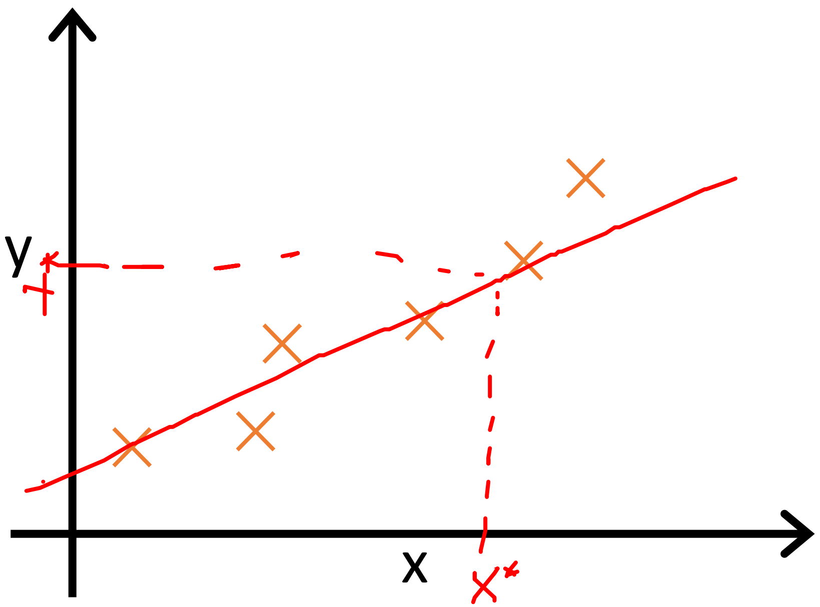
Linearna regresija

Uvod

- Nadgledano učenje – zadati su ulaz i za njih dobijeni izlazi (m primera)

$$\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$$



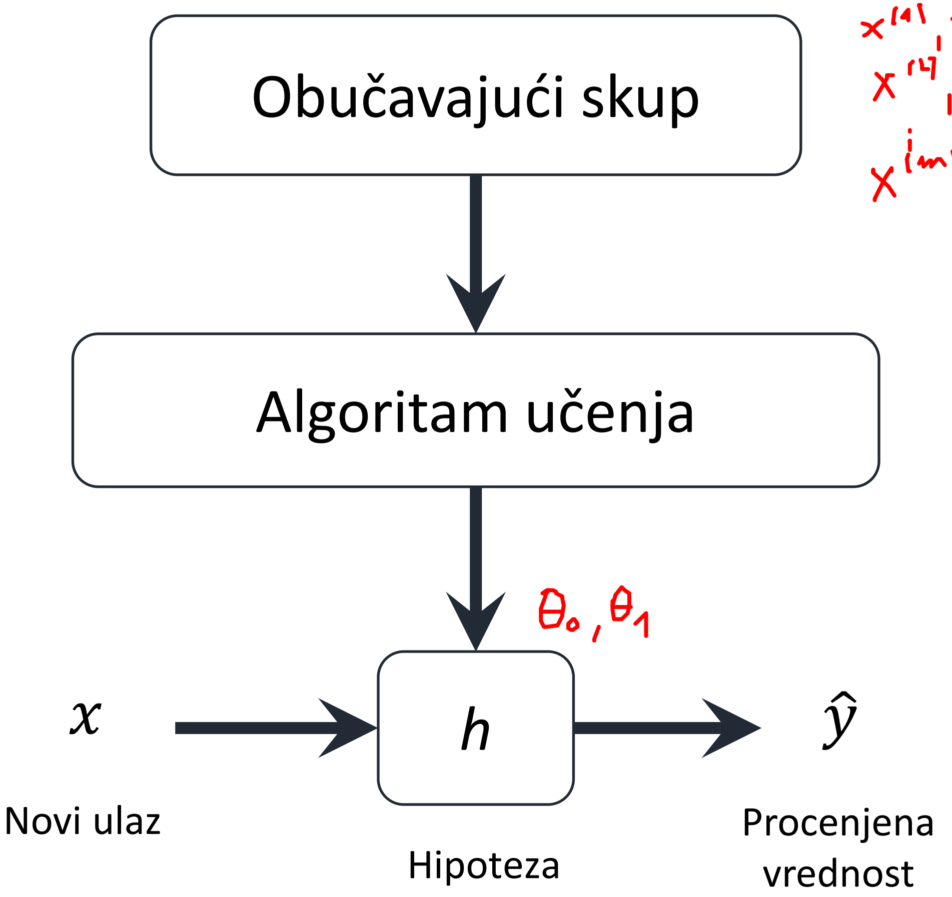
$X \xrightarrow{h} Y$

$X^* \rightarrow Y^*$

Linerna regresija

- Za jednostavniji slučaj sa jednim promenljivom (ulazom):

$$\hat{y} = h_\theta(x) = \theta_0 + \theta_1 x$$



$x^{(1)}, y^{(1)}$
$x^{(2)}, y^{(2)}$
$x^{(m)}, y^{(m)}$

$h_\theta(x) = \underset{\theta_0}{1} + \underset{\theta_1}{2}x$

$x^* = 5$
$y^* = 1 + 2 \cdot 5 = 11$

Optimizacioni kriterijum

- Definiše se kriterijum

$= h_{\theta_0,\theta_1}(x^{(i)})$

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

- Kako odrediti parametre θ_0 i θ_1 da se dobije min J ?

- Za minimizaciju greške se može koristiti gradijentni postupak

Gradijentni postupak

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

- Iterativni postupak

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

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

$x_0^{(i)} = 1$

- Primena na Linearnu regresiju sa jednom promenljivom: ponavlja se do konvergencije

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

$J(\theta) < \varepsilon$

$$\theta_1 = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)} \right) x_1^{(i)}$$

Linearna regresija za više promenljivih

- Hipoteza $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$
- Najčešće se uvodi dodatna promenljiva

$$x_0 = 1$$

Pa se dobije da je

$$h_\theta(x) = \theta^T x = x^T \theta = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n$$

gde je

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}, \qquad x = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$$

Gradijentni postupak za linearnu regresiju sa više promenljivih

- Za linearnu regresiju sa više promenljivih

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \qquad j=1,2,\dots,n$$

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

$$h_\theta\left(x^{(i)}\right) = \theta^T x^{(i)} = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)} + \cdots + \theta_n x_n^{(i)}, \qquad i = 1,2,\dots,m$$

pri čemu je $x_0^{(i)} = 1$

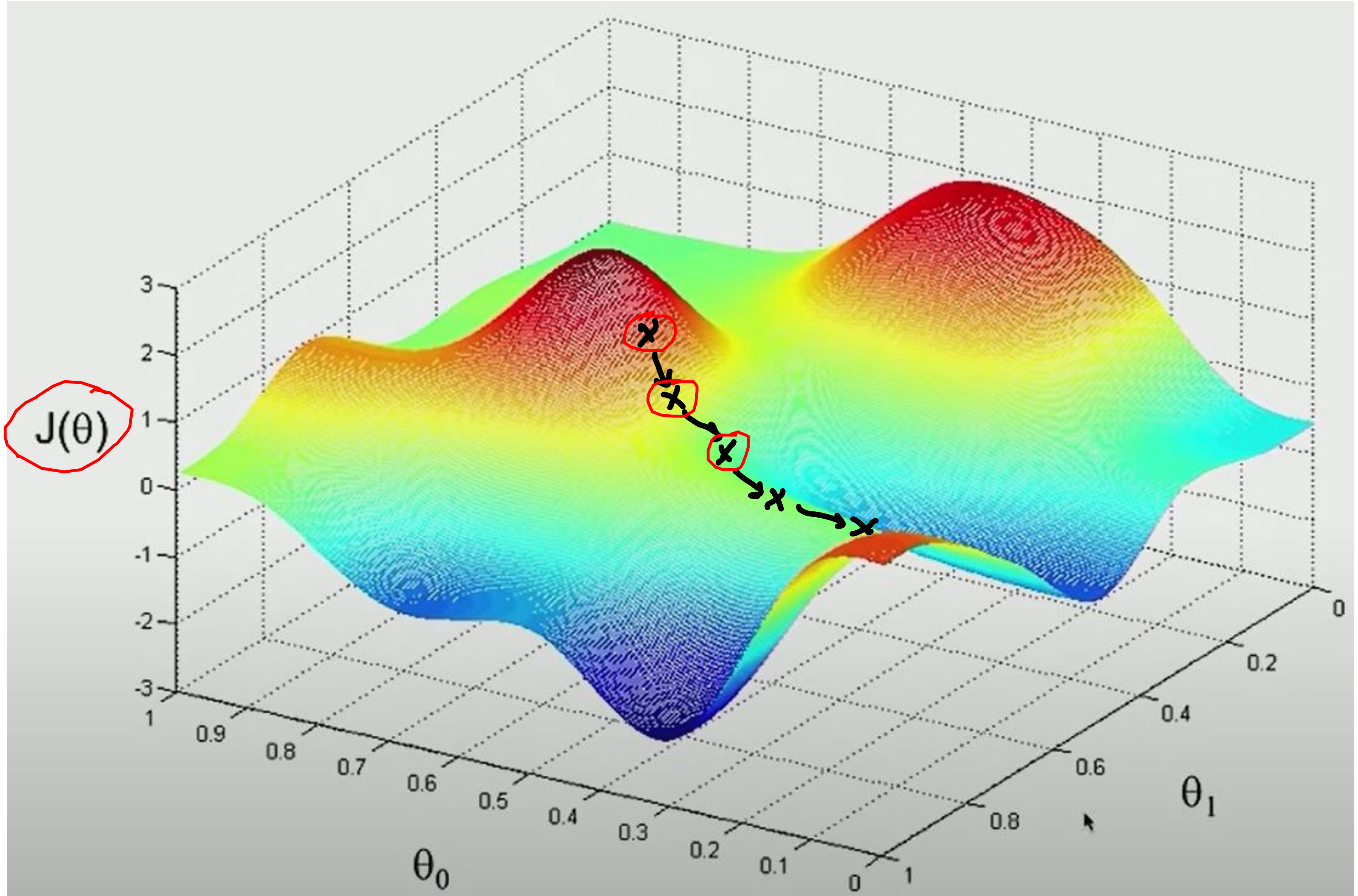
$$\theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_0^{(i)}, \qquad \theta_1 = \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_1^{(i)}$$

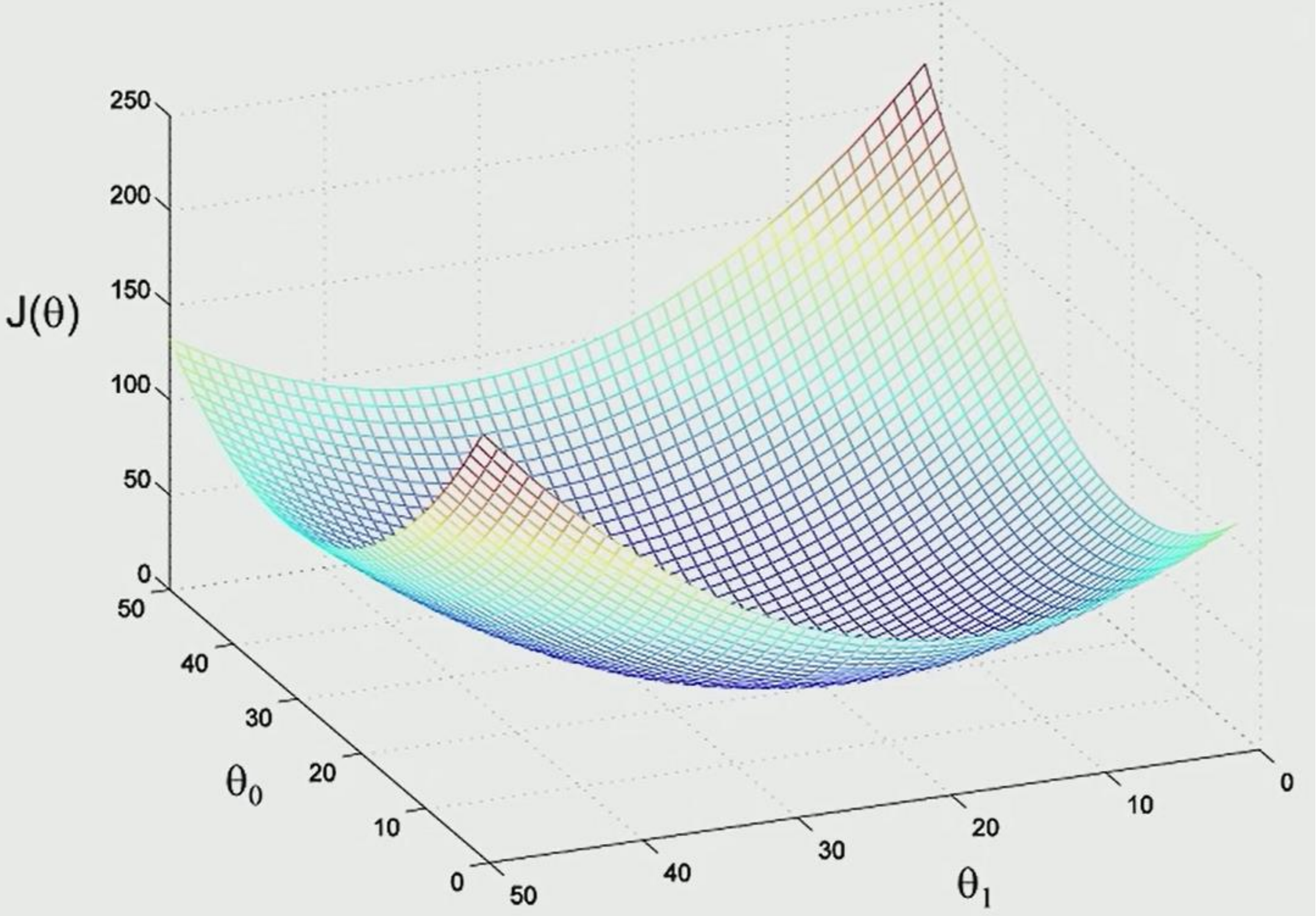
$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

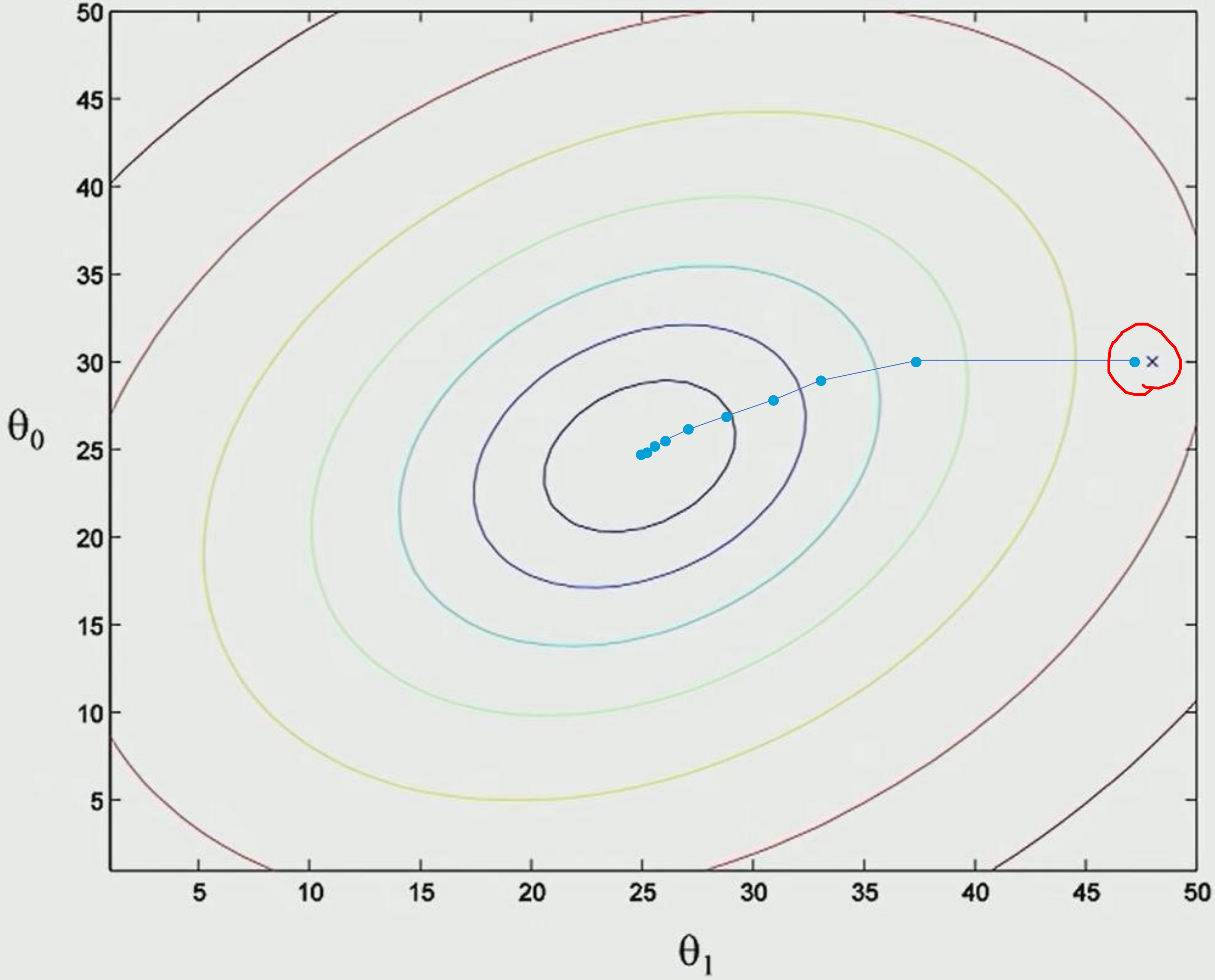
Vrste gradijentnog algoritma

- Paketni ili „šaržni“ (*Batch Gradient Descent*)
  - Svi podaci se uzimaju odjednom u obzir
  - Nedostatak
    - Za veliku količinu podataka (npr. 10^6) svi se moraju učitati iz baze i sporije dolazi do rešenja
- Stohastički (*Stochastic Gradient Descent*)
  - Izvršava se u svakom koraku za npr. jedan slučajan podatak
    - Npr. Za cenu kuće, 1. korak za kuću 1, 2. korak za kuću 2, ...
  - Za veliko m je brži od Paketnog gradijentnog algoritma

Gradijentni postupak

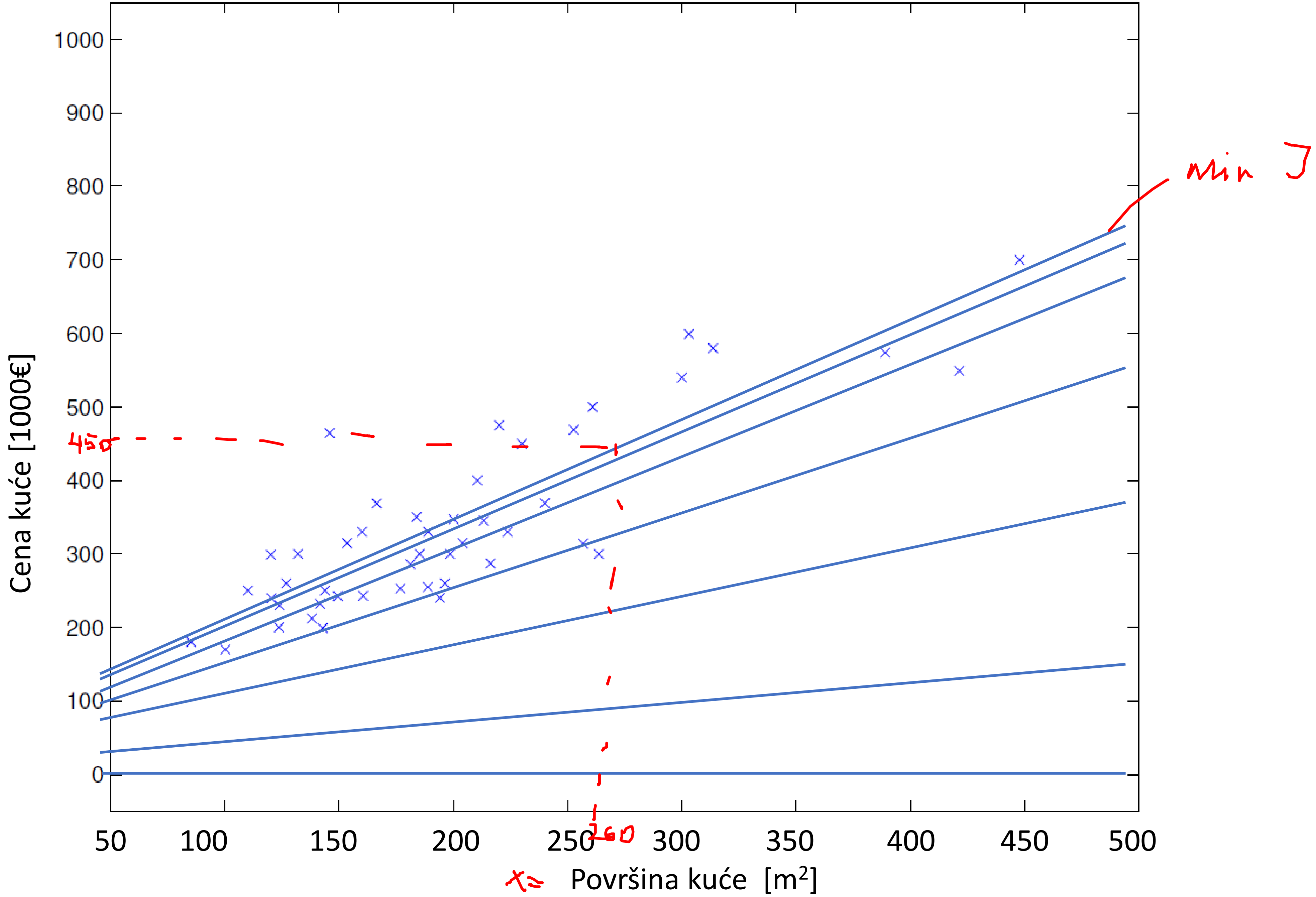



J(θ)
250
200
150
100
50
0
θ0
50
40
30
20
10
0
θ1
50
40
30
20
10
0

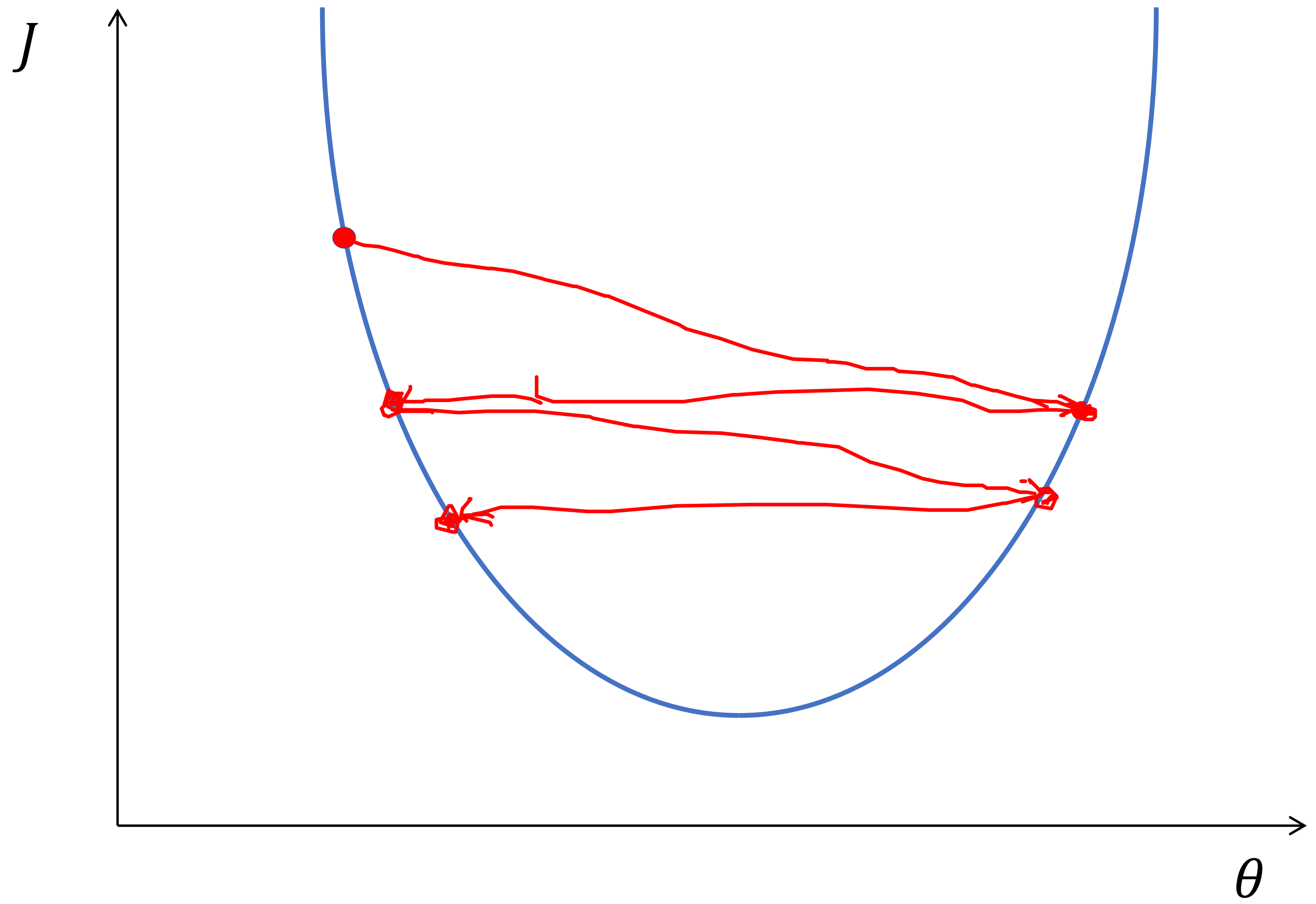


50
45
40
35
30
25
20
15
10
5
θ_0
5 10 15 20 25 30 35 40 45 50
θ_1

Prikaz estimacije po iteracijama

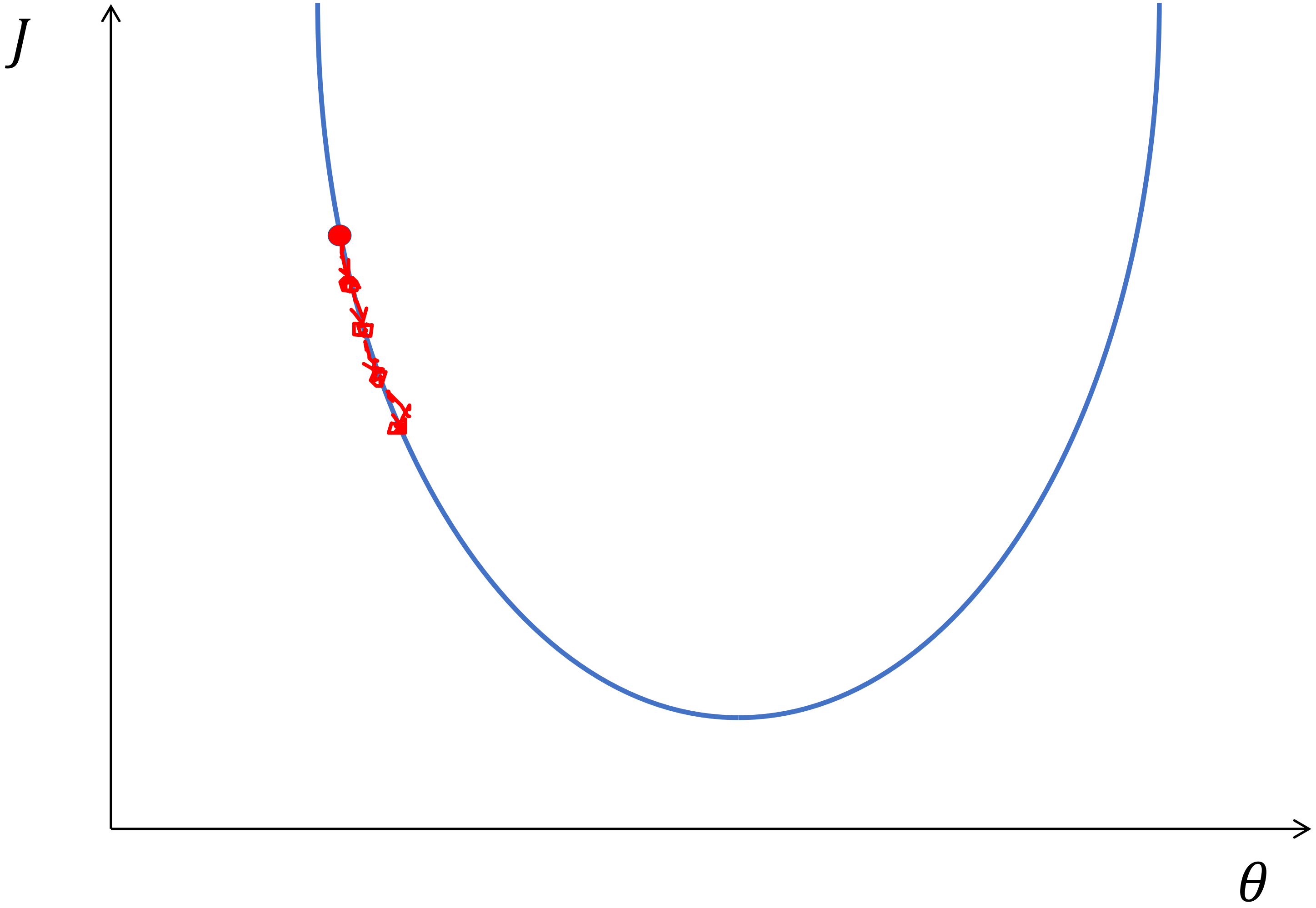


Kako izabrati faktor učenja α?



α

Kako izabrati faktor učenja α?



α mala

α 0.01
0.1

Linearna regresija za više promenljivih

- Obučavajući skup (X, Y) se sastoji iz više primera $(x^{(i)}, y^{(i)})$, $i \in \{1,2,\ldots,m\}$, $x^{(1)}$ tako npr.

Površina (m^2)	Broj spavaćih soba	Spratnost	Starost kuće (god)	Cena (€1000)
250	5	1	15	460
145	3	1	20	232
162	4	2	10	315
93	2	1	32	178

$X^{(1)}$

$x_0^{(1)} = 1$

$Y^{(1)}$

Linearna regresija za više promenljivih

- Obučavajući skup (X, Y) se sastoji iz više primera $(x^{(i)}, y^{(i)})$, $i \in \{1,2,\ldots,m\}$, $x^{(1)}$ tako npr.

X	x_0	x_1	x_2	x_3	x_4	y
$x^{(1)}$	1	250	5	1	15	460
$x^{(2)}$	1	145	3	1	20	232
$x^{(3)}$	1	162	4	2	10	315
$x^{(4)}$	1	93	2	1	32	178

$x_3^{(1)}$

$h_\theta(x) = \theta_0 x_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \\ x^{(4)} \end{bmatrix} = \begin{bmatrix} 1 & 250 & 5 & 1 & 15 \\ 1 & 145 & 3 & 1 & 20 \\ 1 & 162 & 4 & 2 & 10 \\ 1 & 93 & 2 & 1 & 32 \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \end{bmatrix} \quad y = \begin{bmatrix} 460 \\ 232 \\ 315 \\ 178 \end{bmatrix}$$

$\hat{y} = X\,\theta$

Normalna jednačina

Omogućuje analitičko određivanje parametara θ

$$X = \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(m)} \end{bmatrix} = \begin{bmatrix} x_0^{(1)} & x_1^{(1)} & \cdots & x_n^{(1)} \\ x_0^{(2)} & x_1^{(2)} & & x_n^{(2)} \\ x_0^{(3)} & x_1^{(3)} & \cdots & x_n^{(3)} \\ \vdots & \vdots & \vdots & \vdots \\ x_0^{(m)} & x_1^{(m)} & \cdots & x_n^{(m)} \end{bmatrix} \quad \theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix} \quad y = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathrm{y} = X\,\theta$$

Normalna jednačina

- Polazi od optimizacionog kriterijuma $J(\theta) = \frac{1}{2}(\underbrace{X\theta}_{\hat{y}} - y)^T(X\theta - y)$

- Predstavlja drugačiji način minimizacije $J(\theta)$

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

- Dobije se da je

$$\theta = (X^T X)^{-1} X^T y$$

Kako izabrati algoritam za minimizaciju greške?

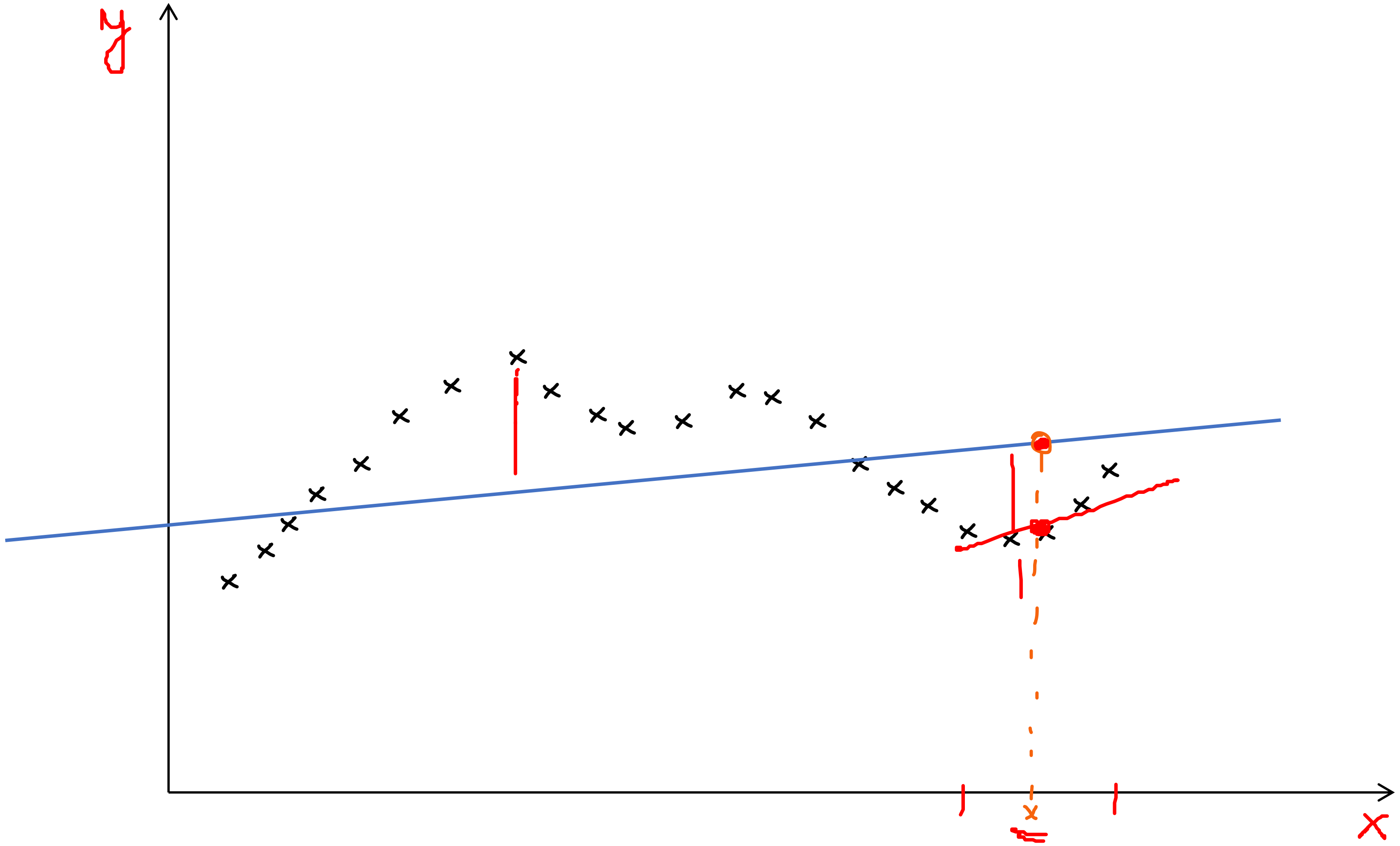
Gradijenti postupak

- Prednosti
  - Dobro radi i za veliko *n*
- Nedostaci
  - Izbor α
  - Iterativni postupak (često je potreban veliki broj iteracija)

Normalna jednačina

- Prednosti
  - Ne treba birati α
  - Rešenje se dobija direktnim analitičkim proračunom
- Nedostaci
  - Potrebno odrediti inverznu matricu $X^T X$, problem za veliko *n* (složenost $O(n^3)$)
  - Spor - kada je n jako veliko
  - Matrica nema inverznu $X^T X$

Lokalno ponderisana linearna regresija



Lokalno ponderisana linearna regresija

- Fituje se θ za minimizaciju

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

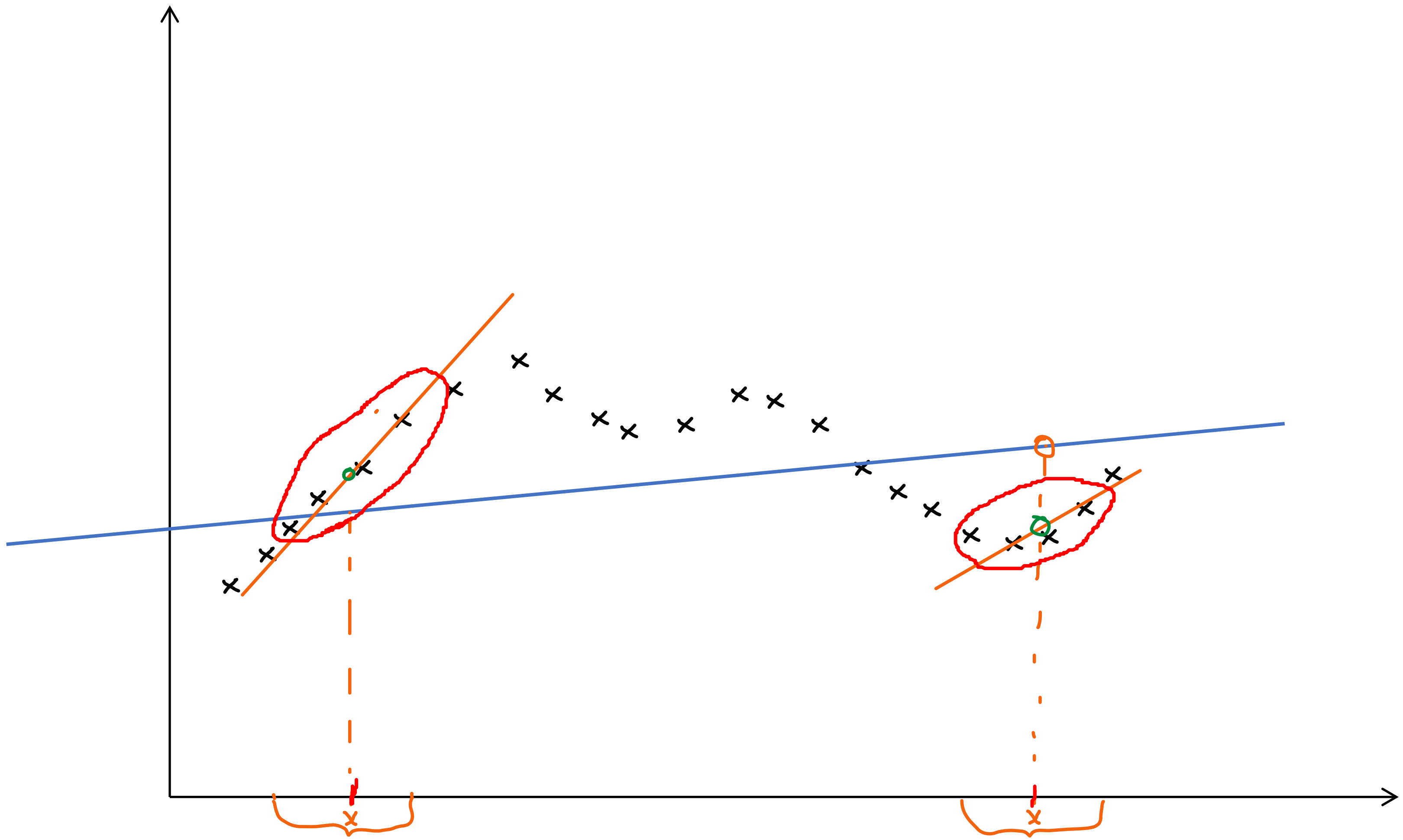
gde je $w^{(i)}$ težinska funkcija

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2}\right)$$

Ako je $\left|x^{(i)} - x\right|$ veliko, $w^{(i)} \approx 0$ – mali uticaj na J

$\left|x^{(i)} - x\right|$ malo, $w^{(i)} \approx 1$

Lokalno ponderisana linearna regresija



Lokalno ponderisana linearna regresija

- Za izbor Gausove krive za w uvodu se parametar τ

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

Za veće τ, w je manje – spušta se i podiže Gausova kriva