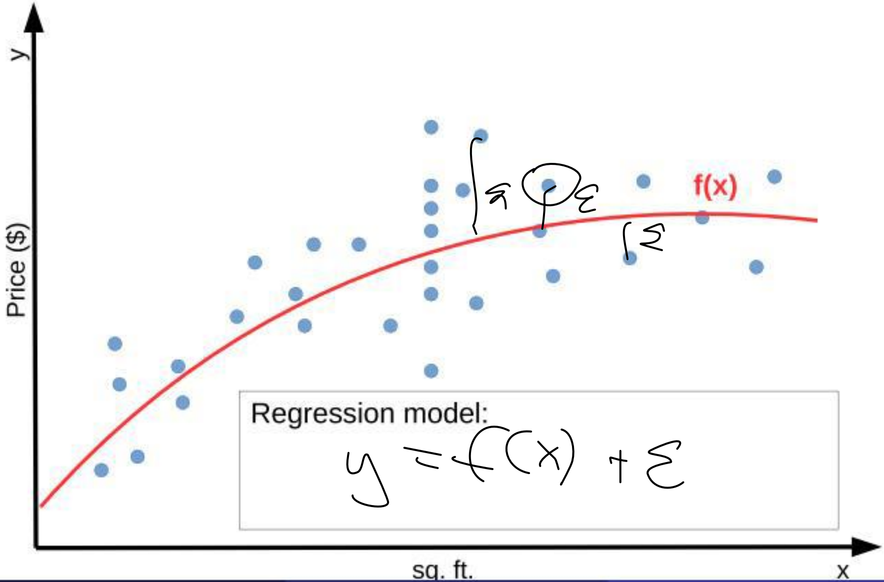
...

$(x_m = 150\ m^2, y_m = 100\ 000\$)$

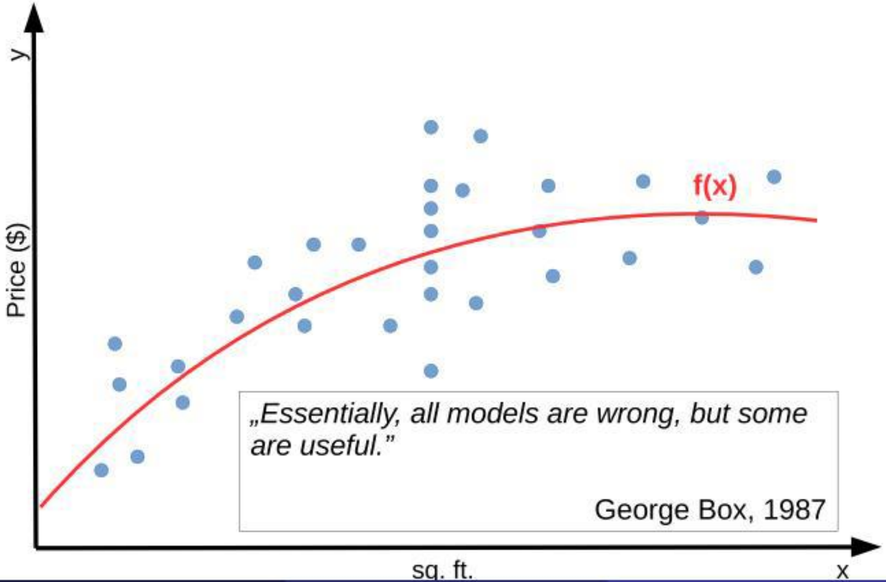
Model



Model

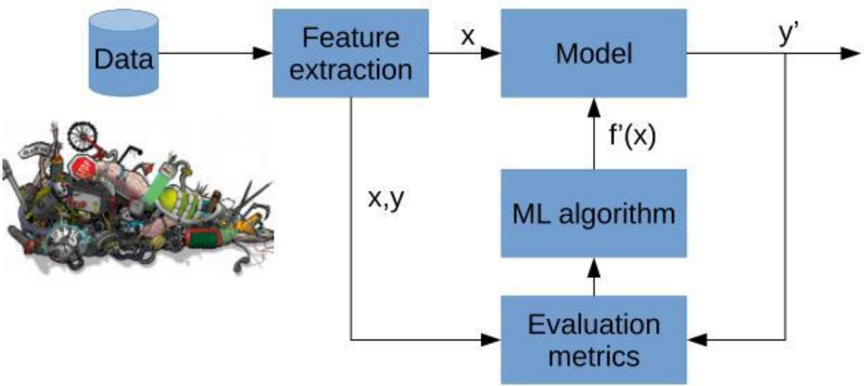


Model

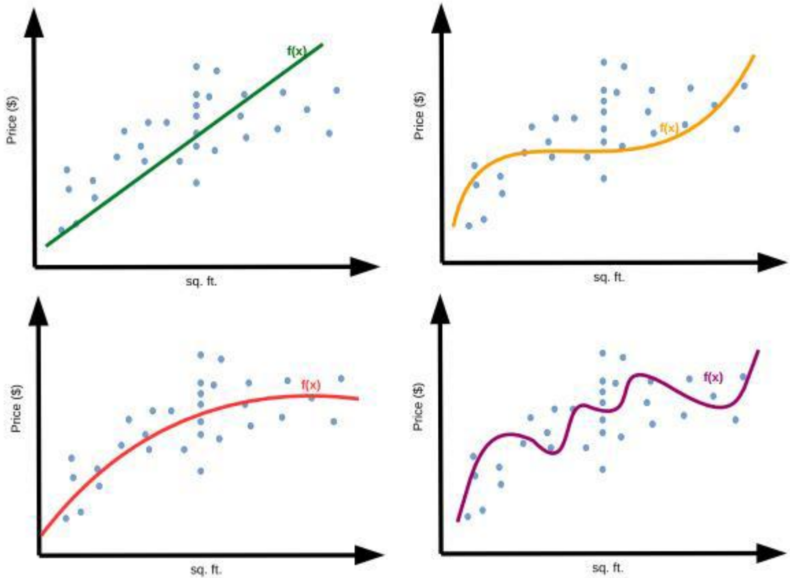


y
Price ($)
f(x)
„Essentially, all models are wrong, but some are useful."
George Box, 1987
sq. ft.
x

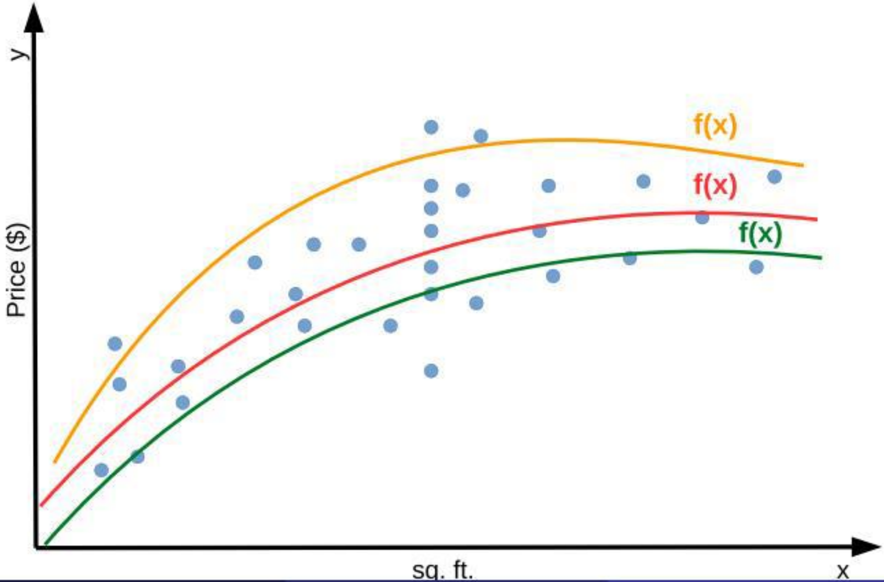
How to get the model



Which model?

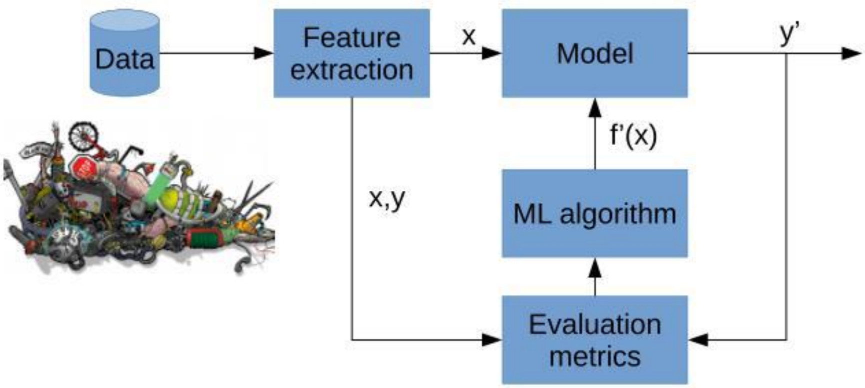


Which model?

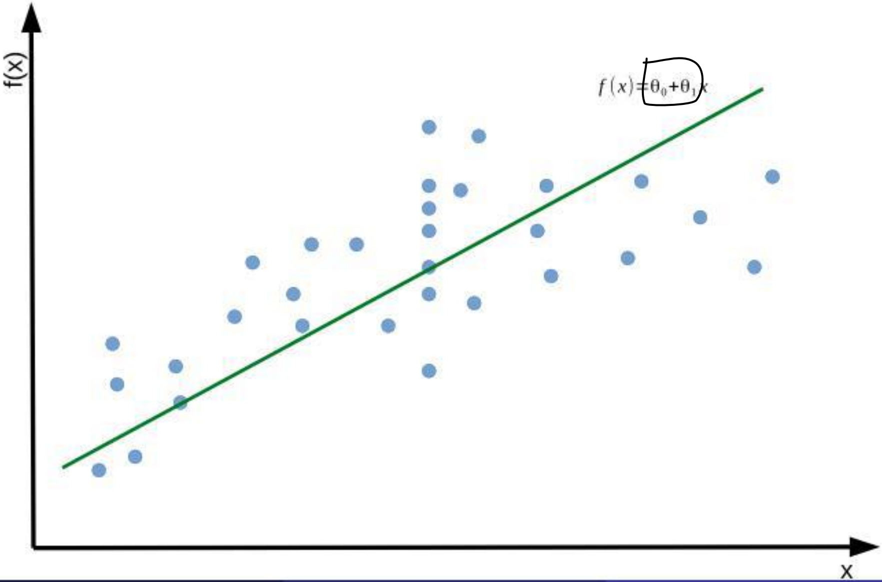


Presentation Outline

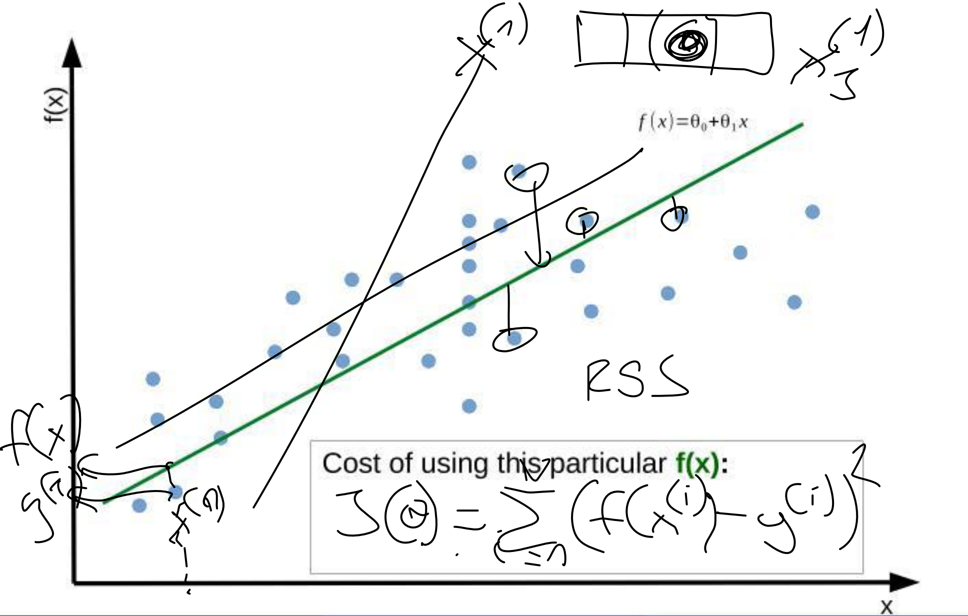
Representing a model



Representing a model



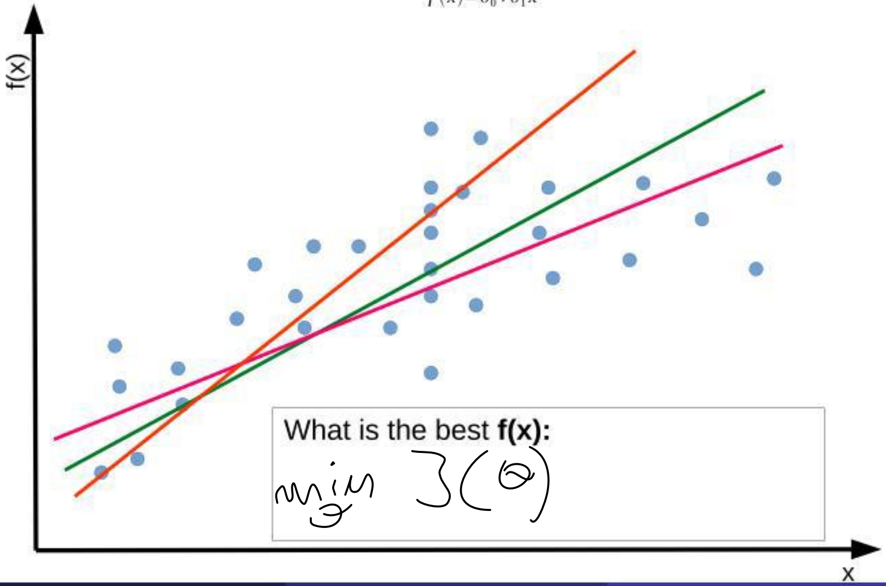
Cost of using a model



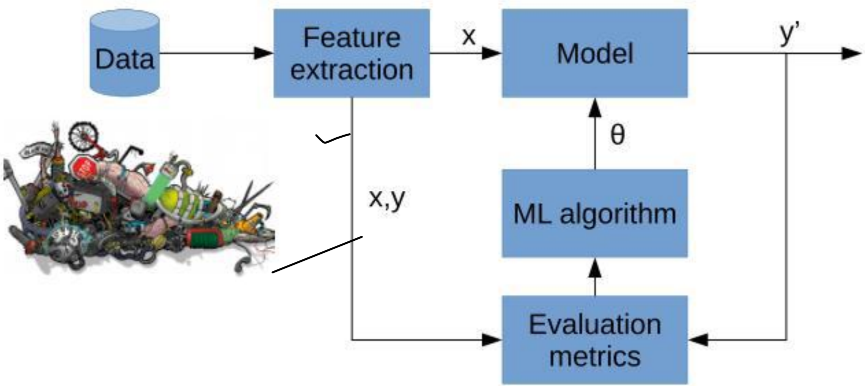
$$J(\theta) = \sum_{i=1}^{n} \left(f(x^{(i)}) - y^{(i)}\right)^2$$

Cost of using a model

$$f(x) = \theta_0 + \theta_1 x$$



Cost of using a model



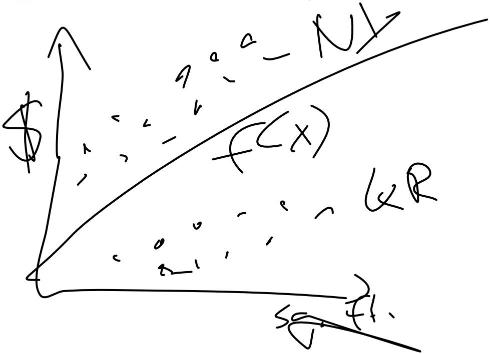
Outline

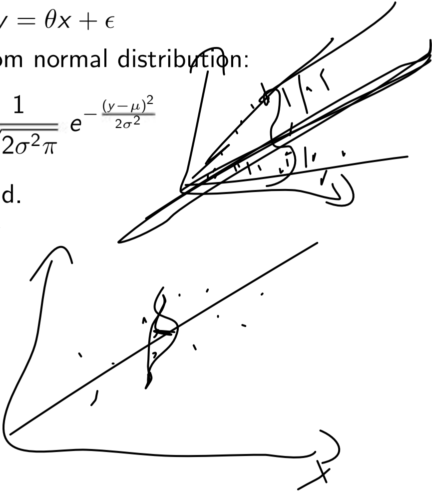
Gaussian interpretation

- Prediction is linear function and noise: $y = \theta x + \epsilon$
- We assume that the noise ϵ is drawn from normal distribution:

$$\mathcal{N}(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

- We assume that training samples are i.i.d.





Gaussian interpretation

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

- We want to learn

$$\mathcal{P}(y \mid \theta, x, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(y-\sum_i \theta_i x_i)^2}{2\sigma^2}}$$

- The best P is when probability is max for every training set:

$$\operatorname{argmax}_{\theta} \ln \mathcal{P}(D \mid \theta, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} \prod_j^N e^{-\frac{(y^{(j)} - \sum_i \theta_i x_i^{(j)})^2}{2\sigma^2}} =$$

$$= \ln \ldots$$

$$\sum^N \ln e^{-\frac{(y^{(i)} \ldots)^2}{2\sigma^2}}$$

$$+\sum \frac{(y^{(i)} - f(x))^2}{\cancel{2\sigma^2}} = RSS$$

Other cost functions

- Residual sum of squares:

$$RSS(w) = \sum_{i}^{N}(y^{(i)} - \theta x^{(i)})^2$$

- Mean squared error:

$$\frac{1}{N}\sum_{i}^{N}(y^{(i)} - \theta x^{(i)})^2$$

- Log error:

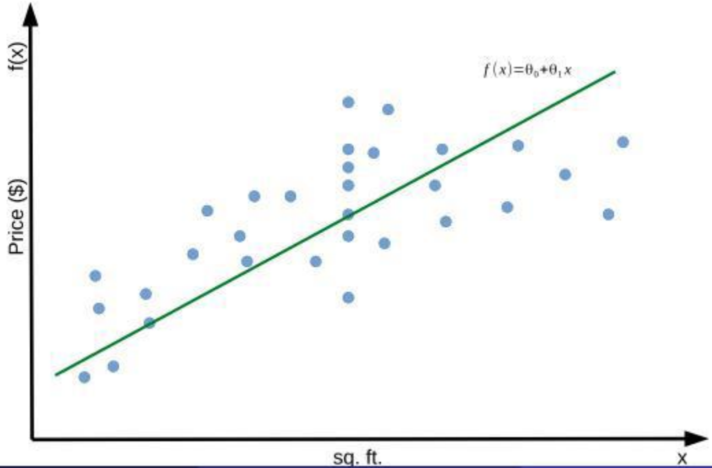
$$\frac{1}{N}\sum_{i=1}^{N}(-y^{(i)} log(\theta x^{(i)}) - (1 - y^{(i)}) log(1 - \theta x^{(i)})))$$

- Asymetric:

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha - \mathbb{1}_{(y^{(i)} - f(x^{(i)}))<0}\right| \cdot \left(y^{(i)} - f(x^{(i)})\right)^2$$

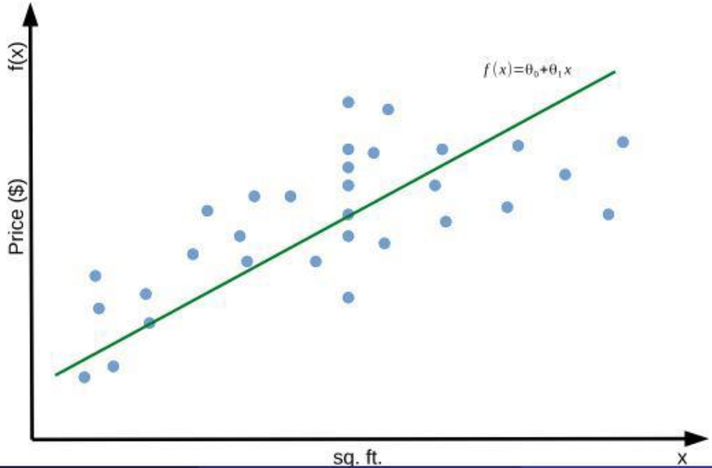
Assymetric cost function

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha - \mathbb{1}_{(y^{(i)}-f(x^{(i)}))<0}\right| \cdot \left(y^{(i)} - f(x^{(i)})\right)^2$$



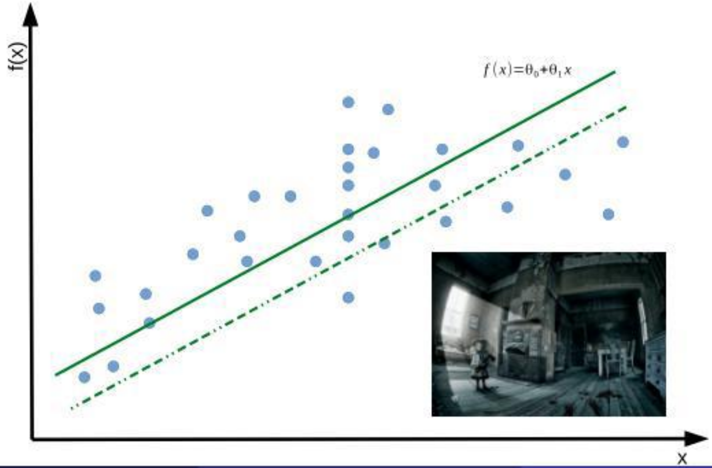
Assymetric cost function

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha - \mathbb{1}_{(y^{(i)}-f(x^{(i)}))<0}\right| \cdot \left(y^{(i)} - f(x^{(i)})\right)^2$$



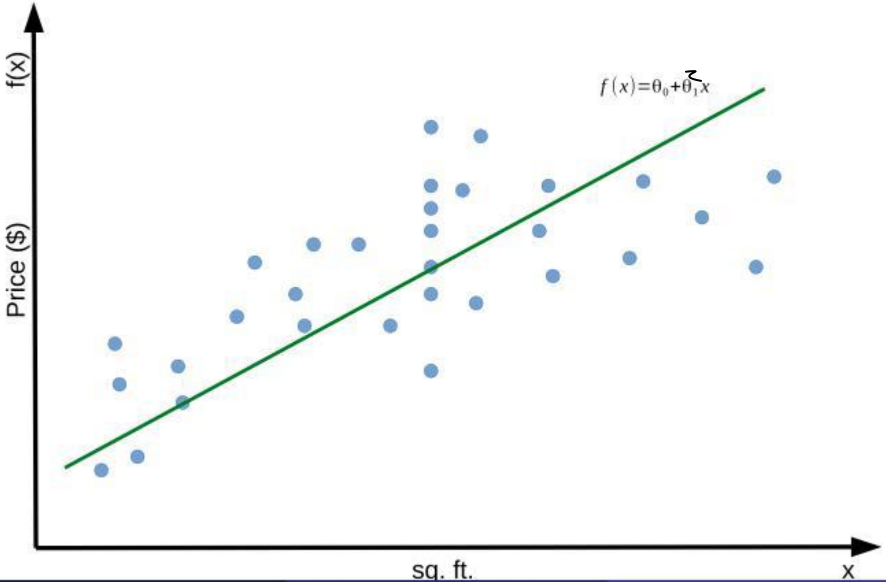
Assymetric cost function

$$\frac{1}{n}\sum_{i=1}^{n}\left|\alpha - \mathbb{1}_{(y^{(i)}-f(x^{(i)}))<0}\right| \cdot \left(y^{(i)} - f(x^{(i)})\right)^2$$

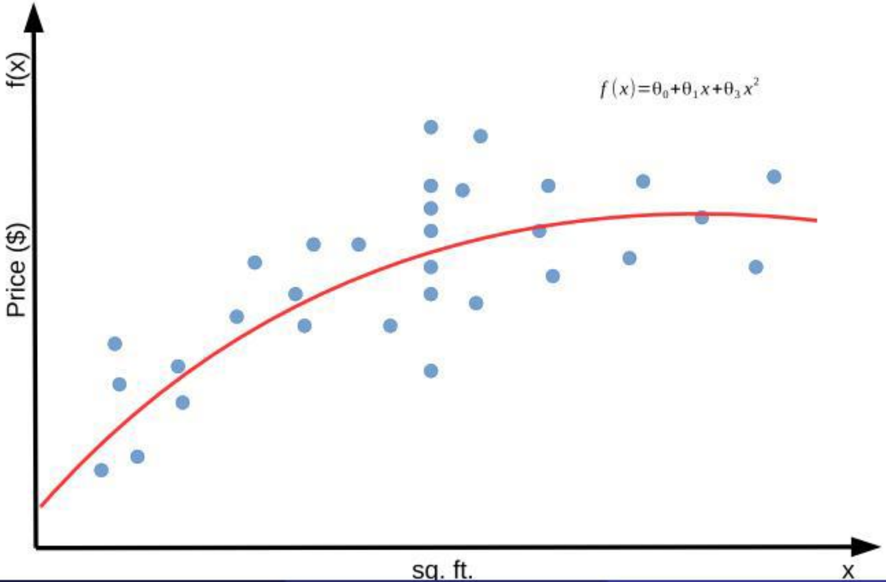


Outline

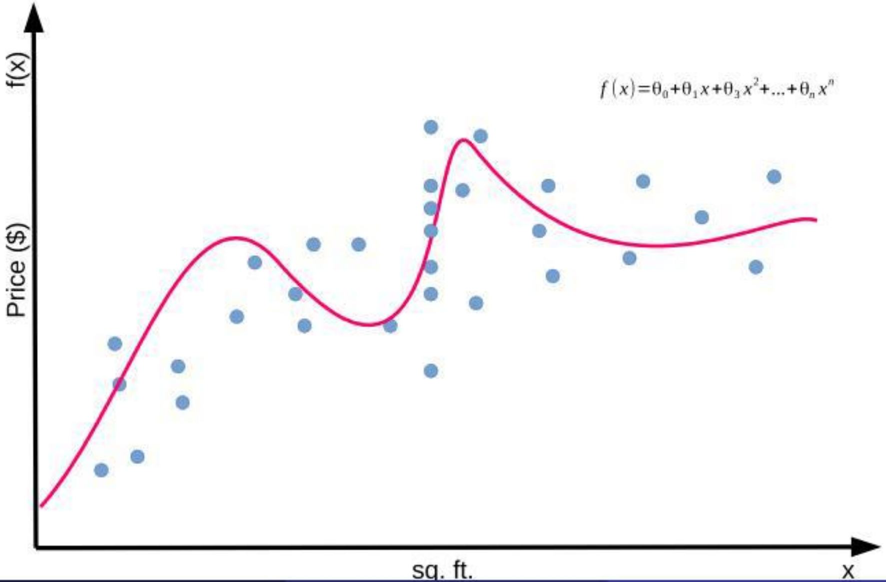
"Linear" is only regression



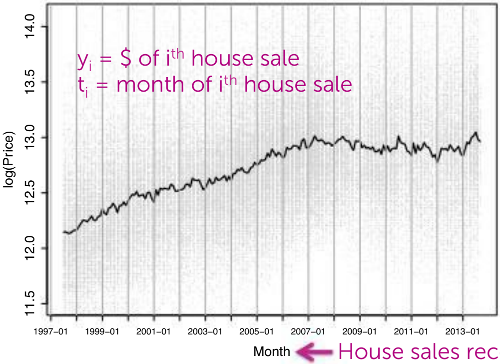
"Linear" is only regression



"Linear" is only regression



Motivating application: Detrending time series

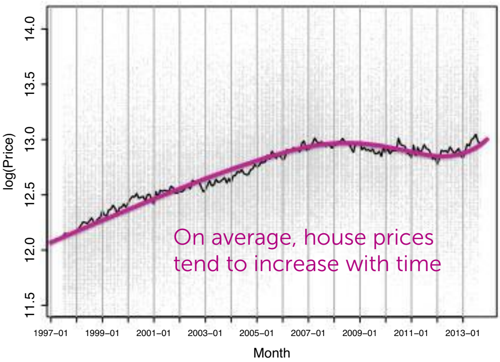


y_i = \$ of i^{th} house sale

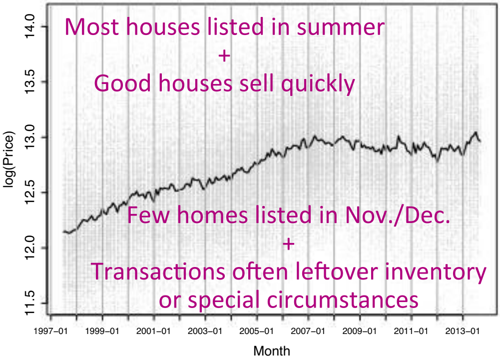
t_i = month of i^{th} house sale

House sales recorded monthly

Trends over time



Seasonality



An example detrending

Model: $x_0 = 1$

$y_i = \theta_0 + \theta_1 t_i + \theta_2 \sin(2\pi t_i / 12 - \Phi) + \varepsilon_i$

Unknown phase/shift

Linear trend

Seasonal component = Sinusoid with period 12 (resets annually)

Trigonometric identity: $\sin(a-b)=\sin(a)\cos(b)-\cos(a)\sin(b)$

→ $\sin(2\pi t_i / 12 - \Phi) = \sin(2\pi t_i / 12)\underbrace{\cos(\Phi)}_{\theta_2} - \cos(2\pi t_i / 12)\underbrace{\sin(\Phi)}_{\theta_3}$

An example detrending

Equivalently,

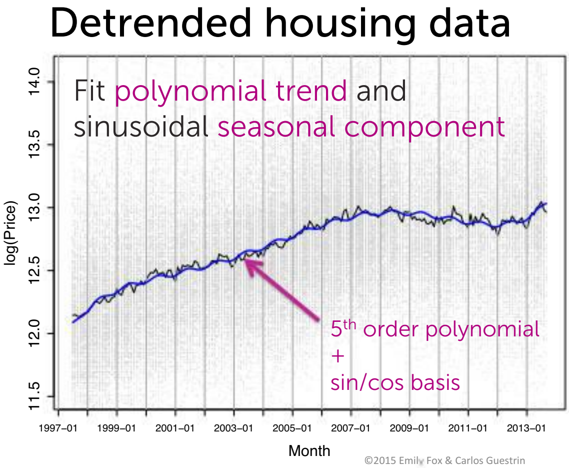
$$y_i = \theta_0 + \theta_1 t_i + \theta_2 \sin(2\pi t_i / 12) + \theta_3 \cos(2\pi t_i / 12) + \epsilon_i$$

feature 1 = 1 (constant)

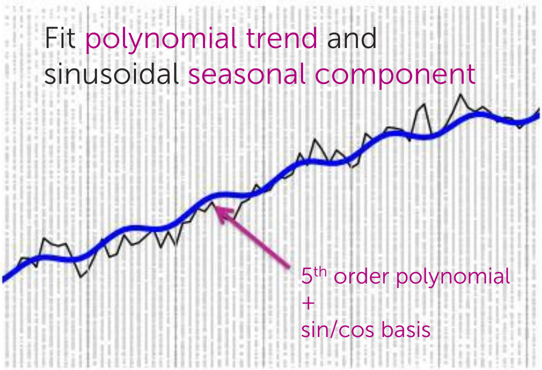
feature 2 = t

feature 3 = $\sin(2\pi t/12)$

feature 4 = $\cos(2\pi t/12)$


Detrended housing data
Fit polynomial trend and sinusoidal seasonal component
5th order polynomial + sin/cos basis
log(Price)
14.0
13.5
13.0
12.5
12.0
11.5
1997–01
1999–01
2001–01
2003–01
2005–01
2007–01
2009–01
2011–01
2013–01
Month
©2015 Emily Fox & Carlos Guestrin

Zoom in...

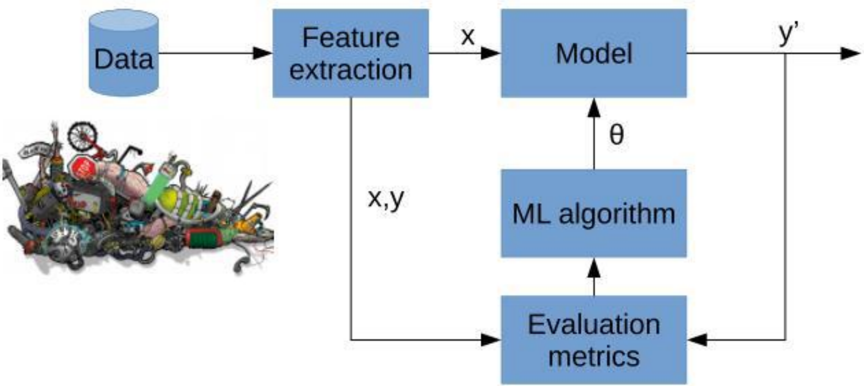

Fit polynomial trend and
sinusoidal seasonal component
5th order polynomial
+
sin/cos basis

Even more complex features

Adding more dimensions to the problem can be done mainly in two ways:

- By addning aditional features (like number of bathrooms)
- By adding artificial features being just a combinations or functions of existing ones (squared, log, etc.)
- To allow more compact notation, usually, we add fake *zero-th* feature that equals one, and is multiplied by θ_0.

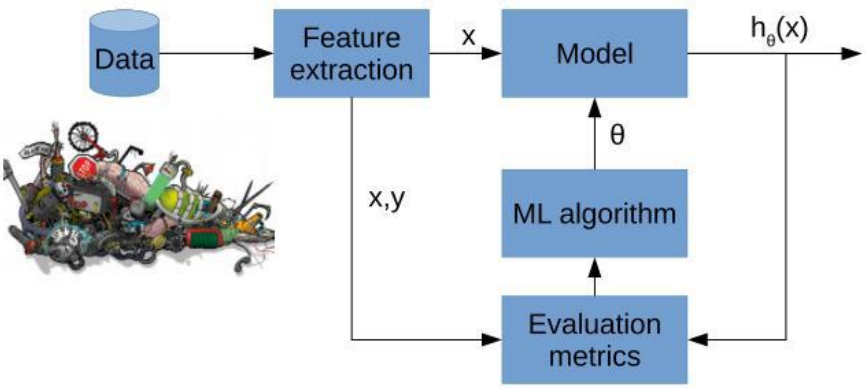
Note on notation



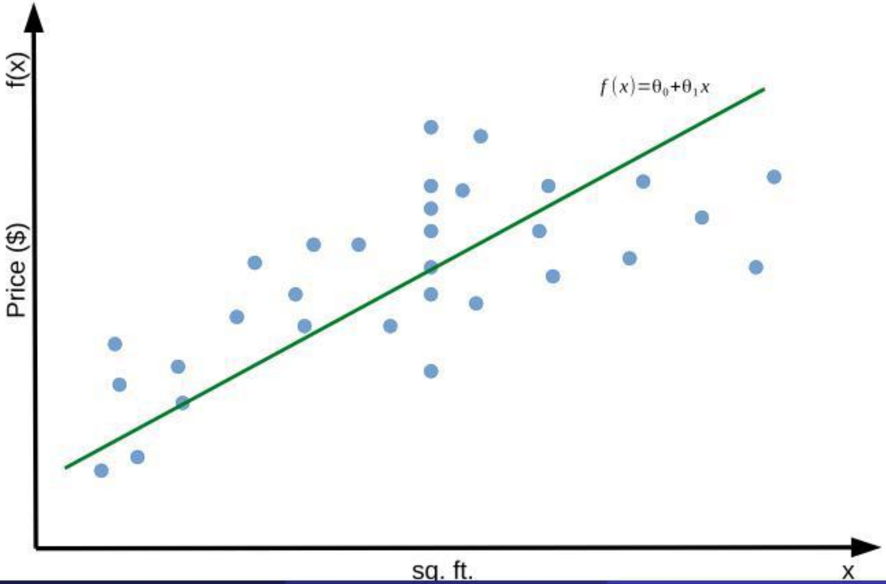
Note on notation

$\theta > \theta x^2$

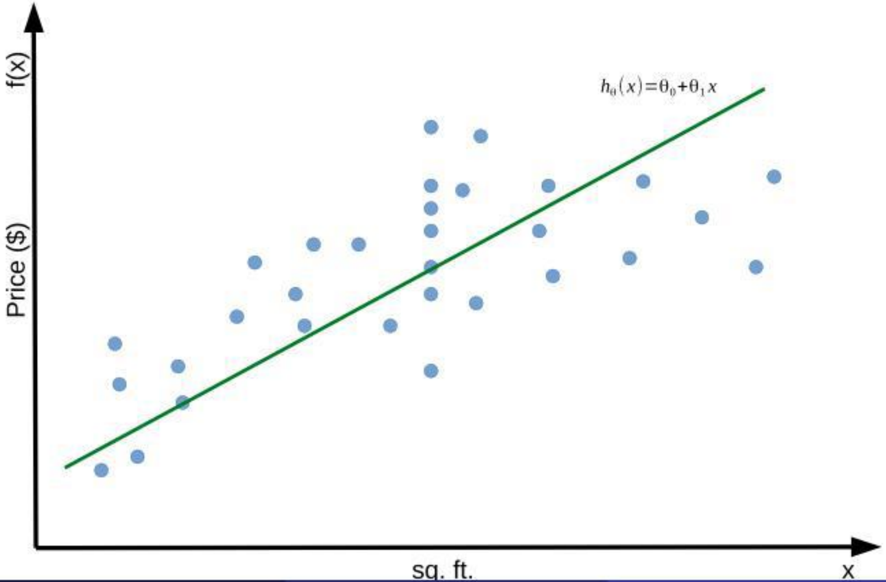
sin



Note on notation



Note on notation



Note on notation

- N – number of observations
- D – number of dimensions
- $h_\theta(x)$ – a hypothesis to be learned
- X – training set, where rows represents training instances
- $x^{(j)}$ – *j-th* training instance. For example one instance constains many feature values:
  - $x_1^{(j)}$ – sq. ft.
  - $x_2^{(j)}$ – no. of bathrooms
  - $x(j)_i$ – some uber cool feature
- θ – vector of coefficients to learn (each θ_i corresponds to each x_i)
- y – ground truth values for training set. $y^{(j)}$ value corresponds to *j-th* row from X, thus to $x^{(j)}$.
- We will denote $J(\theta)$ as a cost function which in our case will be RSS. Therefore

$$J(\theta) = \sum_{i=1}^{N} (h_{theta}(x^{(i)}) - y^{(i)})^2$$

$f(x)$ X

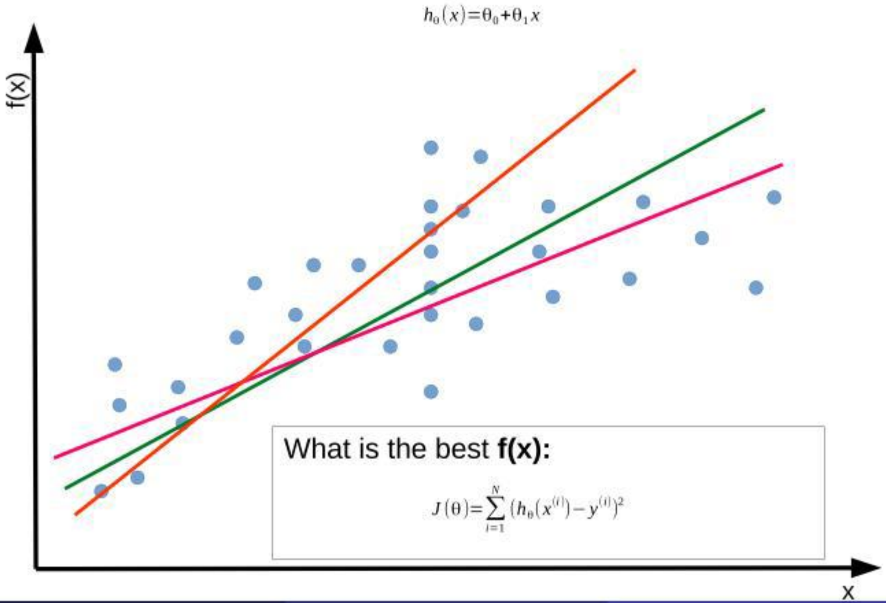
sq #bath

y

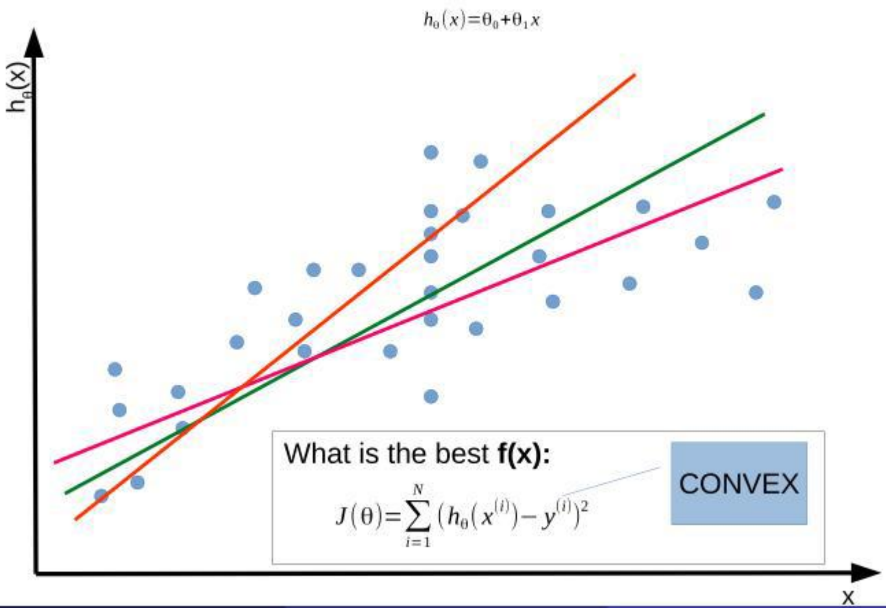
$\theta = [\theta_0, \theta_1, \theta_2 \dots]$

Presentation Outline

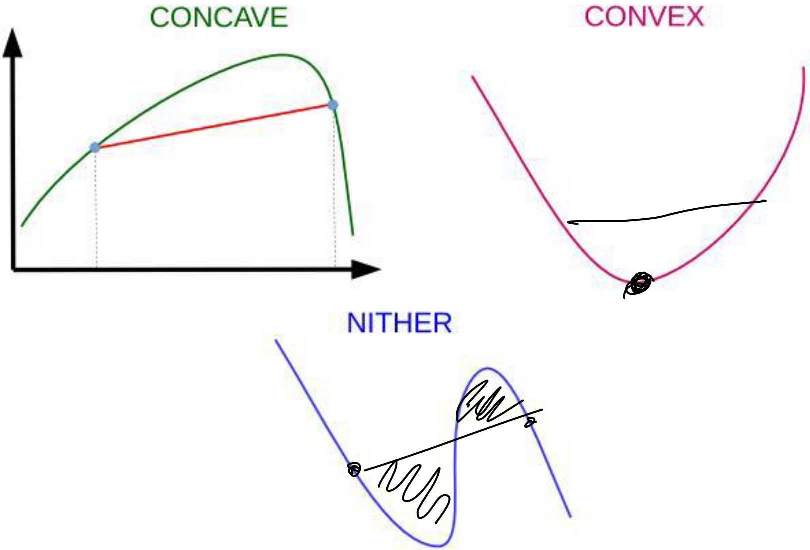
Minimizing the error



Minimizing the error



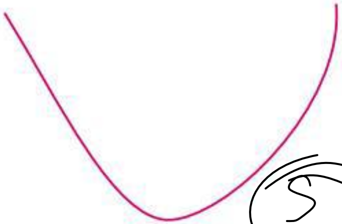
Minimizing the error



Outline

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

CONVEX



$g(\theta)=5-(\theta-10)^2$

5

$$0=\frac{\partial}{\partial\theta}=0-2(\theta-10)$$

$$0=-2(\theta-10)$$

$$\boxed{\theta=10}$$

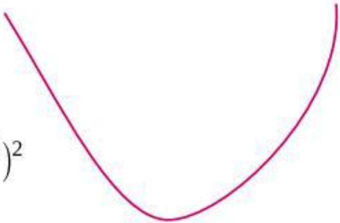
Calculate gradient

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

$$J(\theta)=\sum_{i=1}^{N}\left(\theta_0 x_0^{(i)}+\theta_1 x_1^{(i)}-y^{(i)}\right)^2$$

1

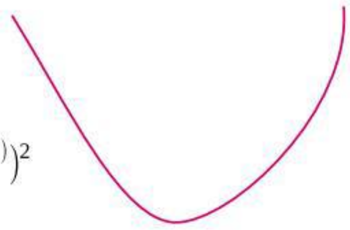
CONVEX



$g(\theta)=5-(\theta-10)^2$

Calculate gradient

CONVEX



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

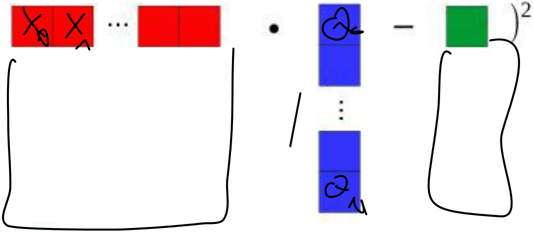
$$J(\theta)=\sum_{i=1}^{N}\left(\theta_0 \boxed{x_0^{(i)}}+\theta_1 \cancel{x_1^{(i)}}-\cancel{y^{(i)}}\right)^2$$

$$\nabla J(\theta)=\begin{bmatrix} 2\sum_{i=1}^{N}\left(\theta_0 x_0^{(i)}+\theta_1 x_1^{(i)}-y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\theta_0 x_0^{(i)}+\theta_1 x_1^{(i)}-y^{(i)}\right)x_1^{(i)} \end{bmatrix} = 0$$

Calculate gradient

$$J(\theta)=\sum_{i=1}^{N}(\theta_0 x_0^{(i)}+\theta_1 x_1^{(i)}-y^{(i)})^2$$

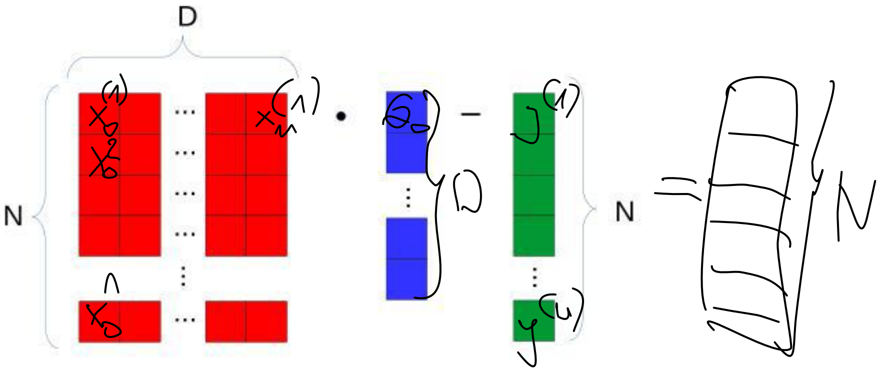
$$J(\theta)=\sum_{i=1}^{N}(x_0^{(i)}\theta_0+x_1^{(i)}\theta_1-y^{(i)})^2$$

$$J(\theta)=\sum_{i=1}^{N}($$  $$)^2$$

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

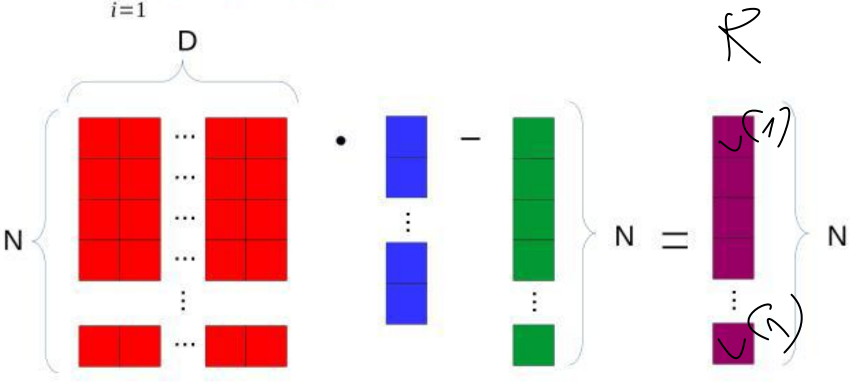
Calculate gradient

$$J(\theta) = \sum_{i=1}^{N} \left((x^{(i)})^T \theta - y^{(i)} \right)^2$$



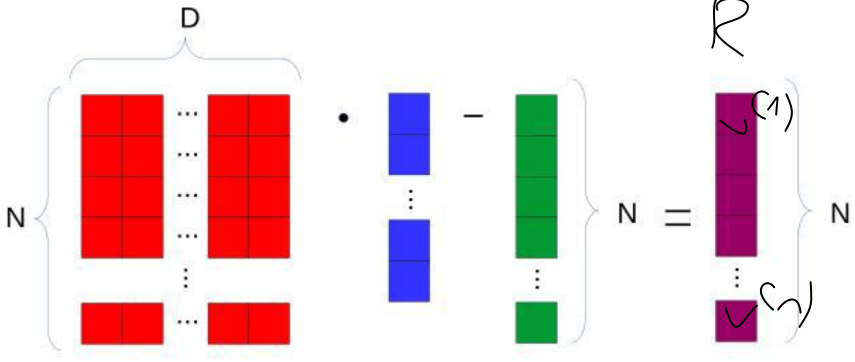
Calculate gradient

$$J(\theta) = \sum_{i=1}^{N} \left((x^{(i)})^T \theta - y^{(i)} \right)^2$$



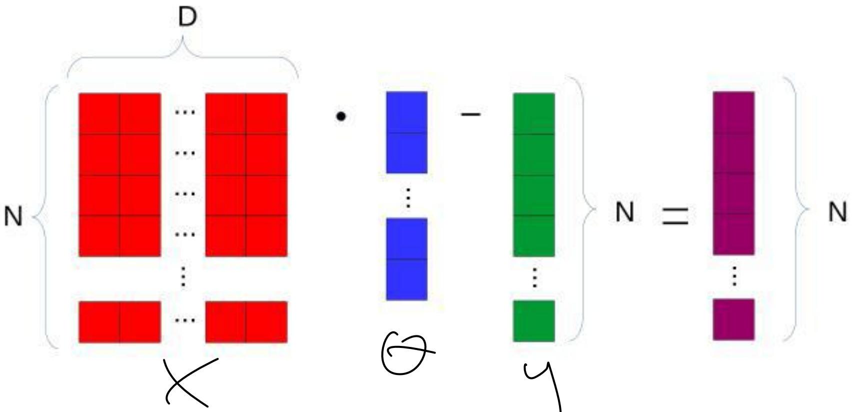
Calculate gradient

$$J(\theta)=\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta-y^{(i)}\right)^2=\sum_{i=1}^{N}\left(r^{(i)}\right)^2 = R^T R$$



Calculate gradient

$$J(\theta) = \sum_{i=1}^{N} \left((x^{(i)})^T \theta - y^{(i)} \right)^2 = \overbrace{(X\theta - y)}^{R}{}^T (X\theta - y)$$

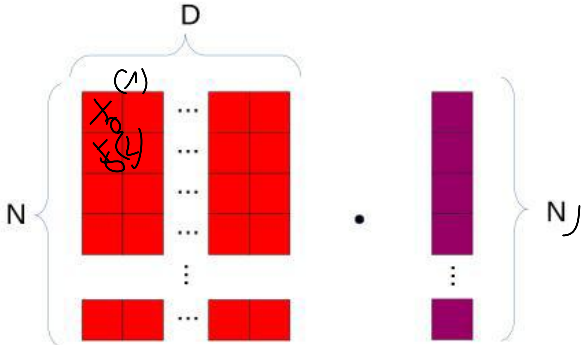


Calculate gradient

$$\nabla J(\theta) = \begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_1^{(i)} \end{bmatrix} = \begin{bmatrix} 2\sum_{i=1}^{N}\left(x_0^{(i)} r^{(i)}\right) \\ 2\sum_{i=1}^{N}\left(x_1^{(i)} r^{(i)}\right) \end{bmatrix}$$

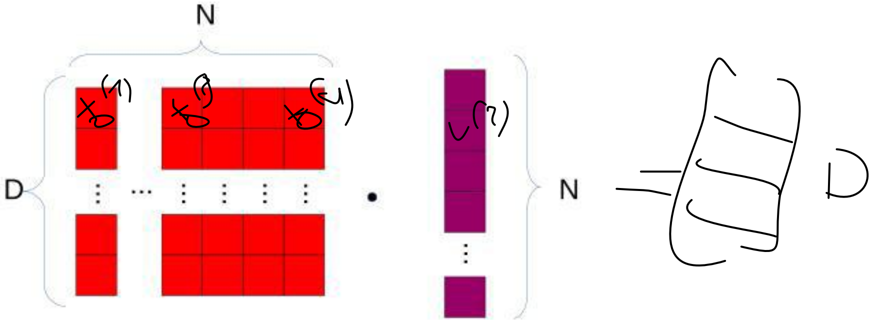
Calculate gradient

$$\nabla J(\theta)=\begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^{T}\theta-y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^{T}\theta-y^{(i)}\right)x_1^{(i)} \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{N}\left(x_0^{(i)}r^{(i)}\right) \\ \sum_{i=1}^{N}\left(x_1^{(i)}r^{(i)}\right) \end{bmatrix}$$



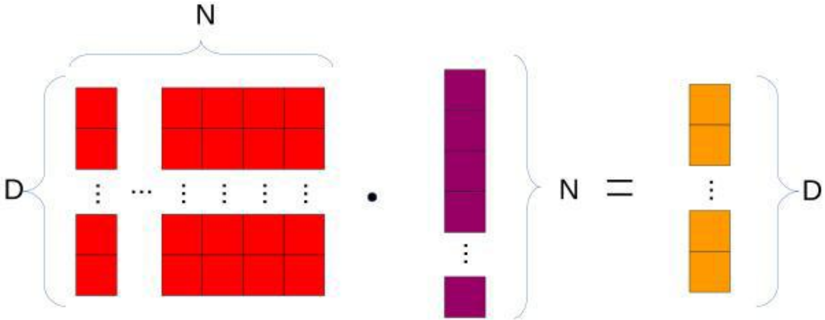
Calculate gradient

$$\nabla J(\theta) = \begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_1^{(i)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N}\left(x_0^{(i)} r^{(i)}\right) \\ \sum_{i=1}^{N}\left(x_1^{(i)} r^{(i)}\right) \end{bmatrix}$$



Calculate gradient

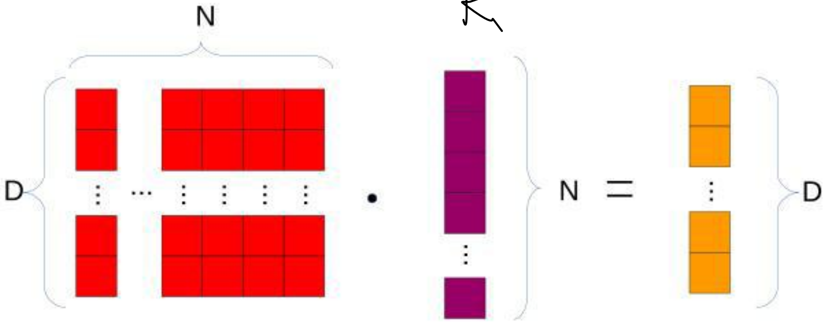
$$\nabla J(\theta) = \begin{bmatrix} 2\sum_{i=1}^{N} \left((x^{(i)})^T \theta - y^{(i)} \right) x_0^{(i)} \\ 2\sum_{i=1}^{N} \left((x^{(i)})^T \theta - y^{(i)} \right) x_1^{(i)} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{N} \left(x_0^{(i)} r^{(i)} \right) \\ \sum_{i=1}^{N} \left(x_1^{(i)} r^{(i)} \right) \end{bmatrix} = X^T R$$



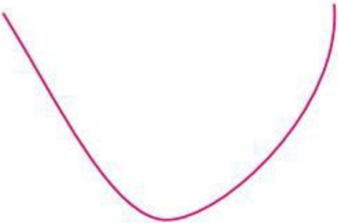
Calculate gradient

$$\nabla J(\theta)=\begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^{T}\theta-y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^{T}\theta-y^{(i)}\right)x_1^{(i)} \end{bmatrix}=2X^{T}\overbrace{(X\theta-y)}^{R}$$

R



CONVEX



$g(\theta)=5-(\theta-10)^2$

$$J(\theta)=(X\theta-y)^T(X\theta-y)$$

$$\nabla J(\theta)=2X^T(X\theta-y)$$

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

$$2X^T(X\theta-y)=0$$

$$2X^TX\theta-2X^Ty=0$$

$$\cancel{2}X^TX\theta=\cancel{2}X^Ty$$

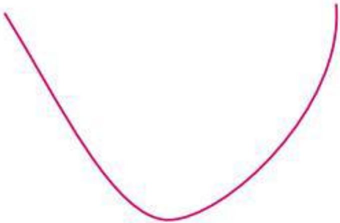
$$\theta=(X^TX)^{-1}X^Ty$$

$$AA^{-1}=I$$

$$(X^TX)^{-1}$$

Calculate gradient

CONVEX



$g(\theta)=5-(\theta-10)^2$

$$J(\theta)=(X\theta-y)^T(X\theta-y)$$

$$\nabla J(\theta)=2X^T(X\theta-y)$$

$$\theta=(X^TX)^{-1}X^Ty$$

h_1 30

h_2 60

D

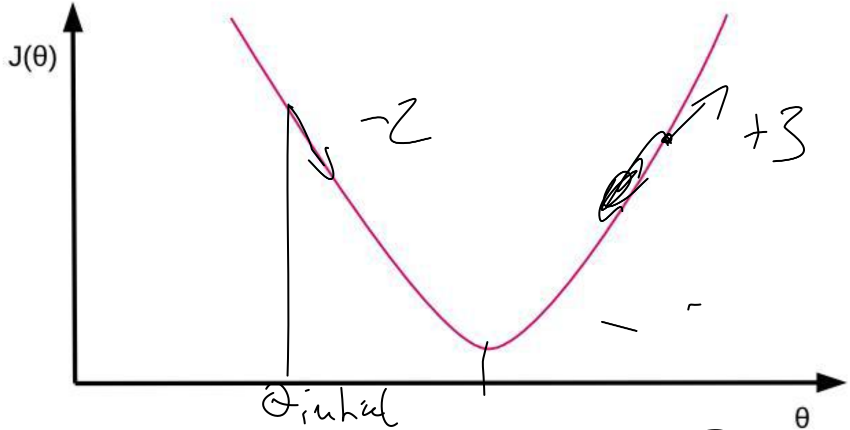
D

Invertible if: $h_2 = 2h_1$

Complexity of inverse: $O(D^3)$

Outline

Gradient descent

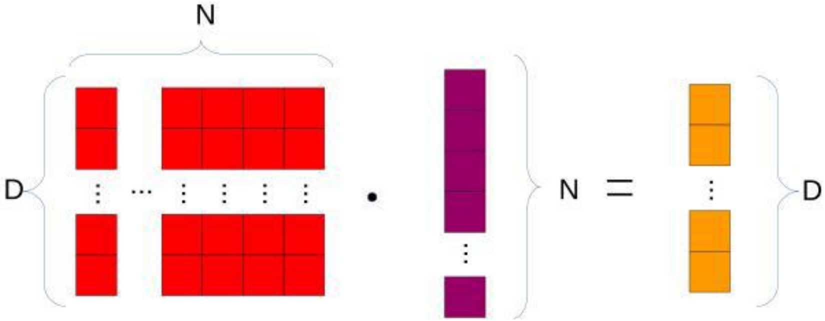


$$\theta = \theta - \frac{\partial J(\theta)}{\partial \theta}$$

$\leftarrow$

Gradient descent

$$\nabla J(\theta)=\begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta-y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta-y^{(i)}\right)x_1^{(i)} \end{bmatrix}=2X^T(X\theta-y)$$

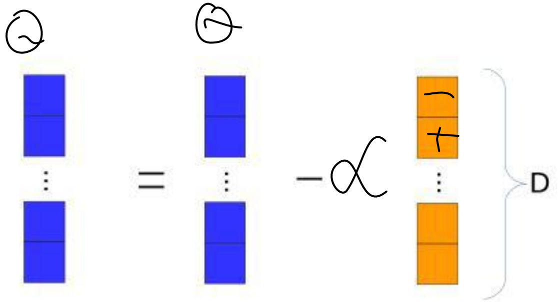


Gradient descent

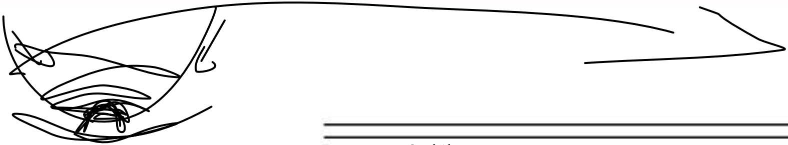
$$\nabla J(\theta) = \begin{bmatrix} 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_0^{(i)} \\ 2\sum_{i=1}^{N}\left(\left(x^{(i)}\right)^T\theta - y^{(i)}\right)x_1^{(i)} \end{bmatrix} = 2X^T(X\theta - y)$$

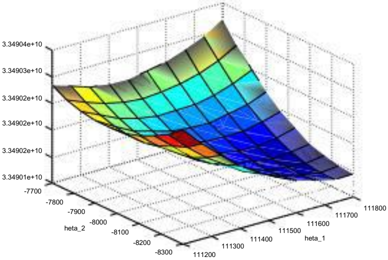
0 − 1

0 −



Gradient descent

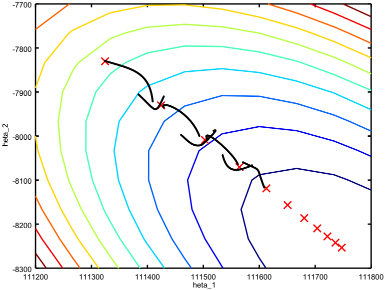




---

while $|\frac{\partial J(\theta)}{\partial \theta}| > \epsilon$ **do**
    for $i \in \{1, \ldots, D\}$ **do**
        $\frac{\partial J(\theta)}{\partial \theta_i} = 2 \sum_j^N (\theta x^{(j)} - y^{(j)}) x_i^{(j)}$
    end
    for $i \in \{1, \ldots, D\}$ **do**
        $\theta_i = \theta_i - \alpha \frac{\partial J(\theta)}{\partial \theta_i}$
    end
end

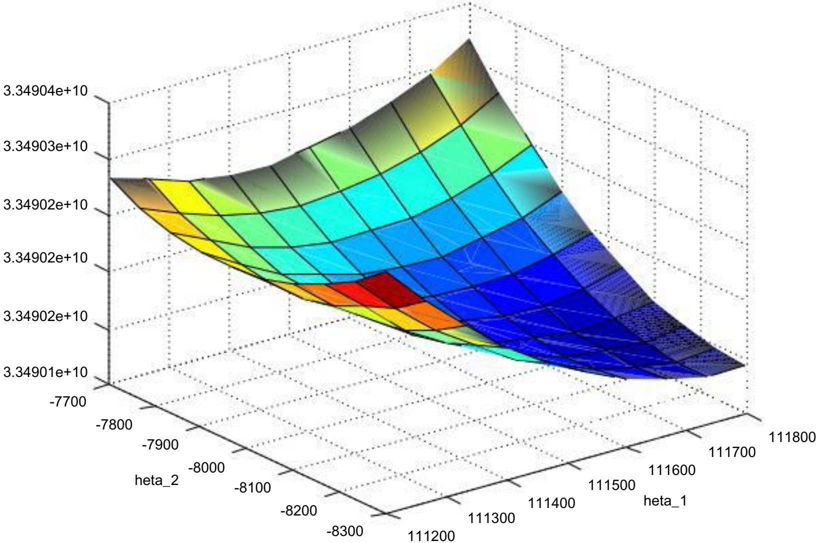
1000



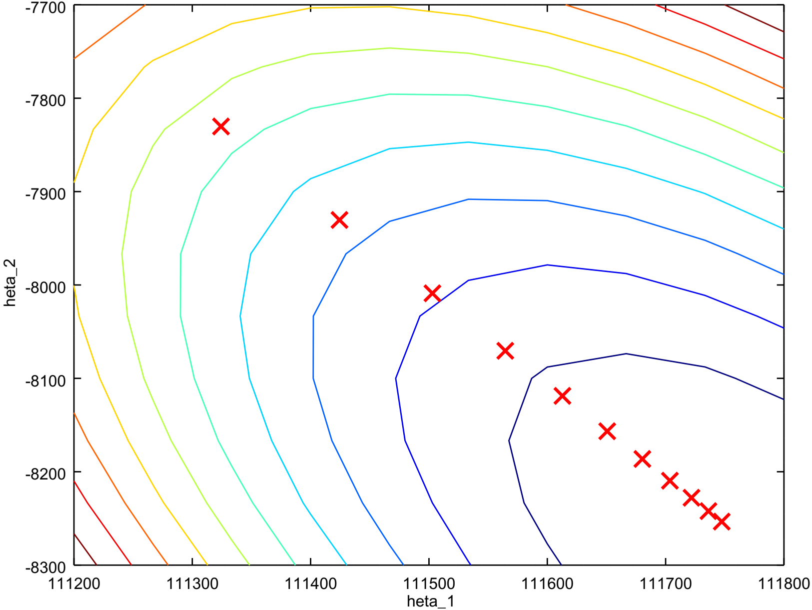
while $|\frac{\partial J(\theta)}{\partial \theta}| > \epsilon$ **do**
  for $i \in \{1, \ldots, D\}$ **do**
    $\frac{\partial J(\theta)}{\partial \theta_i} = 2\sum_j^N(\theta x^{(j)} - y^{(j)})x_i^{(j)}$
  end
  for $i \in \{1, \ldots, D\}$ **do**
    $\theta_i = \theta_i - \alpha\frac{\partial J(\theta)}{\partial \theta_i}$
  end
end

Outline

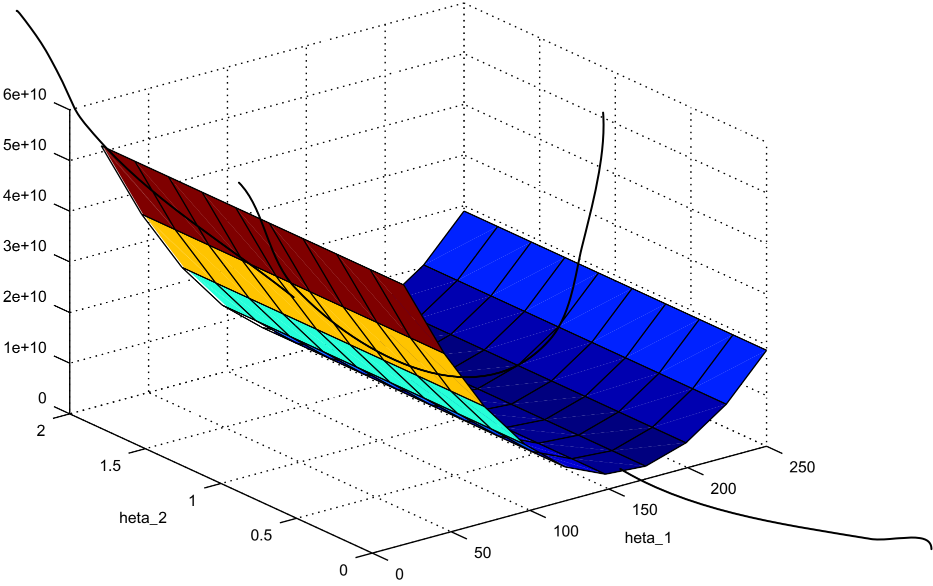
Convergence of gradient descent



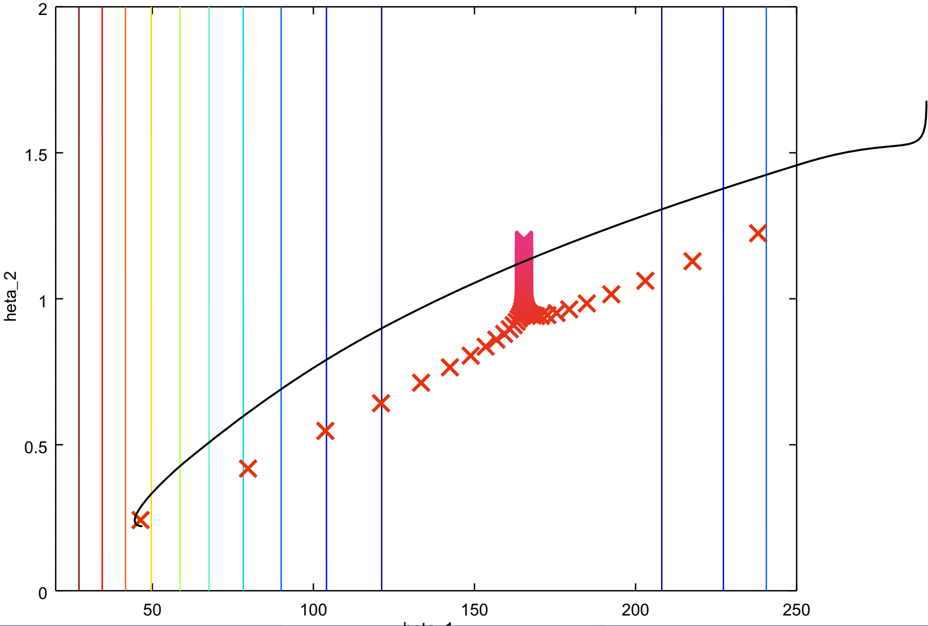
Convergence of gradient descent



Convergence of gradient descent



Convergence of gradient descent



Features scaling

Min-Max scaling

- Values afetr scaling are within fixed range. Usually $[0; 1]$.
- We use following equation:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Mean standarization

- Values are oriented around zero-mean and standard deviation 1
- We use the following equation:

$$x_i = \frac{x_i - \mu}{\sigma}$$

Thank you!

Szymon Bobek
Institute of Applied Computer Science
AGH University of Science and Technology
21 March 2017

http://geist.agh.edu.pl