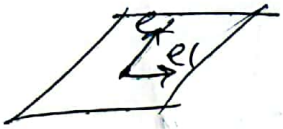
$f(x)$ $\quad f: \mathbb{R}^n \to \mathbb{R}$

$D[u]f : \mathbb{R}^n \to \mathbb{R}$

$D[v]\,D[u]f : \mathbb{R}^n \to \mathbb{R}$

$$D^2[v,u]f\Big|_{x_0} = D[v]\,D[u]f\Big|_{x_0} \quad \text{bilinear}$$



$$= v^T H u$$

H → Hessian Matrix

$$H_{ij} = e_i^T H e_j = D[e_i]\,D[e_j]f = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f$$

$H \to H(x_0)$

$$= \frac{\partial^2}{\partial x_i \partial x_j} f$$

$$D[v]\,D[u]f\Big|_{x_0} = D[v]\,u^T \nabla f\Big|_{x_0} = D[v] \sum_{i=1}^{n} u_i \frac{\partial f}{\partial x_i}$$

$$= \sum_{i=1}^{n} u_i\, D[v] \frac{\partial f}{\partial x_i} = \sum_i u_i\, v^T \nabla \frac{\partial f}{\partial x_i}$$

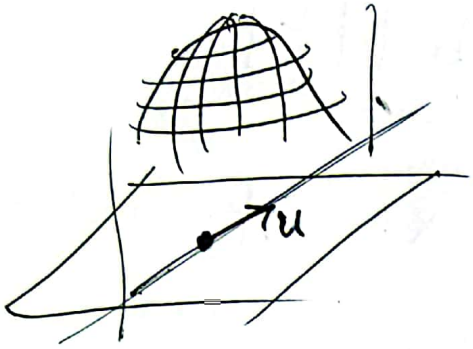
$$= \sum_i u_i \cdot v^T \begin{bmatrix} \frac{\partial f}{\partial x_1}\frac{\partial f}{\partial x_i} \\ \frac{\partial f}{\partial x_2}\frac{\partial f}{\partial x_i} \\ \frac{\partial f}{\partial x_n}\frac{\partial f}{\partial x_i} \end{bmatrix}$$

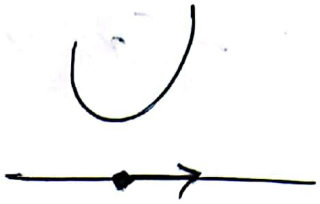
$$= v^T H u$$

$f \in C^2$ twice continuously differentiable

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i} \Rightarrow H = H^T$$

$$f(x,y) = \begin{cases} x^2 y & y<0 \\ -x^2 y & y>0 \end{cases}$$

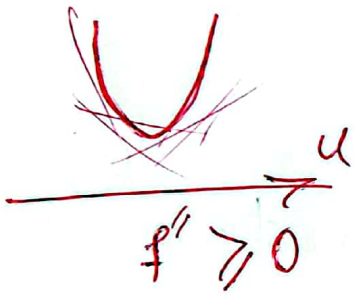


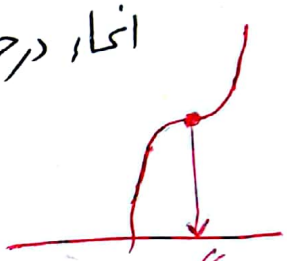


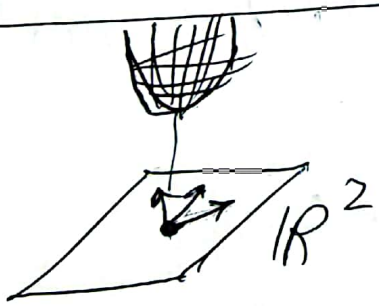
$$D^2[u,u]\Big|_{x_0} = D[u]D[u]f(x_0)$$

$$= u^T H u$$

اتجاه دلخواه u







$f \in C^2$

$D[u]D[u]f > 0$ for all $u \neq 0$

$u^T H u > 0$

$\Rightarrow H$ is positive definite

$\Rightarrow H$ is Negative definite

Quadratic Form

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = f(x_1, x_2) = a x_1^2 + b x_2^2 + c x_1 x_2$$

$$f(x) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \underbrace{\begin{bmatrix} a & c/2 \\ c/2 & b \end{bmatrix}} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= x^T A x \qquad A \text{ symmetric}$$

Ex $f: \mathbb{R}^n \longrightarrow \mathbb{R}$. $f(x) = x^T A x$ quadratic form

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

Quadratic function

$$f(x) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = a x_1^2 + b x_2^2 + c x_1 x_2 + d x_1 + e x_2 + f$$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & c/2 \\ c/2 & b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} d & e \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + f$$

$f: \mathbb{R}^n \longrightarrow \mathbb{R}$ $\quad f(x) = x^T A x + \vec{b}^T x + c$

A: $n \times n$ symmetric; $\vec{b} \in \mathbb{R}^n$; $c \in \mathbb{R}$

$$D[u] f = \frac{d}{d\alpha} (x+\alpha u)^T A (x+\alpha u) + \vec{b}^T (x+\alpha u) + c \Big|_{\alpha=0}$$

$$= u^T A x + x^T A u + b^T u$$

$$(A^T = A) \quad = 2 u^T A x + b^T u$$

$$D[v] D[u] f = D[v] (2 u^T A x + b^T u) = \frac{d}{d\alpha} 2 u^T A (x + \alpha v) + b^T u$$

$$= 2 u^T A v = 2 v^T A^T u = 2 v^T A u = v^T (2A) u$$

Hessian

A positive definite $\Rightarrow$  f

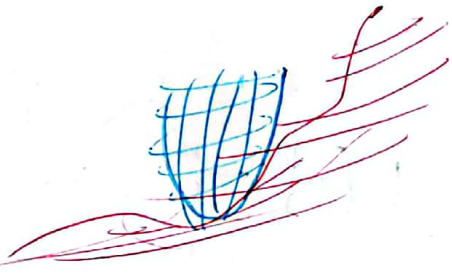
$$f(x) = x^T A x + b^T x + c$$

$$\nabla f = 2Ax + b = 0$$

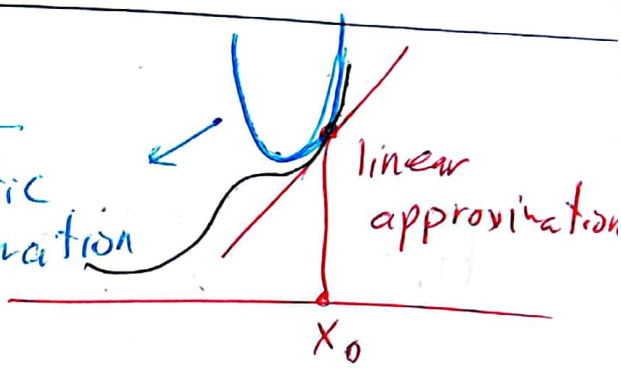
$$\Rightarrow -\tfrac{1}{2} A^{-1} b = \begin{cases} \text{minimum} & \text{if } A \text{ is PD} \\ \text{maximum} & \text{if } A \text{ is ND} \end{cases}$$

$$f(x) = \frac{1}{2} x^T A x + b^T x + c$$

min ~~$\frac{1}{2}$~~ $-\frac{1}{2} A^{-1} b$ if $A > 0$



quadratic approximation



$$\nabla f = Ax + b$$

$$H = A$$

Any function $f \in C^2$ $\quad f: \mathbb{R}^n \to \mathbb{R}$

$f(x_0)$ $\quad \nabla f(x_0)$ $\quad Hf(x_0)$

$$f(x) \simeq f(x_0) + \nabla f(x_0)^T (x - x_0) + \frac{1}{2} (x - x_0)^T H_f(x_0) (x - x_0)$$

$$f(x_0) + \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} (x_i - x_{0i}) + \frac{1}{2} \sum_i \sum_j \frac{\partial^2 f}{\partial x_j \partial x_i} (x_i - x_{0i})(x_j - x_{0j})$$

$$+ \frac{1}{3!} \sum \sum \sum (x_i - x_{0i})(x_j - x_{0j})(x_k - x_{0k}) \frac{\partial^3 f}{\partial x_k \partial x_j \partial x_i}$$

$$+ \frac{1}{4!}$$

$$D[v]\,D[u]\,f(x) = v^T \underbrace{H(x)}_{\text{Hessian}}\, u$$

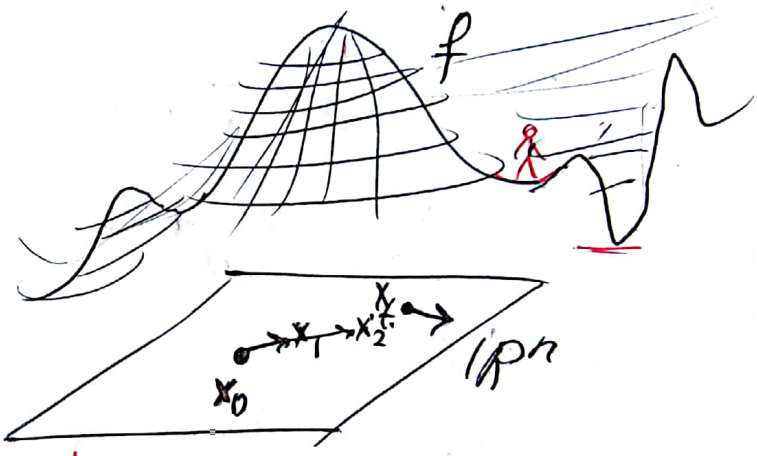
---

minimize $f: \mathbb{R}^n \longrightarrow \mathbb{R}$

$\nabla f(x) = 0 \qquad x = ?$

n equations
n unknowns

$f: \mathbb{R}^n \longrightarrow \mathbb{R}$



$x_0 =$ random

while ( )

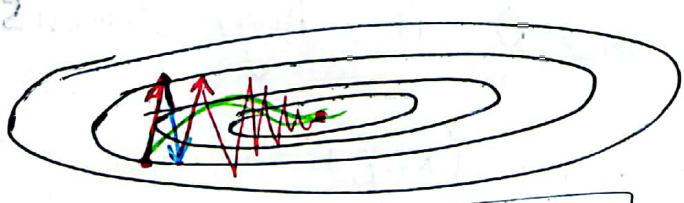
$x_{t+1} \leftarrow x_t - \alpha \nabla f(x_t)$

$t = t+1$

α → learning rate
step size

while $(\|x_t - x_{t+1}\| > \varepsilon)$

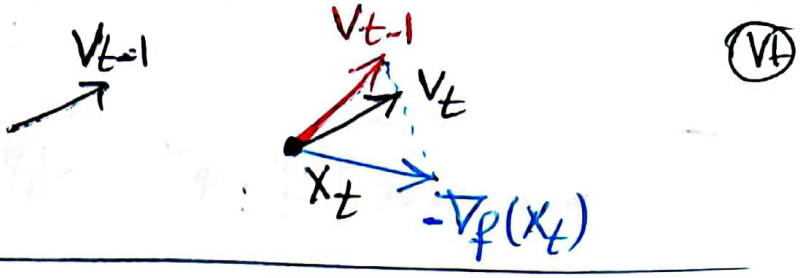
while $(\|\nabla x_t\| > \varepsilon)$



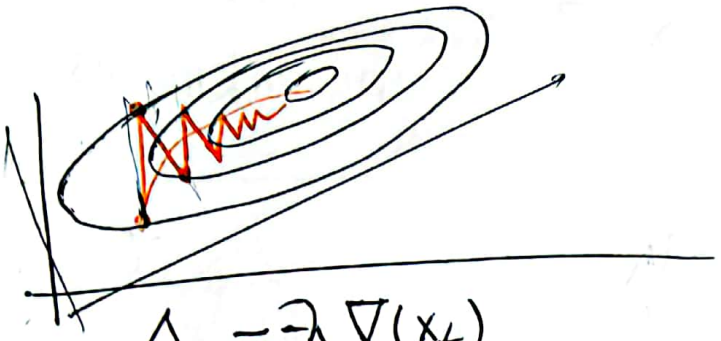
$x_{t+1} \leftarrow x_t + \underbrace{v_t - \nabla f(x_t)}_{\text{momentum}}$

$V_t = \beta V_t + (1-\beta)(-\nabla f(x_t))$ $\beta = 0.9$ MA21

$x_{t+1} = x_t + \alpha V_t$



MA22 (I)



$\beta = 0.9$

$\Delta_t = -\lambda \nabla(x_t)$ $\Delta_t = \beta \Delta_{t-1} - \lambda \nabla(x_t)$

$x_{t+1} = x_t + \Delta_t$ $x_{t+1} = x_t + \Delta_t$

---

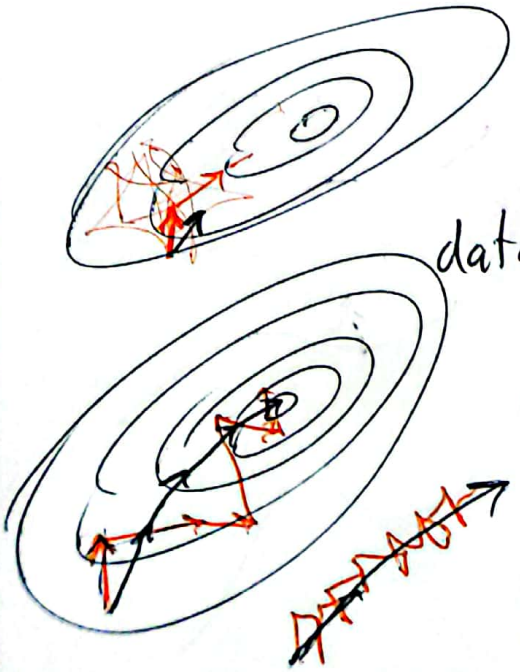
cost function:

$$C(\theta) = \sum_{i=1}^{N} C_i(\theta) = \sum_{i=1}^{N} d(f_\theta(x_i), y_i) = \sum_{i=1}^{N} \| f_\theta(x_i) - y_i \|_2^2$$

$$\frac{\partial C(\theta)}{\partial \theta} = \nabla C(\theta) = \sum_{i=1}^{N} \nabla C_i(\theta)$$

In modern ML, N is typically very large

also θ is high dimensional

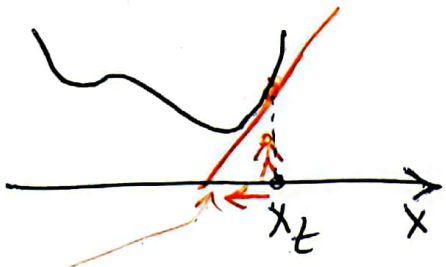


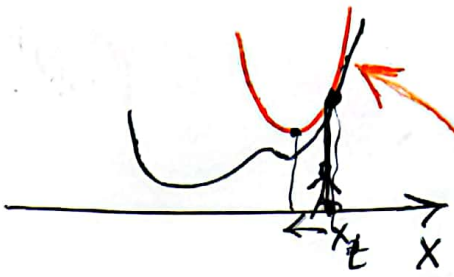
data: $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$

mini-batch 1, mini-batch 2

SGD stochastic Gradient Descent

SGD+momentum / RMS Prop / ~~AMAP~~ ADAM





تقریب درجه ۲

quadratic approximation

$f(x)$ $\quad f: \mathbb{R} \to \mathbb{R}$ $\quad x \in \mathbb{R}$

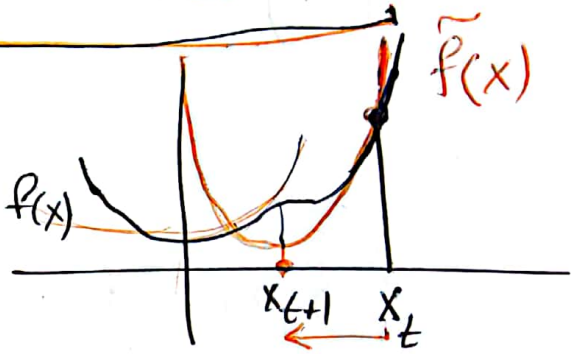
$$f(x) \simeq f(x_t) + f'(x_t)(x - x_t) + \frac{1}{2} f''(x_t)(x - x_t)^2$$

$\tilde{f}(x)$

$$\tilde{f}(x_t) = f(x)$$

$$\tilde{f}'(x_t) = f'(x_t)$$

$$\tilde{f}''(x_t) = f''(x_t)$$



$$\tilde{f}(x) \simeq \underbrace{f(x_t)}_{c} + \underbrace{f'(x_t)}_{b}(x - x_t) + \frac{1}{2}\underbrace{f''(x_t)}_{a}(x - x_t)^2$$

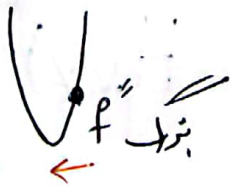
$$\tilde{f}(x) = c + b(x - x_t) + \frac{1}{2} a (x - x_t)^2$$

$$\frac{d}{dx}\tilde{f}(x) = b + a(x - x_t) = 0 \Rightarrow x - x_t = -\frac{b}{a}$$

$$\Rightarrow x = x_t - \frac{b}{a} \Rightarrow \boxed{x_{t+1} \leftarrow x_t - \frac{f'(x_t)}{f''(x_t)}}$$

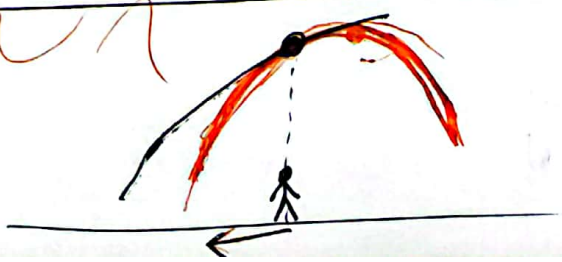
update equation

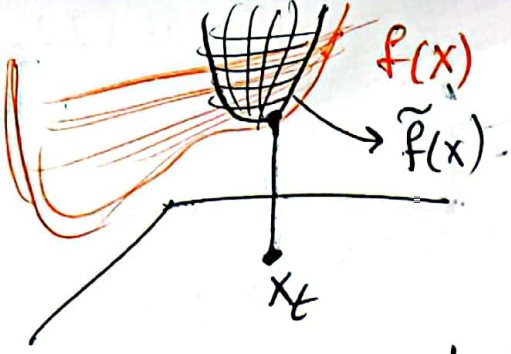
روش نیوتون
Newton's Method



$f''(x_t) > 0$ به سمت کمینه

$f''(x_t) < 0$





$f: \mathbb{R}^n \longrightarrow \mathbb{R}$

Quadratic Approximation

$$f(x) \simeq \underbrace{f(x_t) + \nabla_f(x_t)^T (x-x_t) + \frac{1}{2}(x-x_t)^T H_f(x_t)(x-x_t)}_{\tilde{f}(x)}$$

گرادیان — هسین

$$\tilde{f}(x) = f(x_t) + \nabla(x_t)^T (x-x_t) + \frac{1}{2}(x-x_t)^T H_f(x_t)(x-x_t)$$

$$\Rightarrow \tilde{f}(x) = c + \vec{b}^T (x-x_t) + \frac{1}{2}(x-x_t)^T A (x-x_t)$$

$$\nabla\tilde{f}(x) = 0 \Longrightarrow 0 + \vec{b} + A(x-x_t) = 0$$

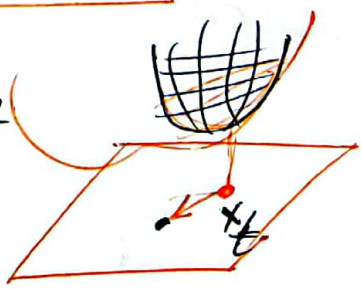
$$\Rightarrow A(x-x_t) = -\vec{b} \Rightarrow$$

$$= (x-x_t) = -A^{-1}\vec{b}$$

$$x = x_t - A^{-1}b$$

$$\boxed{x_{t+1} \leftarrow x_t - H_f(x_t)^{-1} \nabla_f(x_t)}$$

Only if $H(x_t)$ is positive-definite



Compl computing

Newton's Method

$$x_{t+1} \leftarrow x_t - H^{-1}(x_t)\,\nabla(x_t)$$

$f: \mathbb{R}^n \to \mathbb{R}$

$x \in \mathbb{R}^n$

- Computing Gradient: $O(n)$ time
- Computing Hessian: $O(n^2)$ time
- Computing H^{-1}: $O(n^3)$

Quasi-Newton روش‌های شبه نیوتون

Approximate { H using gradients

H^{-1} using gradient

Least Squares

$$\|Ax - b\|_2^2$$

$$C(x) = \|Ax-b\|^2 = \left\| \begin{bmatrix} a_1^T \\ a_2^T \\ \vdots \\ a_n^T \end{bmatrix} x - \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} \right\| = \sum_{i=1}^{n} (a_i^T x - b_i)^2$$

Non-linear Least Squares

$x \in \mathbb{R}^m$

$h_i : \mathbb{R}^m \to \mathbb{R}$

$$C(x) = \sum_{i=1}^{n} (h_i(x) - b_i)^2$$

$$C(x) = \sum_{i=1}^{n} \|h_i(x) - \vec{b_i}\|^2 = \sum_{i=1}^{n} (h_i(x) - \vec{b_i})^T (h_i(x) - \vec{b_i})$$

$h_i : \mathbb{R}^m \to \mathbb{R}^p$, $\vec{b_i} \in \mathbb{R}^p$

$$C(\theta) = \sum_{i=1}^{n} \| f_\theta(x^i) - y^i \|^2$$

$$= \sum_{i=1}^{n} \| f(\theta, x^i) - y^i \|^2$$

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

$C: \mathbb{R}^p \to \mathbb{R} \;/\; \theta \in \mathbb{R}^p \qquad f_i(\theta)$

$$C(\theta) = \sum_{i=1}^{n} \| f_i(\theta) \|^2 = \sum_{i=1}^{n} f_i(\theta)^T f_i(\theta)$$

Non-linear Least squares

$\theta_t \qquad \theta_{t+1} \leftarrow \theta_t + \Delta\theta$

~~$f(\theta)$~~ $f: \mathbb{R}^p \to \mathbb{R}^m$

$$f(\theta_t + \Delta\theta) \simeq f(\theta_t) + J\Big|_{\theta_t} \Delta\theta$$

$$f(\theta) \simeq f(\theta_t) + J\Big|_{\theta_t} (\theta - \theta_t)$$

$$C(\theta) \simeq \tilde{C}(\theta) = f_i(\theta_t)$$

LM

$$C(\theta + \Delta\theta) \simeq \tilde{C}(\theta + \Delta\theta) = \sum_{i=1}^{n} \left(f(\theta_t) + J\Delta\theta \right)^T \left(f(\theta_t) + J\Delta\theta \right)$$

$\tilde{C}: \mathbb{R}^p \to \mathbb{R}$

$$\nabla \tilde{C} = 2 \sum_{i=1}^{n} J_i^T \left(f_i(\theta_t) + J_i \Delta\theta \right) = 0$$

$$= 2 \left(\sum_{i=1}^{n} \left(J_i^T J_i \right) \right) \Delta\theta = \sum_{i=1}^{n} J_i^T f_i(\theta_t)$$