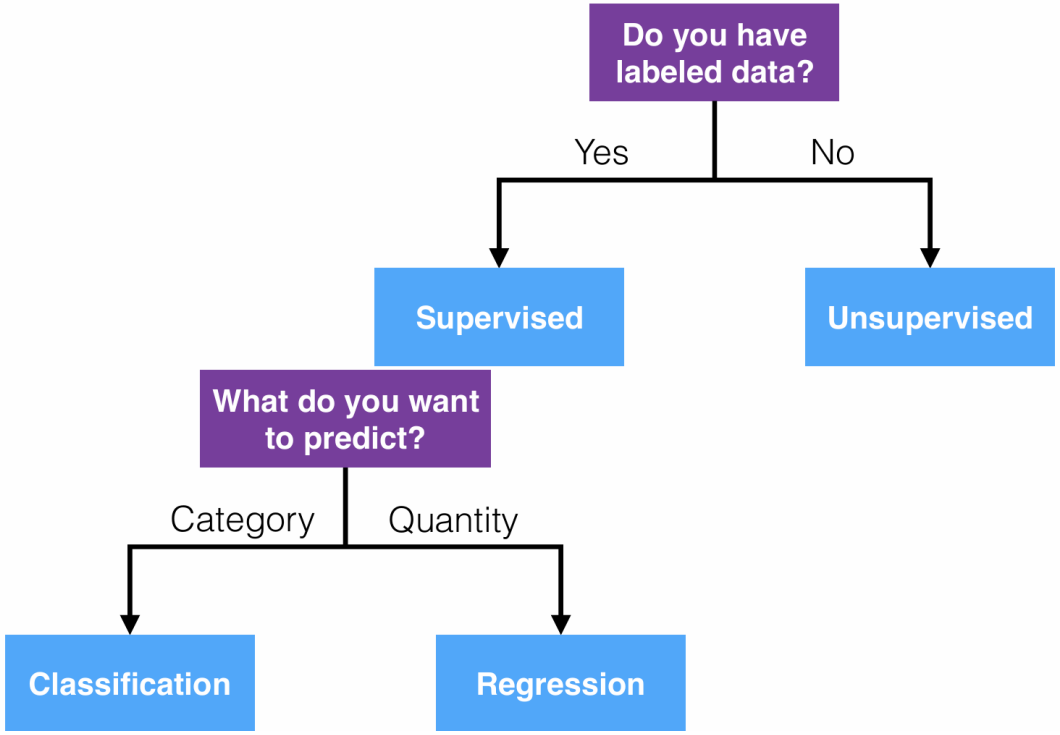
Linear Regression Models

FAIRLEIGH DICKINSON UNIVERSITY

Recall: Types of Machine Learning



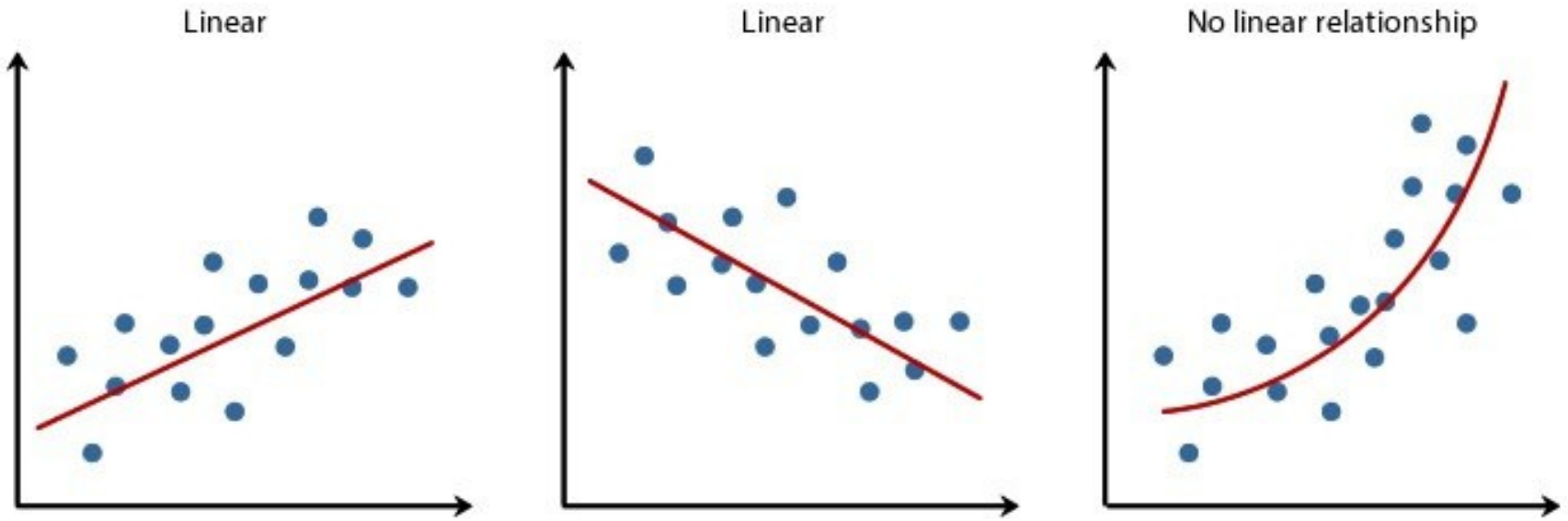
FAIRLEIGH DICKINSON UNIVERSITY

Agenda

- What is linear regression?
- The least-squares (closed-form) solution
  - Simple linear regression
  - Polynomial regression
  - Multivariate linear regression
- Solve linear regression by optimization
  - Gradient descent

What is linear Regression?

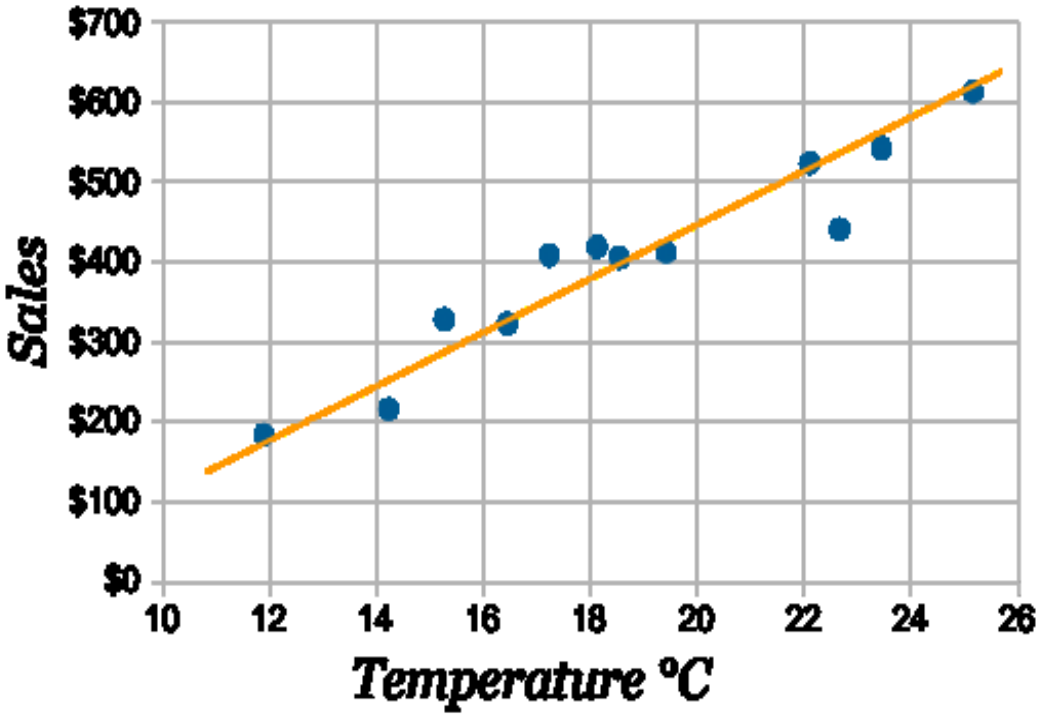
Given a set of observed values of the independent (input) variables and the corresponding values of the dependent (output) variable, determine a relation between the independent variable(s) and a continuous output variable.



FAIRLEIGH DICKINSON UNIVERSITY

Linear regression

- Examples



Linear regression

- Examples

Prices of used cars: example data for regression

Price (US$)	Age (years)	Distance (km)	Weight (pounds)
13500	23	46986	1165
13750	23	72937	1165
13950	24	41711	1165
14950	26	48000	1165
13750	30	38500	1170
12950	32	61000	1170
16900	27	94612	1245
18600	30	75889	1245
21500	27	19700	1185
12950	23	71138	1105

FAIRLEIGH DICKINSON UNIVERSITY

Linear regression

General Approach

- Regression function

$$y = f(x, \theta)$$

- Objective

  Optimize θ such that the approximation error is minimized

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

y_i : target $\hat{y}_i$: model output

Example

$$\text{Price} = a_0 + a_1 \cdot \text{Age} + a_2 \cdot \text{Distance} + a_3 \cdot \text{Weight}$$

$$x = \{\text{Age}, \text{Distance}, \text{Weight}\}$$

$$\theta = \{a_1, a_2, a_3\}$$

Prices of used cars: example data for regression

Price (US$)	Age (years)	Distance (km)	Weight (pounds)
13500	23	46986	1165
13750	23	72937	1165
13950	24	41711	1165
14950	26	48000	1165
13750	30	38500	1170
12950	32	61000	1170
16900	27	94612	1245
18600	30	75889	1245
21500	27	19700	1185
12950	23	71138	1105

FAIRLEIGH DICKINSON UNIVERSITY

The least-squares(closed-form) solution

Different linear regression models

- Simple linear regression
  - Only one continuous independent variable

$$y = a + bx$$

- Polynomial regression
  - Only one continuous independent variable

$$y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$$

- Multivariate linear regression
  - More than one independent variables

$$y = a_0 + a_1 x_1 + \ldots + a_n x_n$$

FAIRLEIGH DICKINSON UNIVERSITY

Simple linear regression

FAIRLEIGH DICKINSON UNIVERSITY

- Ordinary least squares

$$y = \alpha + \beta x$$

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

α is the **y-intercept**.
β is the **slope**.
$\hat{y}_i$: is the **predicted value** of y for a given x_i.
y_i is the **actual (target) value**.

- Objective function

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$= \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

$\partial E / \partial \alpha = 0$

$$\sum_{i=1}^{n} y_i = n\alpha + \beta \sum_{i=1}^{n} x_i$$

$\partial E / \partial \beta = 0$

$$\sum_{i=1}^{n} x_i y_i = \alpha \sum_{i=1}^{n} x_i + \beta \sum_{i=1}^{n} x_i^2$$

Simple linear regression

- Ordinary least squares

$$y = \alpha + \beta x$$

α is the **y-intercept**.
β is the **slope**.
$\hat{y}_i$: is the **predicted value** of y for a given x_i.
y_i is the **actual (target) value**.

Final solution:

$$\beta = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

$$\alpha = \bar{y} - \beta \bar{x}$$

$$\bar{x} = \frac{1}{n} \sum x_i$$

$$\bar{y} = \frac{1}{n} \sum y_i$$

$$\mathrm{Var}(x) = \frac{1}{n-1} \sum (x_i - \bar{x}_i)^2$$

$$\mathrm{Cov}(x, y) = \frac{1}{n-1} \sum (x_i - \bar{x})(y_i - \bar{y})$$

FAIRLEIGH DICKINSON UNIVERSITY

Simple linear regression

FAIRLEIGH DICKINSON UNIVERSITY

- Example

$$y = \alpha + \beta x$$

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

$$n = 5$$

$$\bar{x} = \frac{1}{5}(1.0 + 2.0 + 3.0 + 4.0 + 5.0) = 3.0$$

$$\bar{y} = \frac{1}{5}(1.00 + 2.00 + 1.30 + 3.75 + 2.25) = 2.06$$

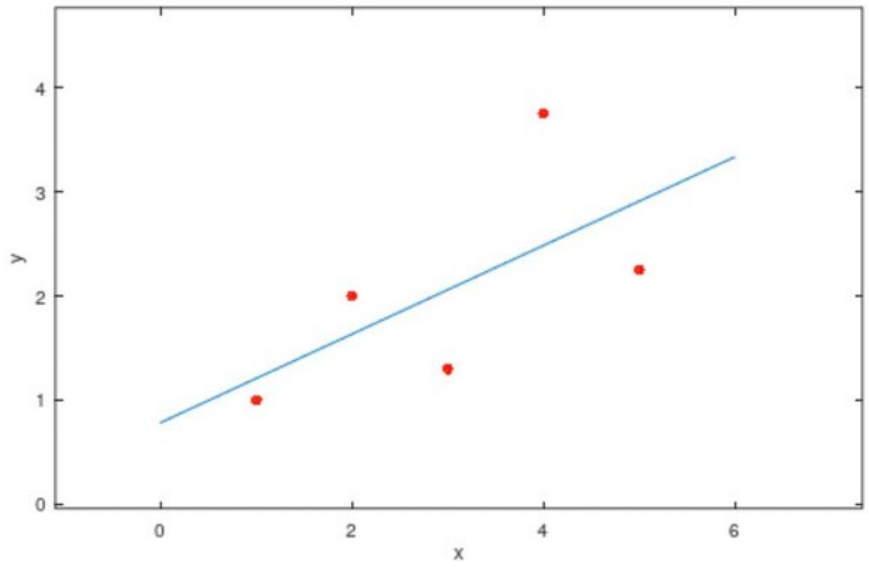
$$\mathrm{Cov}(x, y) = \frac{1}{4}[(1.0 - 3.0)(1.00 - 2.06) + \cdots + (5.0 - 3.0)(2.25 - 2.06)] = 1.0625$$

$$\mathrm{Var}(x) = \frac{1}{4}\left[(1.0 - 3.0)^2 + \cdots + (5.0 - 3.0)^2\right] = 2.5$$

$$b = \frac{1.0625}{2.5} = 0.425$$

$$a = 2.06 - 0.425 \times 3.0 = 0.785$$

$$y = 0.785 + 0.425x$$



Polynomial regression

FAIRLEIGH DICKINSON UNIVERSITY

- Model

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$$

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

- Ordinary least squares
  - Objective function

$$E = \sum_{i=1}^{n} \left[y_i - \left(\alpha_0 + \alpha_1 x_i + \alpha_2 x_i^2 + \cdots + \alpha_k x_i^k \right) \right]^2$$

  - Solution can be obtained by solving

$$\frac{\partial E}{\partial \alpha_i} = 0, \quad \forall i = 0, 1, \ldots, k.$$

Polynomial regression

FAIRLEIGH DICKINSON UNIVERSITY

- Ordinary least squares

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \cdots + \alpha_k x^k$$

$$\frac{\partial E}{\partial \alpha_i} = 0, \quad \forall i = 0, 1, \ldots, k.$$

$$\begin{aligned}
\sum y_i &= \alpha_0 n + \alpha_1 \left(\sum x_i\right) + \cdots + \alpha_k \left(\sum x_i^k\right) \\
\sum y_i x_i &= \alpha_0 \left(\sum x_i\right) + \alpha_1 \left(\sum x_i^2\right) + \cdots + \alpha_k \left(\sum x_i^{k+1}\right) \\
\sum y_i x_i^2 &= \alpha_0 \left(\sum x_i^2\right) + \alpha_1 \left(\sum x_i^3\right) + \cdots + \alpha_k \left(\sum x_i^{k+2}\right) \\
&\vdots \\
\sum y_i x_i^k &= \alpha_0 \left(\sum x_i^k\right) + \alpha_1 \left(\sum x_i^{k+1}\right) + \cdots + \alpha_k \left(\sum x_i^{2k}\right)
\end{aligned}$$

$$\vec{\alpha} = \left(D^T D\right)^{-1} D^T \vec{y}$$

$$\vec{y} = D\vec{\alpha}$$

$$\vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad D = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^k \\ 1 & x_2 & x_2^2 & \cdots & x_2^k \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^k \end{bmatrix}, \text{ and } \vec{\alpha} = \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_k \end{bmatrix}$$

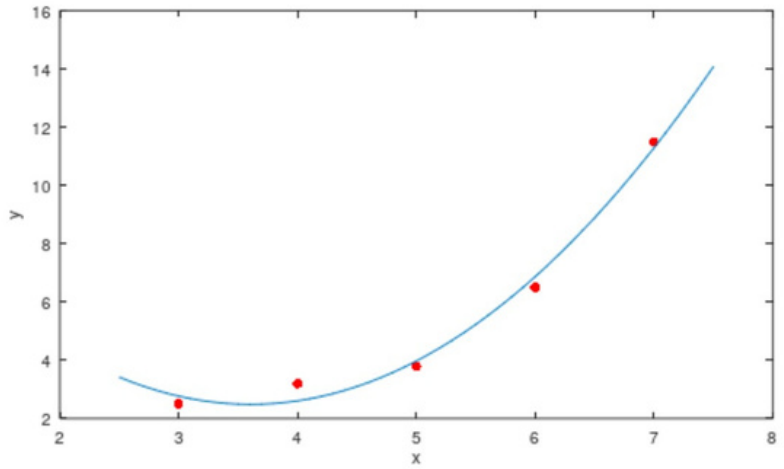
Polynomial regression

FAIRLEIGH DICKINSON UNIVERSITY

- Example

x	3.0	4.0	5.0	6.0	7.0
y	2.5	3.2	3.8	6.5	11.5

$$y = \alpha_0 + \alpha_1 x + \alpha_2 x^2$$



$$\sum y_i = n\alpha_0 + \alpha_1 \left(\sum x_i\right) + \alpha_2 \left(\sum x_i^2\right)$$
$$\sum y_i x_i = \alpha_0 \left(\sum x_i\right) + \alpha_1 \left(\sum x_i^2\right) + \alpha_2 \left(\sum x_i^3\right)$$
$$\sum y_i x_i^2 = \alpha_0 \left(\sum x_i^2\right) + \alpha_1 \left(\sum x_i^3\right) + \alpha_2 \left(\sum x_i^4\right)$$

$$27.5 = 5\alpha_0 + 25\alpha_1 + 135\alpha_2$$
$$158.8 = 25\alpha_0 + 135\alpha_1 + 775\alpha_2$$
$$966.2 = 135\alpha_0 + 775\alpha_1 + 4659\alpha_2$$

$$\alpha_0 = 12.4285714$$
$$\alpha_1 = -5.5128571$$
$$\alpha_2 = 0.7642857$$

$$y = 12.4285714 - 5.5128571x + 0.7642857x^2$$

Multivariate linear regression

FAIRLEIGH DICKINSON UNIVERSITY

- Model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_N x_N$$

- Ordinary least squares

Variables	Values (examples)			
	Example 1	Example 2	...	Example n
x_1	x_{11}	x_{12}	...	x_{1n}
x_2	x_{21}	x_{22}	...	x_{2n}
...				
x_N	x_{N1}	x_{N2}	...	x_{Nn}
y (outcomes)	y_1	y_2	...	y_n

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{N1} \\ 1 & x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{Nn} \end{bmatrix}, \text{ and } B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}$$

$$B = \left(X^T X\right)^{-1} X^T Y$$

Multivariate linear regression

FAIRLEIGH DICKINSON UNIVERSITY

- Example

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

x_1	1	1	2	0
x_2	1	2	2	1
y	3.25	6.5	3.5	5.0

$$Y = \begin{bmatrix} 3.25 \\ 6.5 \\ 3.5 \\ 5.0 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 2 & 2 \\ 1 & 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$$y = 2.0625 - 2.3750x_1 + 3.2500x_2$$

$$X^T X = \begin{bmatrix} 4 & 4 & 6 \\ 4 & 6 & 7 \\ 6 & 7 & 10 \end{bmatrix} \Rightarrow (X^T X)^{-1} = \begin{bmatrix} \frac{11}{4} & \frac{1}{2} & -2 \\ \frac{1}{2} & 1 & -1 \\ -2 & -1 & 2 \end{bmatrix} \Rightarrow B = (X^T X)^{-1} X^T Y = \begin{bmatrix} 2.0625 \\ -2.3750 \\ 3.2500 \end{bmatrix}$$

Gradient Descent

Objective function of linear regression

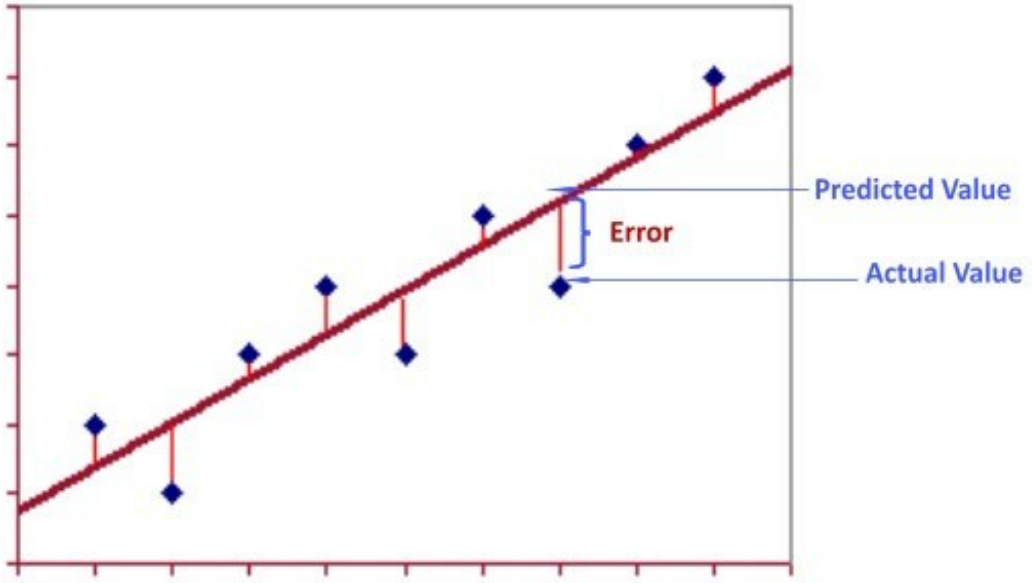
- Linear regression

$$y = f(x, \theta)$$

- Objective function
  - Sum of squared error

$$E = \sum_{i=0}^{n} (y_i - \hat{y}_i)^2$$

y_i :actual value(target) $\hat{y}_i$: predicted value



FAIRLEIGH DICKINSON UNIVERSITY

Objective function of linear regression

FAIRLEIGH DICKINSON UNIVERSITY

- Linear regression

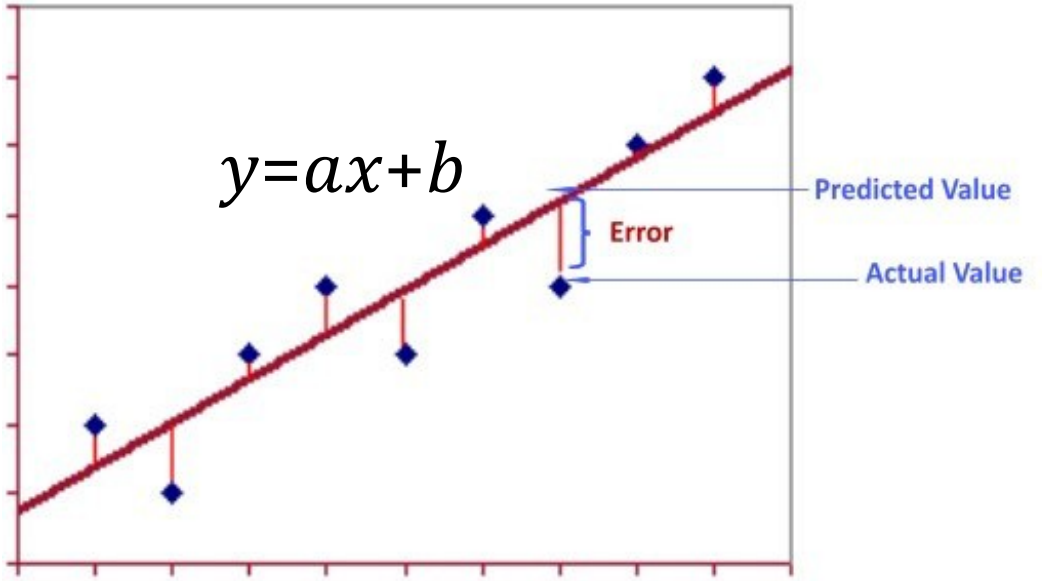
$$y = f(x, \theta)$$

- Objective function
  - Sum of squared error

$$E = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$

  - Solution

$$min \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$



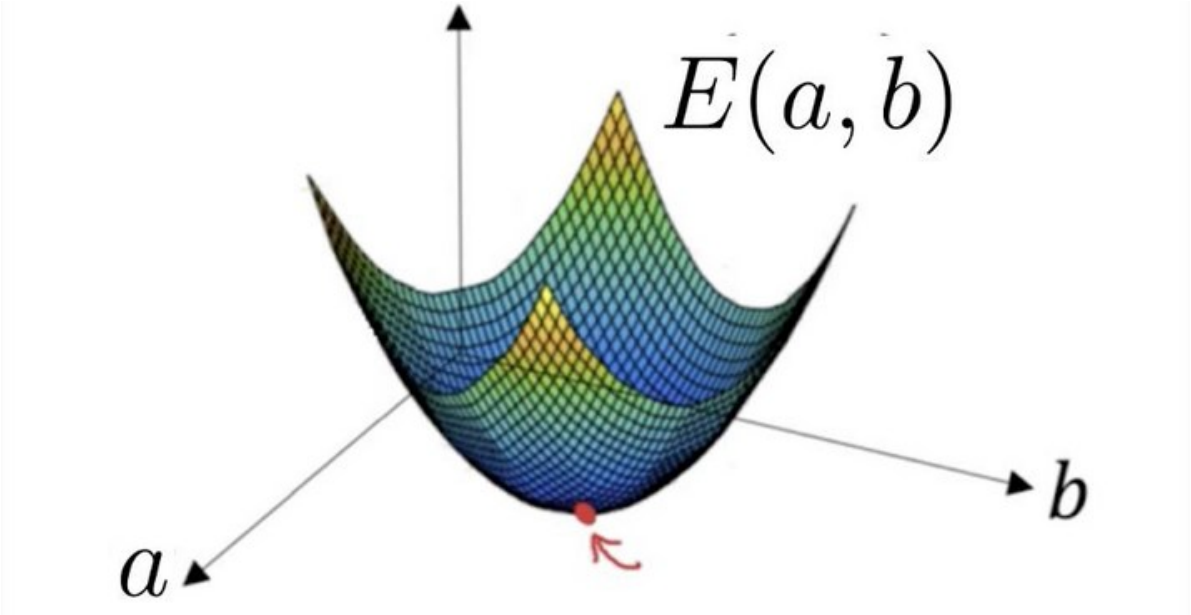
Objective function of linear regression

- Example
- Objective function
  - Sum of squared error

$$E = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$

  - Solution

$$min \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$

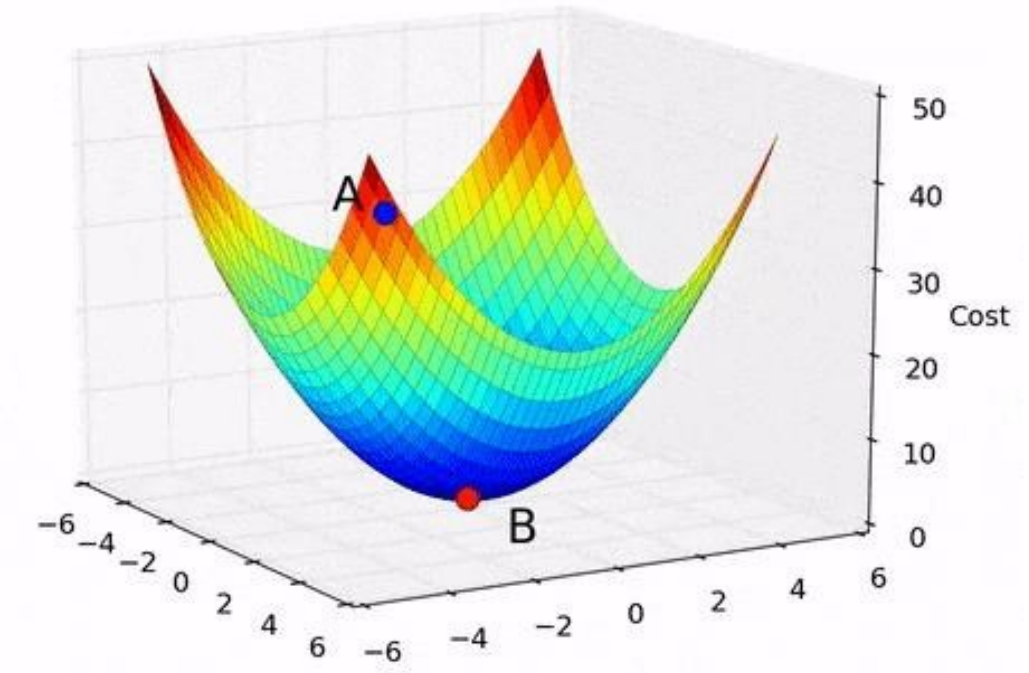


FAIRLEIGH DICKINSON UNIVERSITY

Gradient descent

- Basic idea
  - Take repeated steps in **steepest descent direction** until lowest point is reached
    - The **opposite** direction of the **gradient** of the function at the current point

$$\nabla f(\vec{p}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\vec{p}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\vec{p}) \end{bmatrix}$$



FAIRLEIGH DICKINSON UNIVERSITY

Gradient

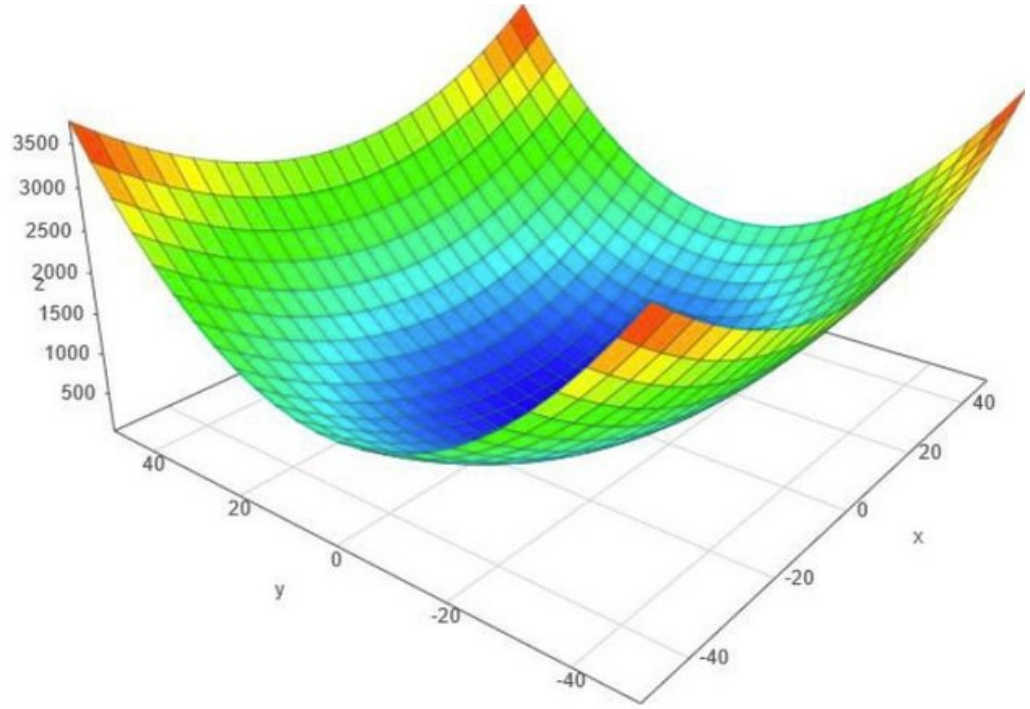
- Example

$$f(x,y) = 0.5x^2 + y^2$$

$$\nabla f(x,y) = \begin{bmatrix} \frac{\partial f}{\partial x}(x,y) \\ \\ \frac{\partial f}{\partial y}(x,y) \end{bmatrix} = \begin{bmatrix} x \\ \\ 2y \end{bmatrix}$$

The gradient at point *p*(10, 10)

$$\nabla f(10,10) = \begin{bmatrix} 10 \\ 20 \end{bmatrix}$$



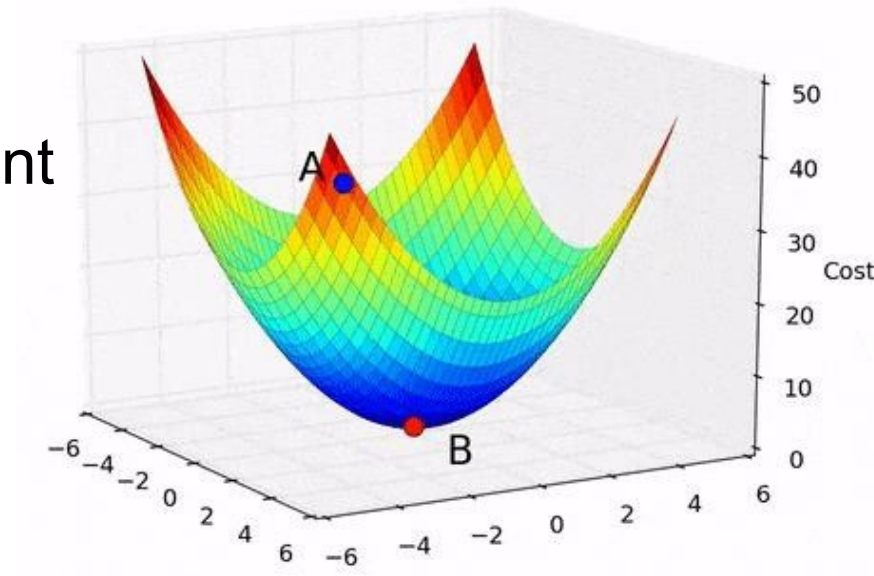
FAIRLEIGH DICKINSON UNIVERSITY

Gradient descent algorithm

- Main steps

1. Start from an initial guess (or even randomly)
2. Calculate the gradient of the function at current point
3. Make a scaled step in the opposite direction to the gradient

$$\vec{p}_{n+1} = \vec{p}_n - \eta \nabla f\left(\vec{p}_n\right)$$



4. Repeat 2) and 3) until one of the criteria is met
   - maximum number of iterations reached
   - step size (or the change of the function value) is smaller than a given tolerance

FAIRLEIGH DICKINSON UNIVERSITY

Gradient descent algorithm

- Example: 1d function

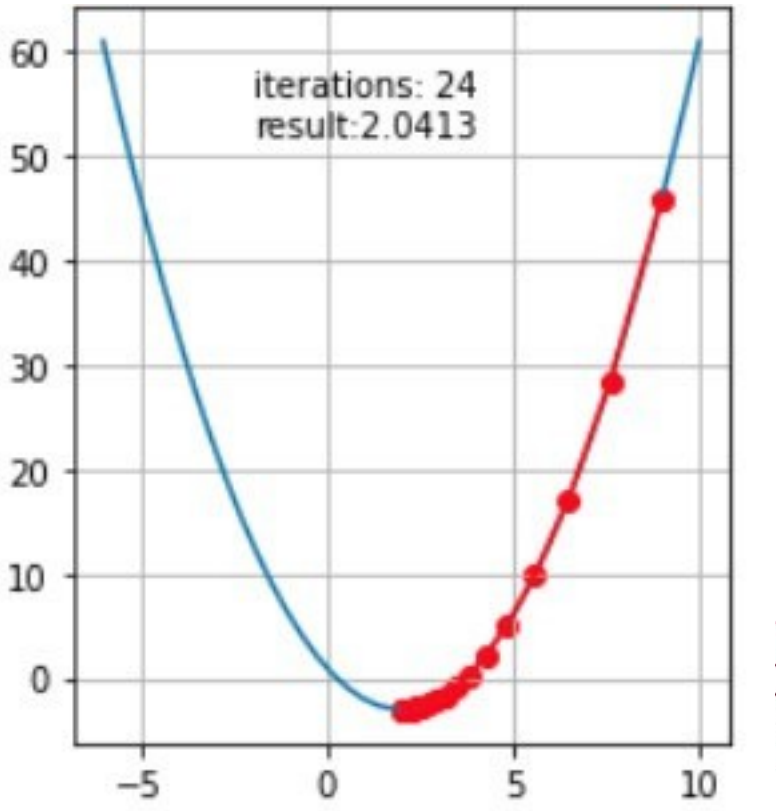
$$f(x) = x^2 - 4x + 1$$

$$\frac{df(x)}{dx} = 2x - 4$$

$$\vec{p}_{n+1} = \vec{p}_n - \eta \nabla f(\vec{p}_n)$$

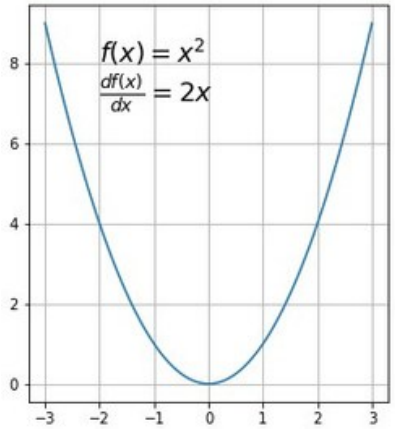
The first few steps

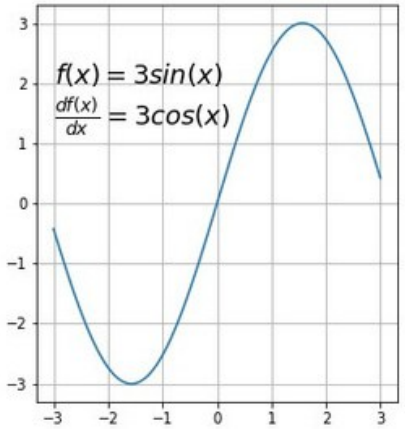
$$x_0 = 9, \qquad f(9) = 46$$
$$x_1 = 9 - 0.1 \times (2 \times 9 - 4) = 7.6, \qquad f(7.6) = 28.36$$
$$x_2 = 7.6 - 0.1 \times (2 \times 7.6 - 4) = 6.48, \qquad f(6.48) = 17.07$$
$$x_3 = 6.48 - 0.1 \times (2 \times 6.48 - 4) = 5.584, \qquad f(5.584) = 9.845$$
$$\ldots$$
$$x_{21} = 2.065, \quad f(2.065) = -2.996$$
$$x_{22} = 2.052, \quad f(2.052) = -2.997$$

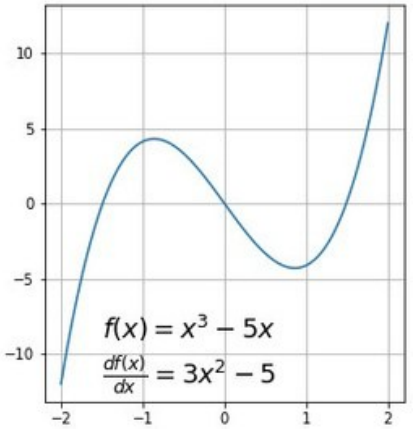


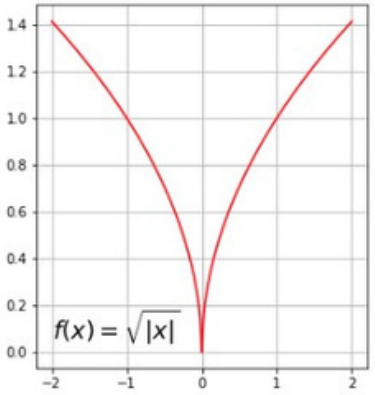
Function requirements

- Requirement #1) Differentiable

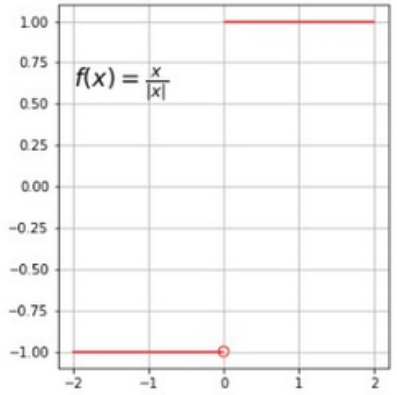




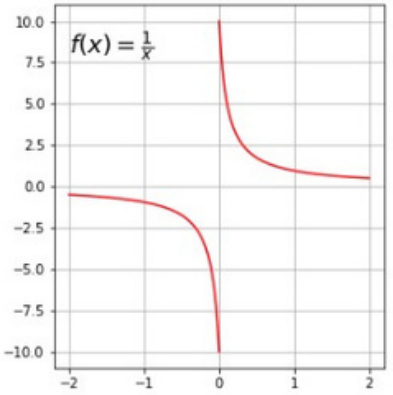




(a) Cusp



(b) Jump discontinuity



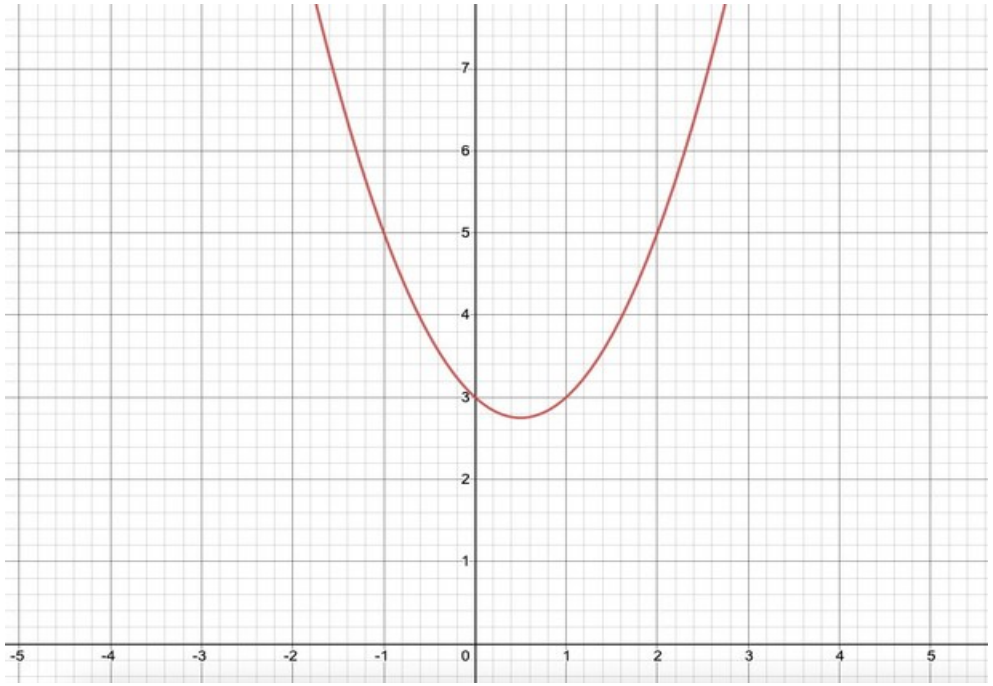
(c) Infinite discontinuity

Function requirements

- Example
  - Differentiable?
  - Convex?

$$f(x) = x^2 - x + 3$$

$$\frac{df(x)}{dx} = 2x - 1, \quad \frac{d^2 f(x)}{dx^2} = 2$$

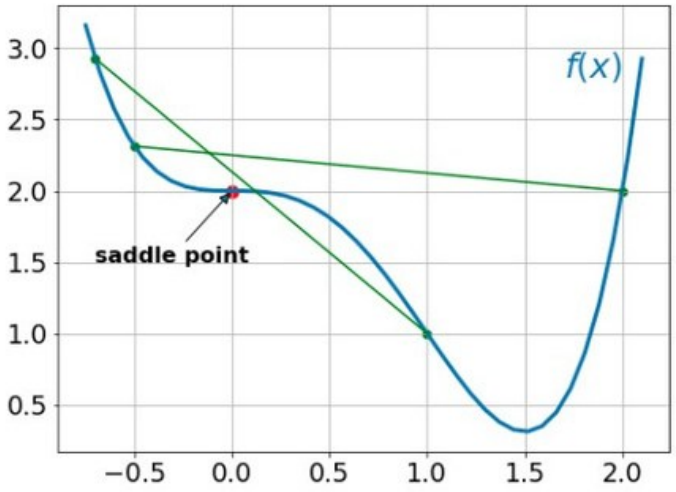


The function has derivative everywhere
The second derivative is always > 0

FAIRLEIGH DICKINSON UNIVERSITY

Function requirements

- Example
  - Differentiable?
  - Convex?



Example of a semi-convex function with a saddle point

$$f(x) = x^4 - 2x^3 + 2$$

$$\frac{df(x)}{dx} = 4x^3 - 6x^2 = x^2(4x - 6)$$

$$\frac{d^2f(x)}{dx^2} = 12x^2 - 12x = 12x(x - 1)$$

- for $x < 0$: function is convex
- for $0 < x < 1$: function is concave
- for $x > 1$: function is convex again

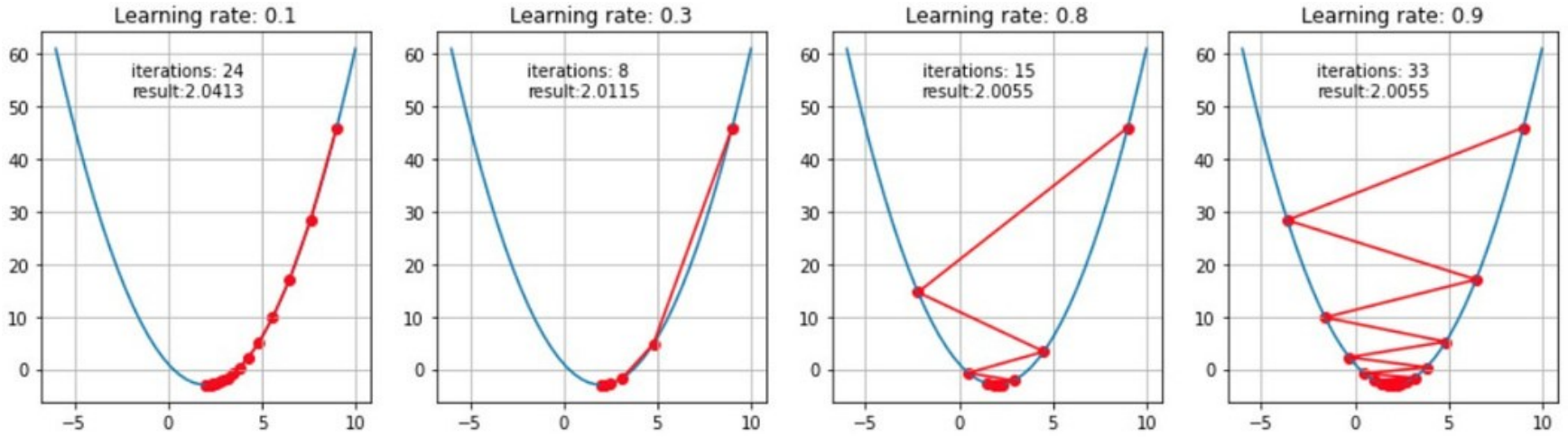
$x = 0$: saddle point
both first and second derivatives equal to zero

Challenges: learning rate

- Parameter update

$$\vec{p}_{n+1} = \vec{p}_n - \eta \nabla f(\vec{p}_n)$$

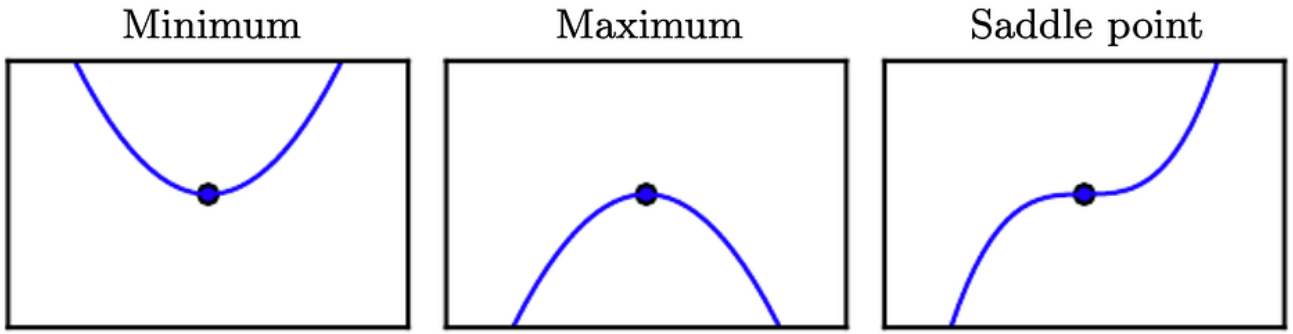
- Learning rate $\boldsymbol{\eta}$: scales the gradient and thus controls the step size
  - Too small: Too slow to converge; may reach maximum iteration before convergence
  - Too big: May not converge to the optimal point (jump around) or even to diverge completely

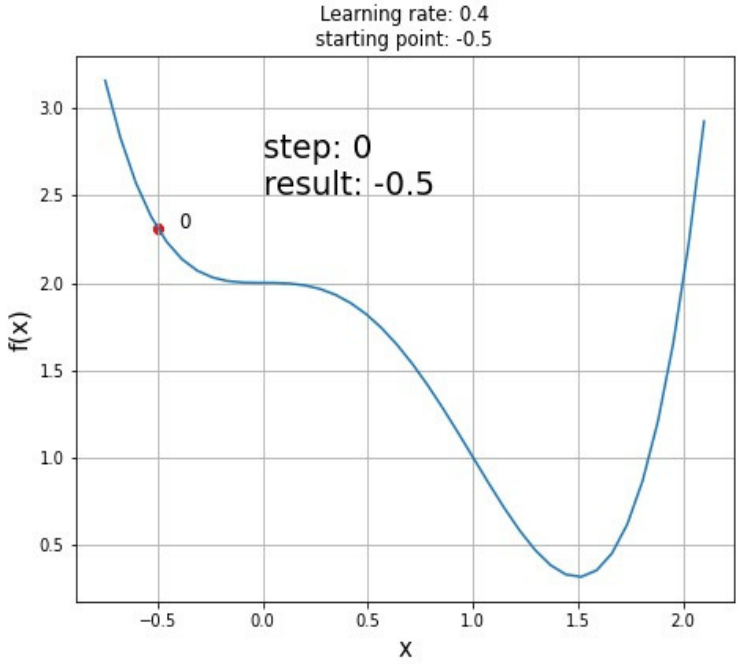


Challenges: saddle points

- Global minimum is not guaranteed

- Saddle point
  - Gradient = 0
  - Neither local minimum nor local maximum

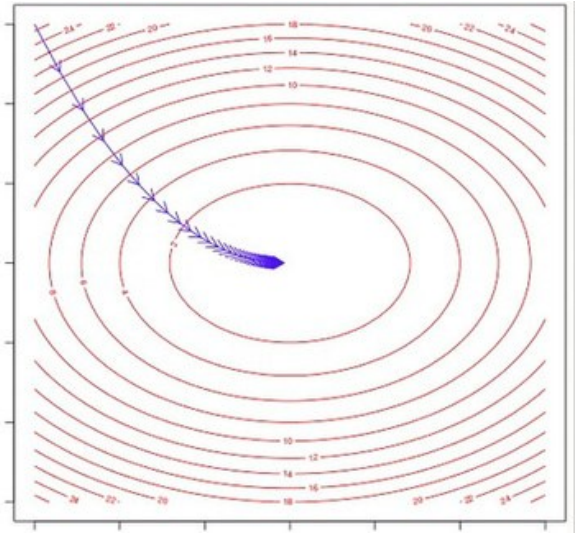


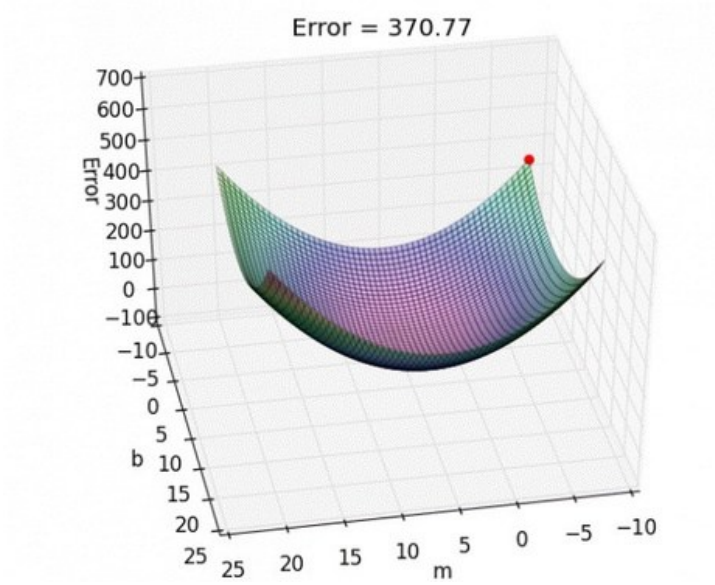


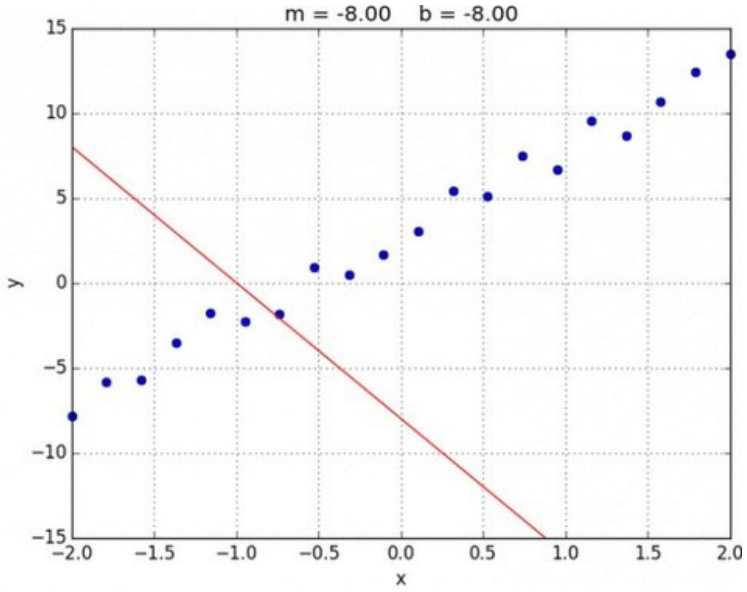
Solve linear regression using GD

- Objective function
  - Always convex

$$f(a,b) = \sum_{i=0}^{n} (y_i - (ax_i + b))^2$$







GD vs Least Square Solution (LSS)

FAIRLEIGH DICKINSON UNIVERSITY

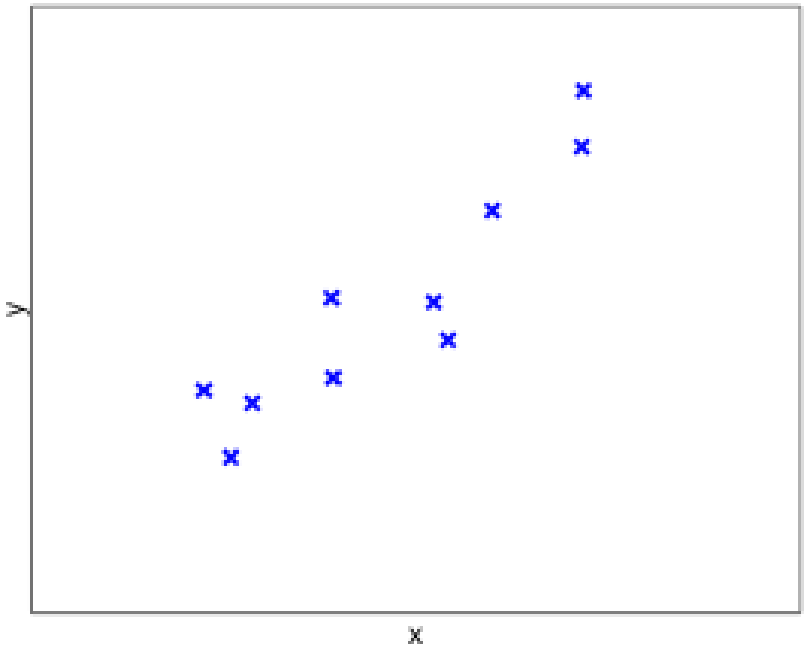
m training samples, n features

GD	LSS
Needs to choose the learning rate	No need to choose learning rate
Needs many iterations	Don't need to iterate
Works well even when n is large	Need to compute (X^TX), which takes about $O(n^3)$
	Slow if n is very large
	Some matrices (e.g., singular) are non-invertible

Recommendation: use the LSS if the number of features in less than 1000, otherwise the Gradient Descent.

Overfitting

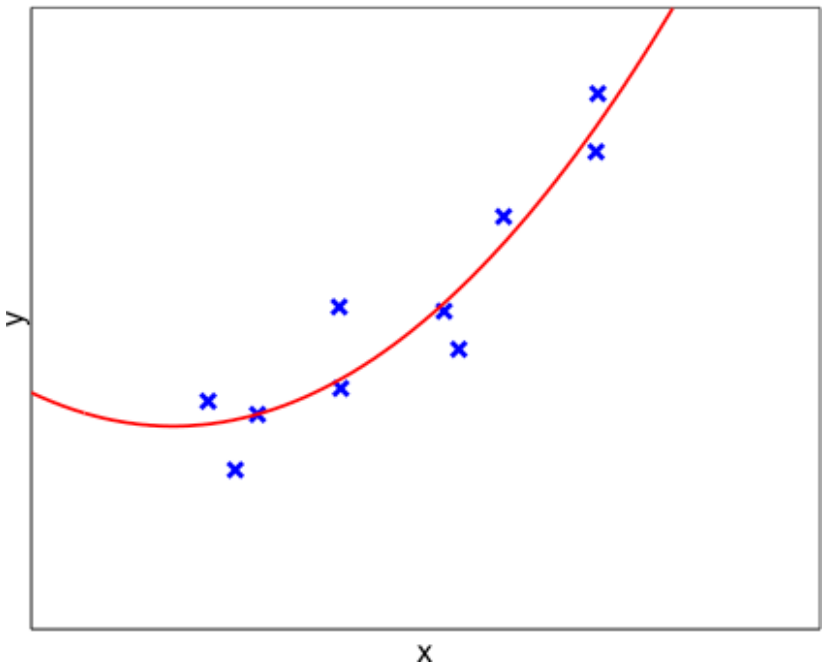
Example of linear regression with polynomial terms



$$f_w(x) = w_0 + w_1 x + w_2 x^2$$

$$X = \begin{matrix} x^2 & x & \\ \end{matrix} \begin{bmatrix} 0.75 & 0.86 & 1 \\ 0.01 & 0.09 & 1 \\ 0.73 & -0.85 & 1 \\ 0.76 & 0.87 & 1 \\ 0.19 & -0.44 & 1 \\ 0.18 & -0.43 & 1 \\ 1.22 & -1.10 & 1 \\ 0.16 & 0.40 & 1 \\ 0.93 & -0.96 & 1 \\ 0.03 & 0.17 & 1 \end{bmatrix} \quad Y = \begin{bmatrix} 2.49 \\ 0.83 \\ -0.25 \\ 3.10 \\ 0.87 \\ 0.02 \\ -0.12 \\ 1.81 \\ -0.83 \\ 0.43 \end{bmatrix}$$

Solving the problem

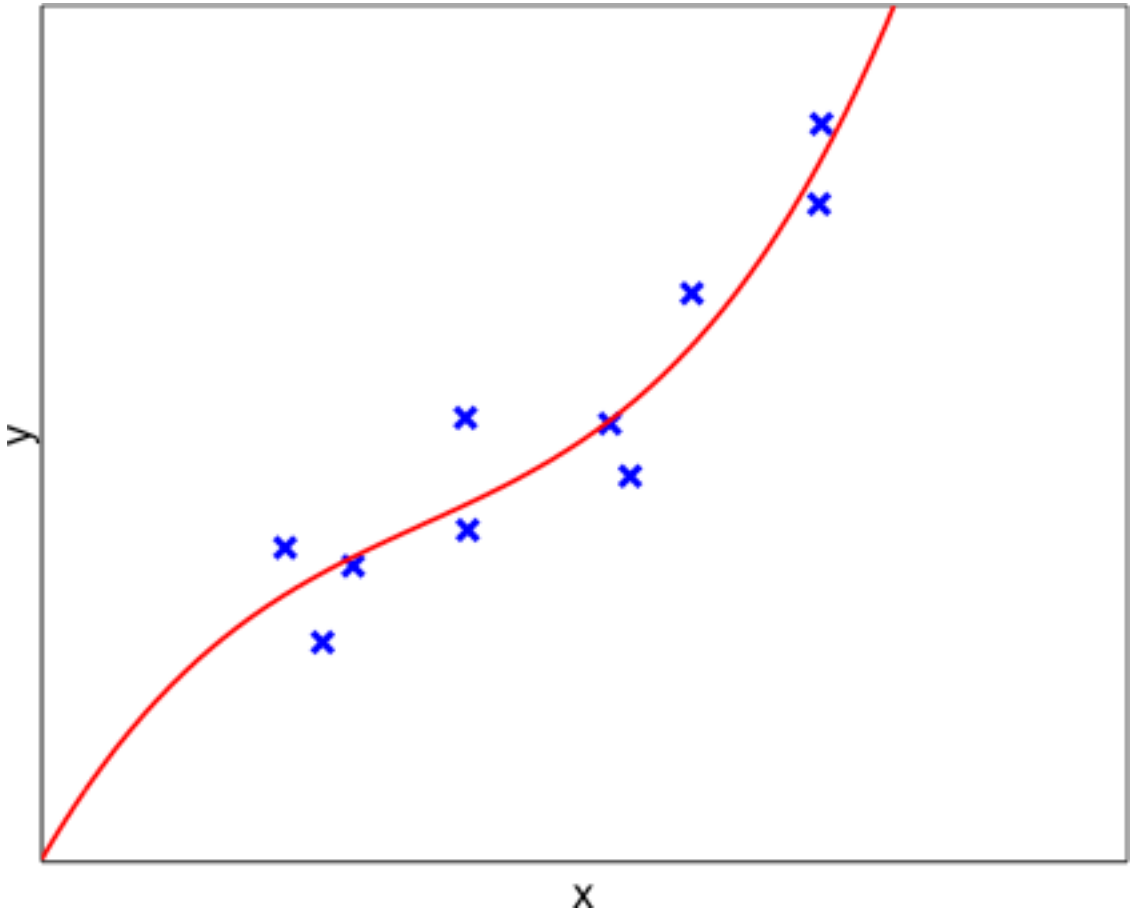


$$\mathbf{w} = (X^T X)^{-1} X^T Y = \begin{bmatrix} 4.11 & -1.64 & 4.95 \\ -1.64 & 4.95 & -1.39 \\ 4.95 & -1.39 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 3.60 \\ 6.49 \\ 8.34 \end{bmatrix} = \begin{bmatrix} 0.68 \\ 1.74 \\ 0.73 \end{bmatrix}$$

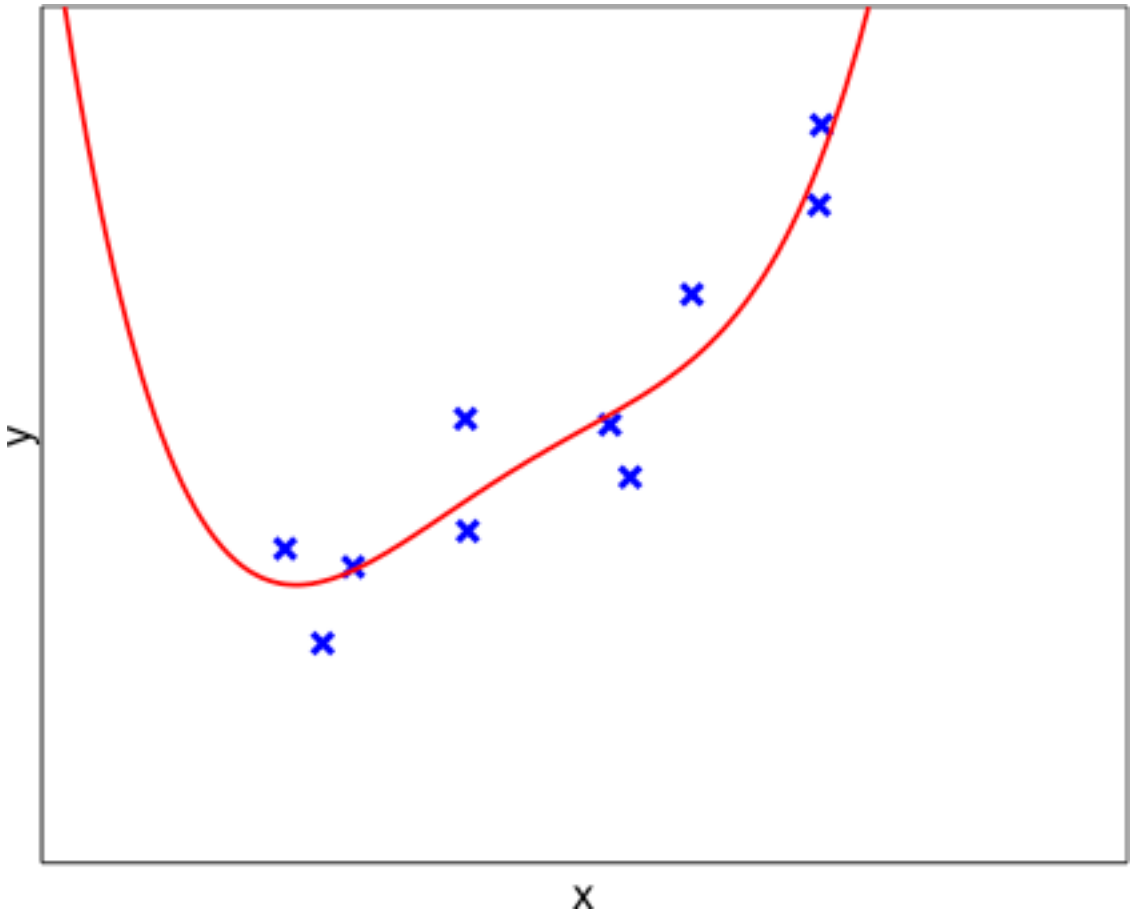
So the best order-2 polynomial is $y = 0.68x^2 + 1.74x + 0.73$.

Compared to $y = 1.6x + 1.05$ for the order-1 polynomial.

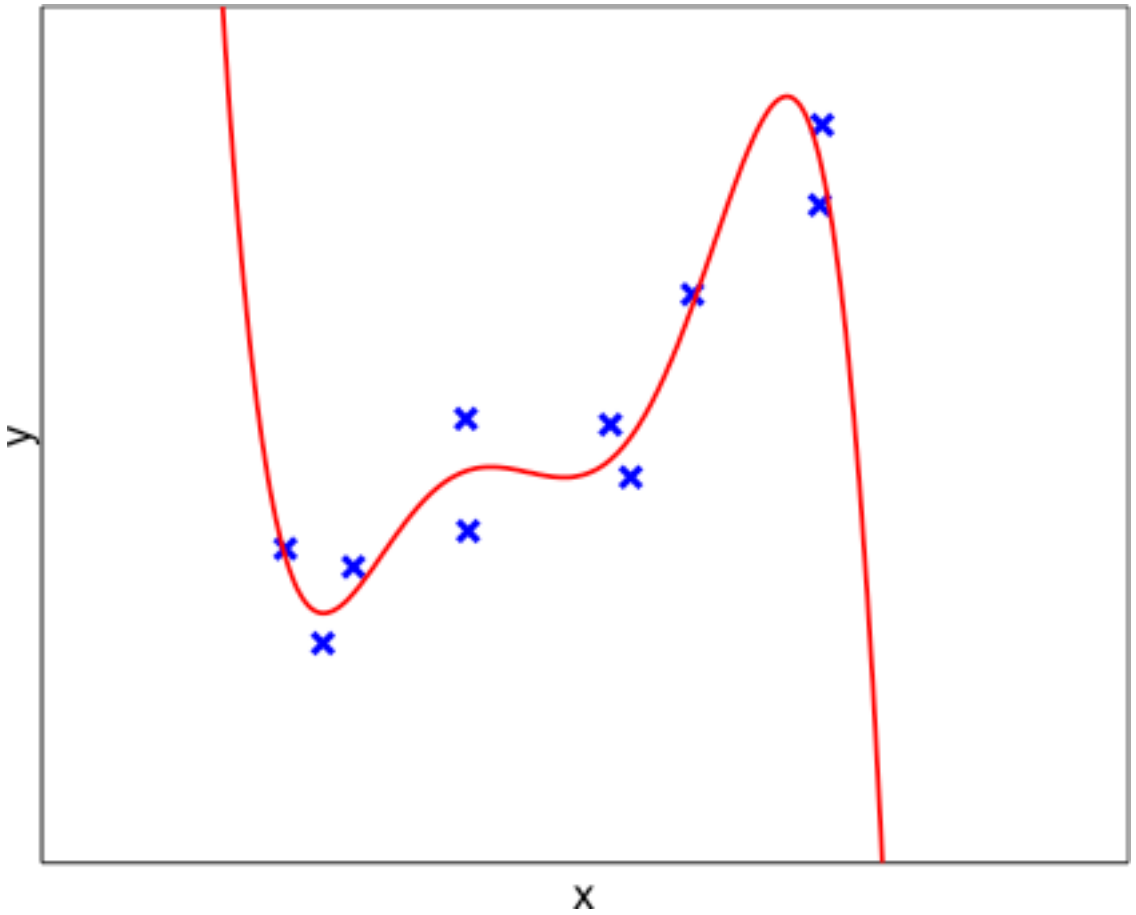
Order-3 fit: Is this better?



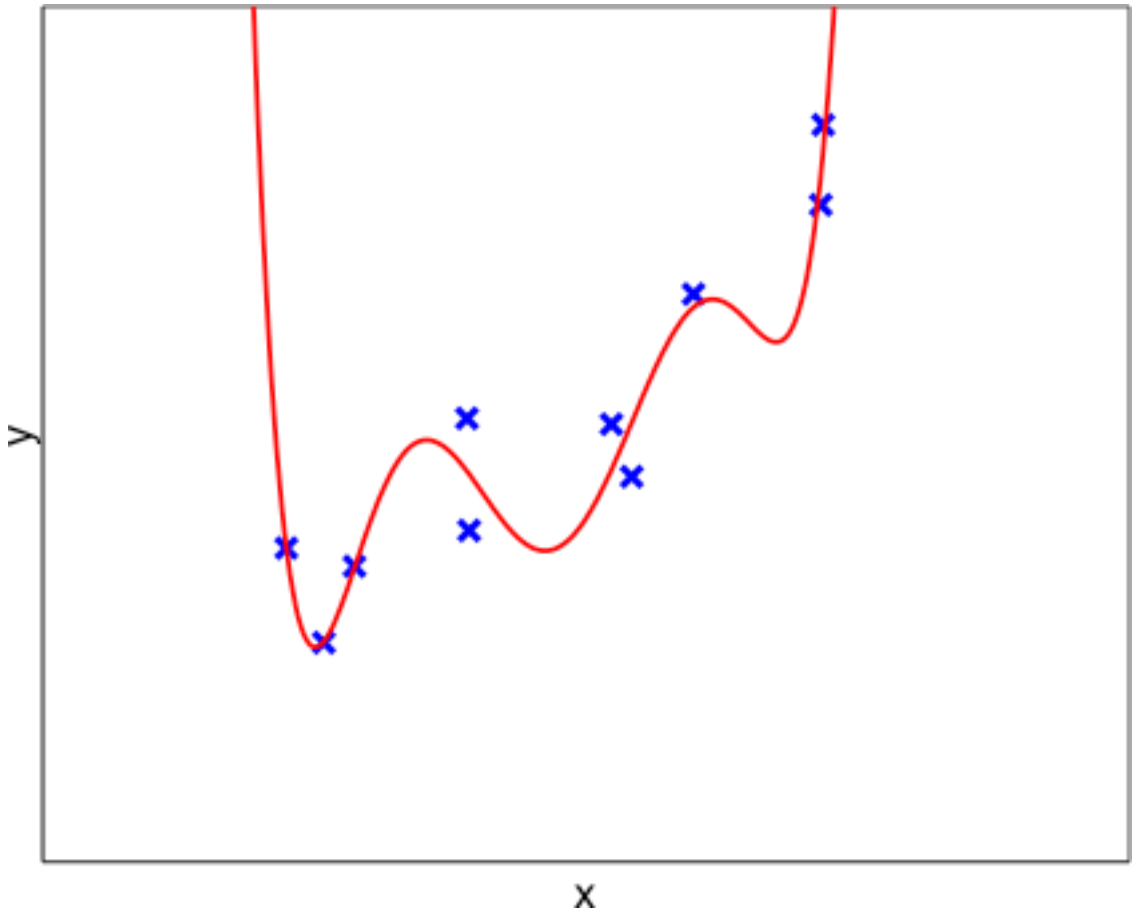
Order-4 fit



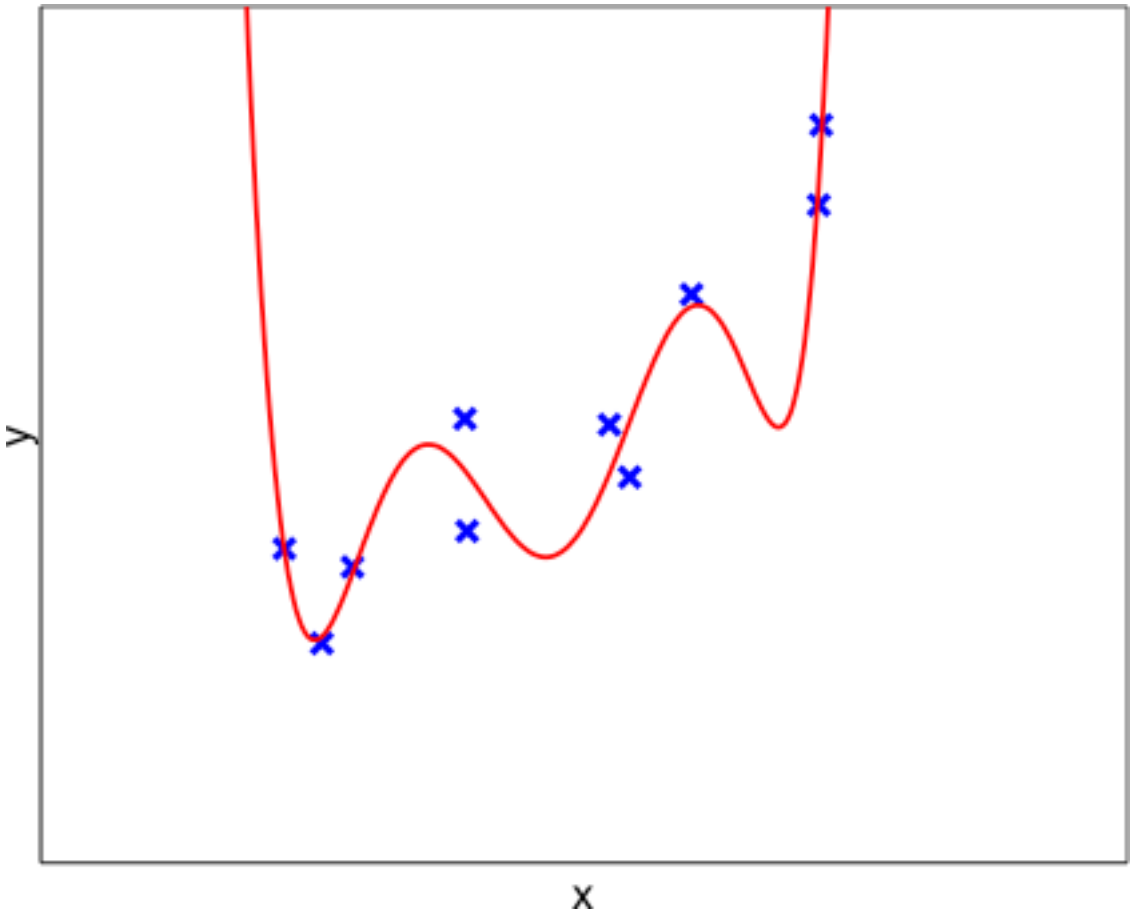
Order-5 fit



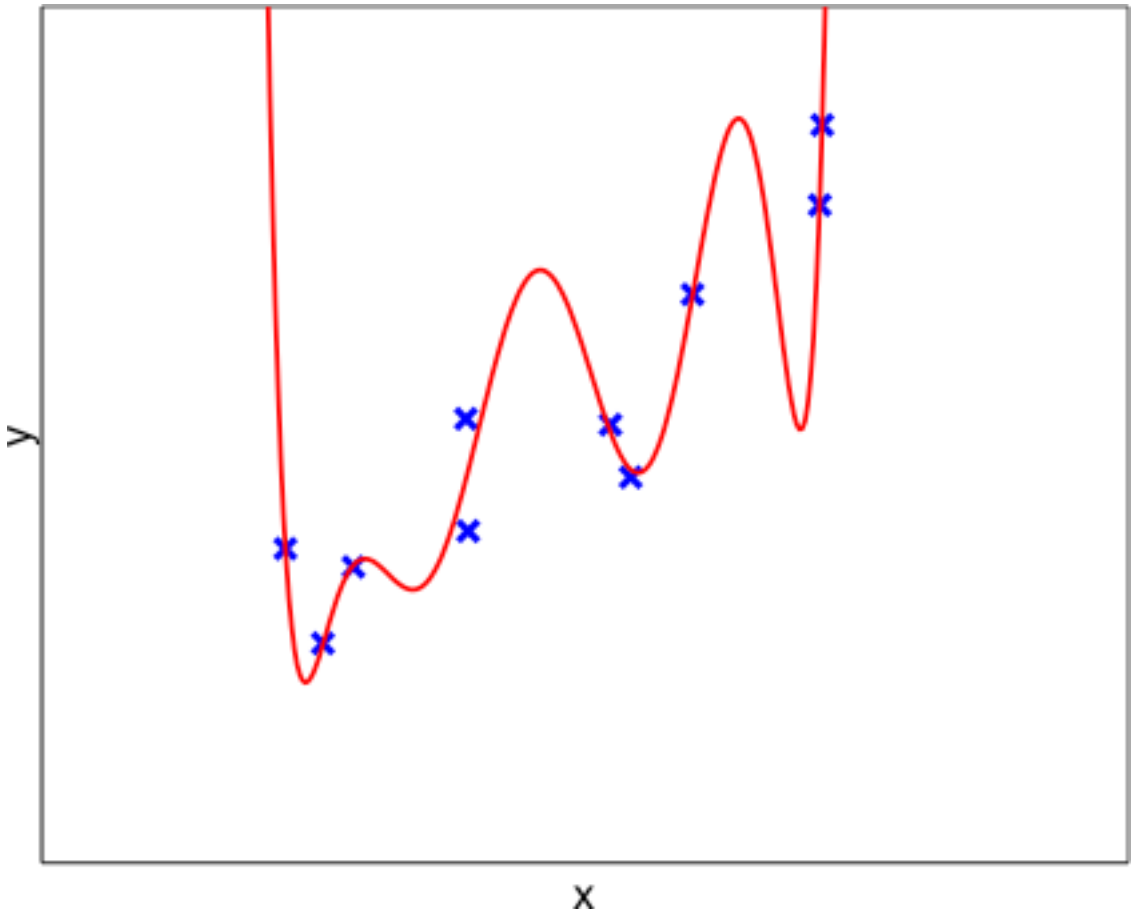
Order-6 fit



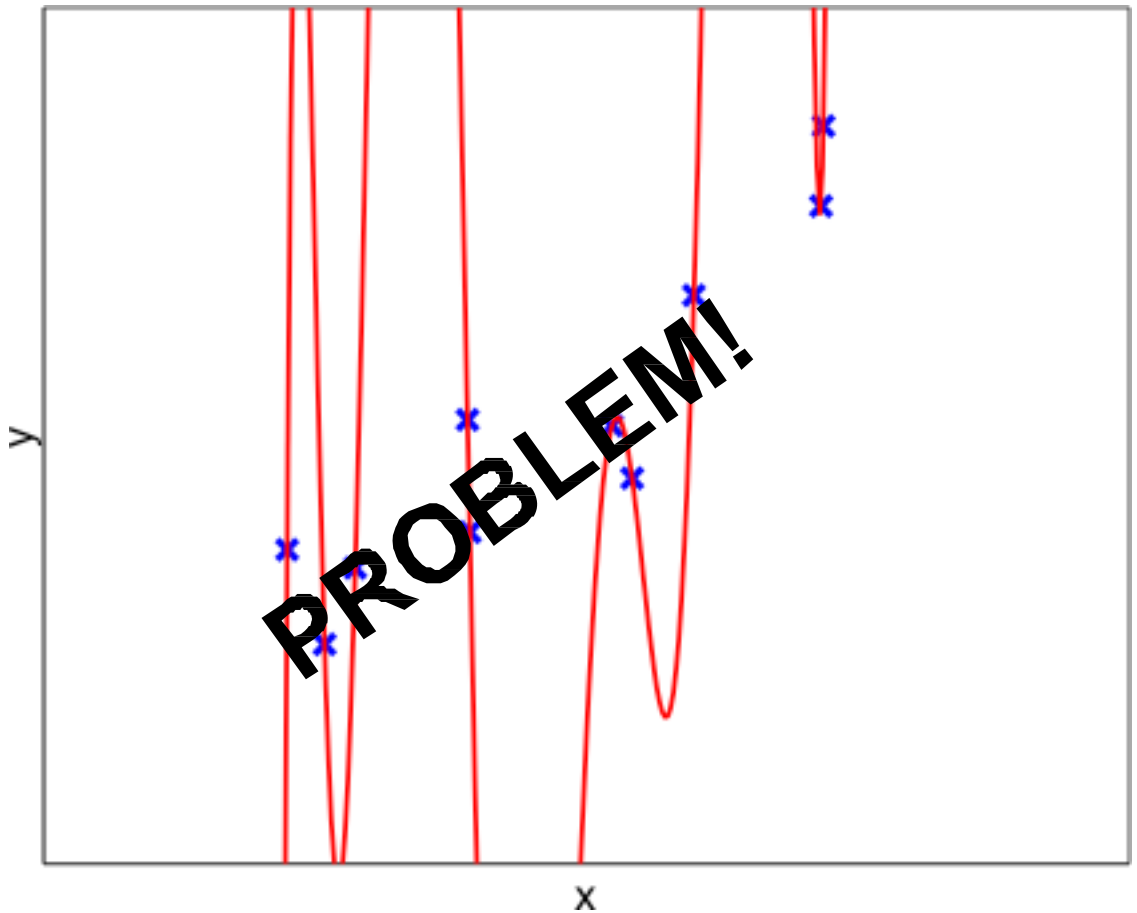
Order-7 fit



Order-8 fit



Order-9 fit

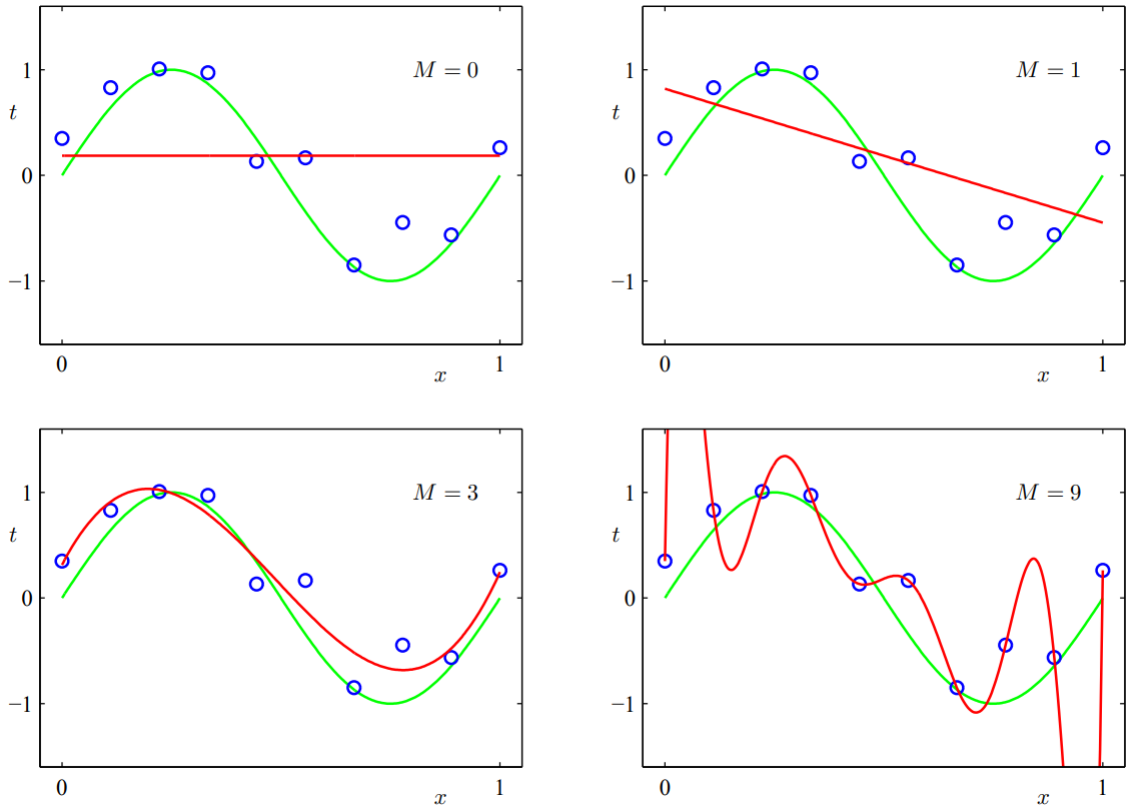


This is overfitting!

Overfitting

- We can find a hypothesis that explains perfectly the training data, but does not generalize well to new data.
- In this example: we have a lot of parameters (weights), so the hypothesis matches the data points exactly, but is wild everywhere else.
- A very important problem in machine learning.

Another overfitting Example



- The higher the degree of the polynomial M, the more degrees of freedom, and the more capacity to "overfit" the training data

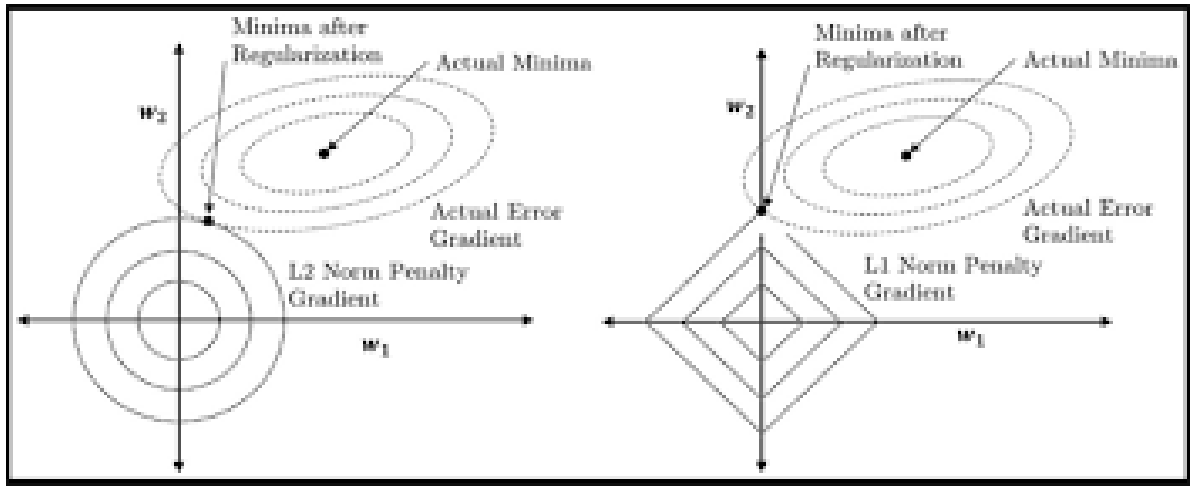
Next Lecture: Prevent Overfitting

L1 Regularization

$$\text{Modified loss function} = \text{Loss function} + \lambda \sum_{i=1}^{n} |W_i|$$

L2 Regularization

$$\text{Modified loss function} = \text{Loss function} + \lambda \sum_{i=1}^{n} W_i^2$$



FAIRLEIGH DICKINSON UNIVERSITY