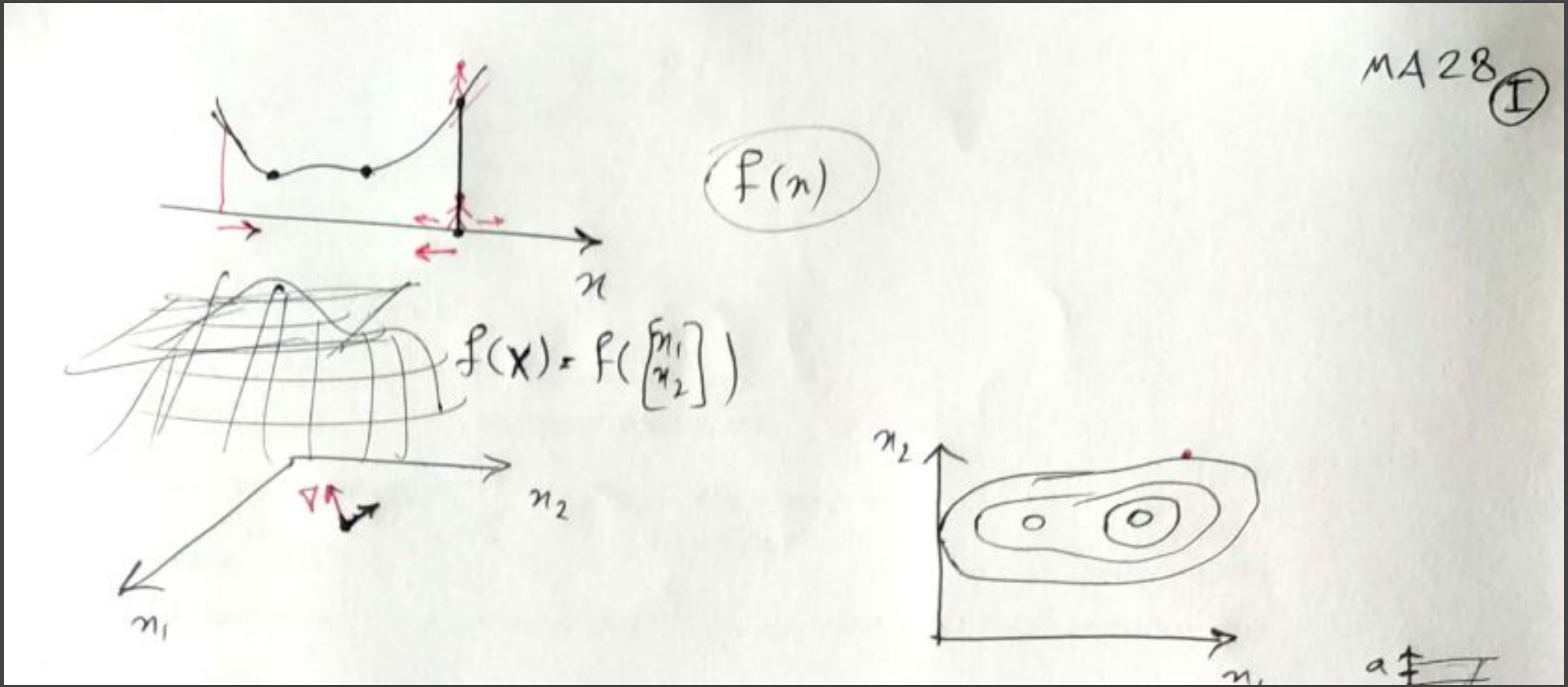
Mathematics for AI

Lecture 28

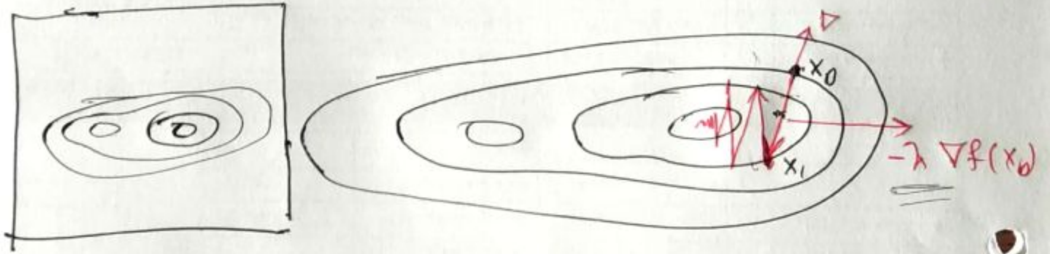
Gradient Descent, SGD, Momentum
Quadratic Approximation, Newton's method

Gradient Descent



Gradient Descent

K. N. Toosi
University of Technology



x_0

$\nabla f(x_0)$

we like to have

$f(x_1) < f(x_0)$

$x_1 \leftarrow x_0 - \lambda \nabla f(x_0)$

step size
learning rate

line search

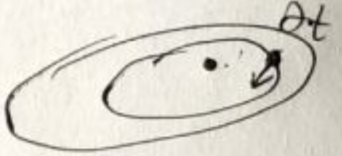
$\boxed{x_{t+1} = x_t - \lambda \nabla f(x_t)}$ gradient descent

Stochastic Gradient Descent (SGD)

K. N. Toosi
University of Technology

MA28 (II)

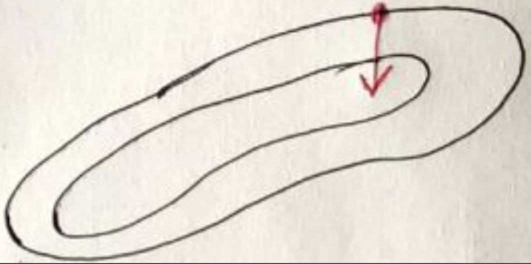
$$C(\theta) = \sum_{i=1}^{N} \| f(\theta, x_i) - y_i \|^2$$



$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

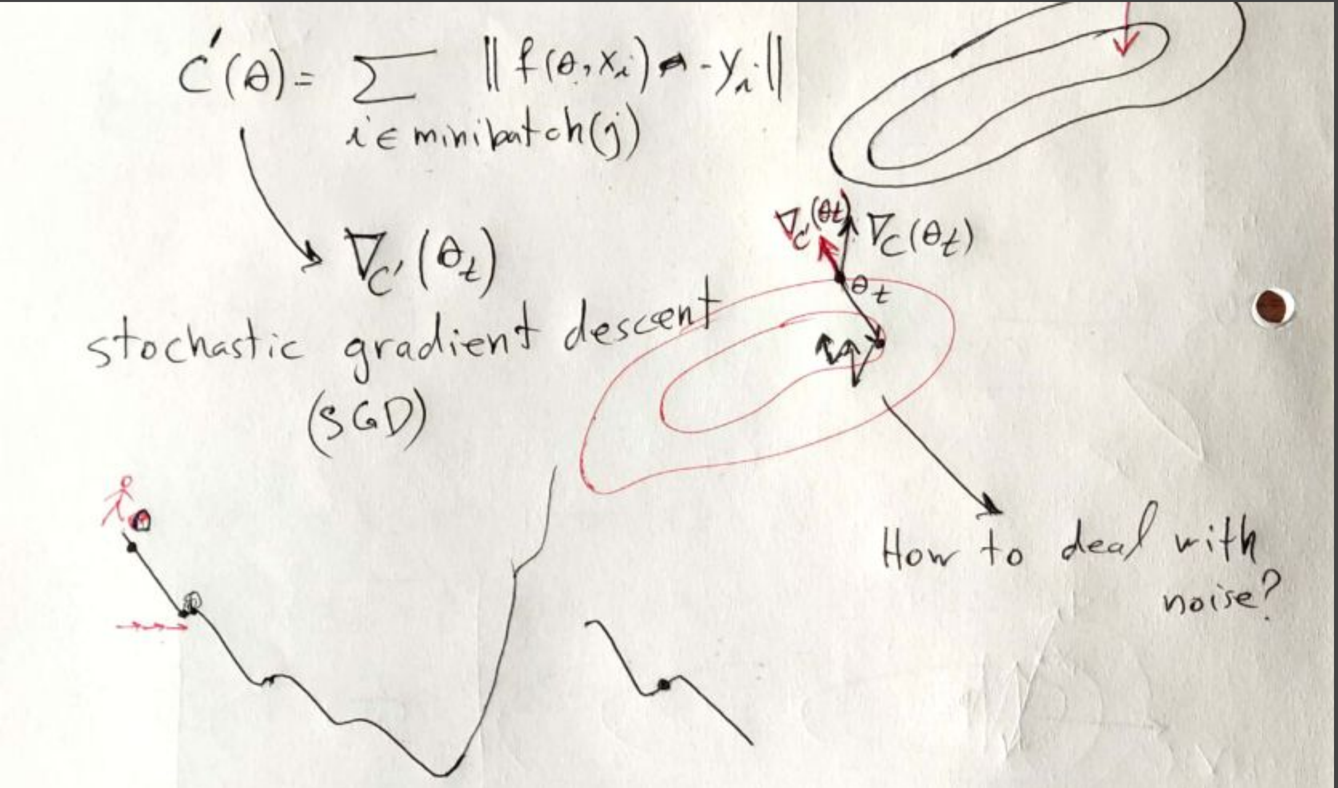
mini-batch

epoch



$$C'(\theta) = \sum_{i \in \text{minibatch}(j)} \| f(\theta, x_i) - y_i \|$$

Stochastic Gradient Descent (SGD)



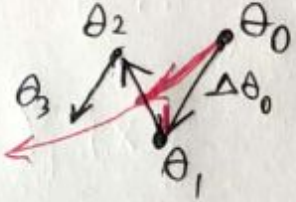
Momentum

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

$$\boxed{\theta_{t+1} \leftarrow \theta_t + \Delta\theta_t}$$

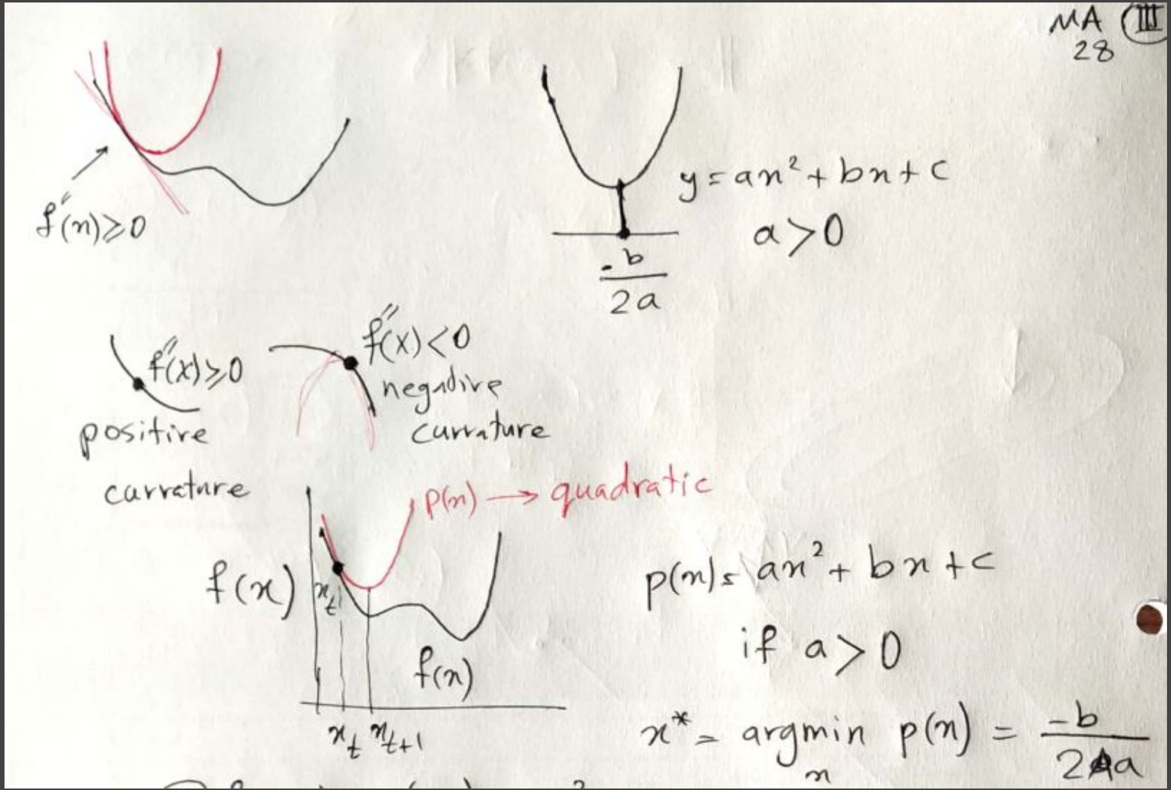
Gradient descent: $\Delta\theta_t = -\lambda \nabla_c(\theta_t)$

Momentum Method: $\Delta\theta_t = -\lambda \nabla_c(\theta_t) + \alpha \Delta\theta_{t-1}$



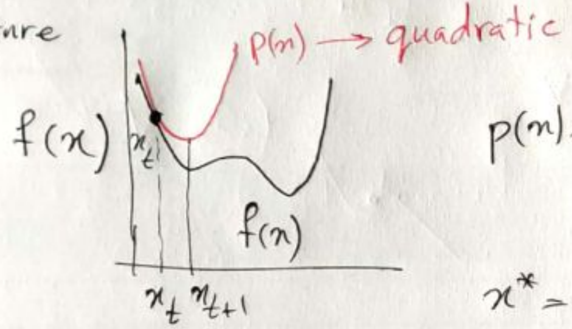
Quadratic Approximation

K. N. Toosi
University of Technology



Quadratic Approximation

curvature



$$p(x) = ax^2 + bx + c$$

if $a > 0$

$$x^* = \underset{x}{\operatorname{argmin}}\ p(x) = \frac{-b}{2a}$$

I $f(x_t) = p(x_t) = a x_t^2 + b x_t + c$

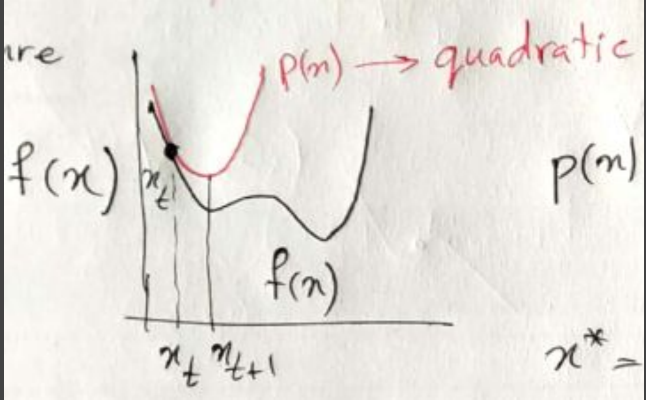
II $f'(x_t) = p'(x_t) = 2a x_t + b$ $\cancel{\Rightarrow b = 2a x_t\, f'(x_t) - 2a x_t}$

III $f''(x_t) = p''(x_t) = 2a \Longrightarrow a = \dfrac{f''(x_t)}{2}$

II $b = f'(x_t) - 2a x_t = f'(x_t) - f''(x_t)\, x_t$

$c = f(x_t) - a x_t^2 - b x_t$

Newton's method



$$\text{III}\quad f''(x_t) = p''(x_t) = 2a \implies a = \frac{f''(x_t)}{2}$$

$$\text{II}\quad b = f'(x_t) - 2a x_t = f'(x_t) - f''(x_t) x_t$$

$$c = f(x_t) - a x_t^2 - b x_t$$ مهم نیست

$$x_{t+1} = \frac{-b}{2a} = \frac{-f'(x_t) + f''(x_t)\, x_t}{f''(x_t)}$$

$$x_{t+1} = x_t - \frac{f'(x_t)}{f''(x_t)}$$

Newton's method

روش نیوتون

$f''(x_t)$ ~~must~~ need to be positive

Multivariate Quadratic Function

K. N. Toosi
University of Technology

$$f(x) = f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = 2x_1^2 + x_1 x_2 + 3x_2^2 + 2x_1 - 4x_2 - 10$$

MA (V

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 10$$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 10$$

$$= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & 1.5 \\ 1.5 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 2 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - 10$$

symmetric

unique

$$= x^T A x + \vec{b}^T x + c$$

Multivariate Quadratic Function

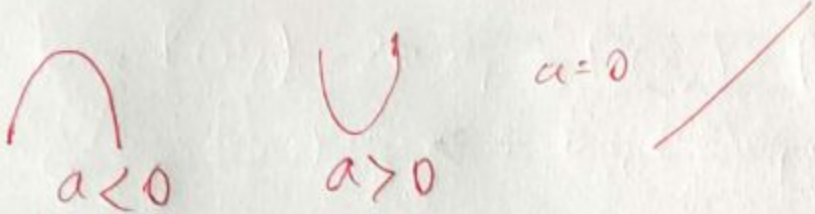
Every quadratic function $p: \mathbb{R}^2 \longrightarrow \mathbb{R}$

can be uniquely represented as

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

where $A \in \mathbb{R}^{n \times n}$ is <u>symmetric</u>, $b \in \mathbb{R}^n$, $c \in \mathbb{R}$

$n = 1$



$p'(x) = 0 -$

$x = -b/2a$

Stationary point of a Quadratic Function

K. N. Toosi
University of Technology

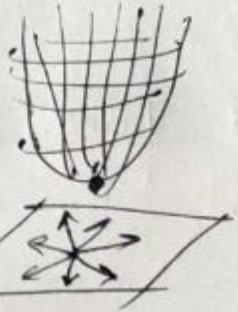
$$p(x) = x^T A x + b^T x + c$$

$$\nabla p(x) = 2Ax + b \qquad \nabla p(x) = 0 \Rightarrow 2Ax = -b$$

$$H(x) = \underline{\underline{2A}} \qquad x^* = -\frac{A^{-1}b}{2}$$

if A nonsingular

x^*



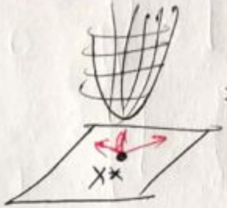
x

When is the stationary point the minimum?

K. N. Toosi
University of Technology

$p(x) = x^T A x + b^T x + c$ MA 29

$\nabla = 0 \qquad x^* = -\frac{1}{2} A^{-1} b$

x^* is a minimum 

$\Longrightarrow$

directional curvature is positive in all directions.

$\underline{u^T H u > 0}$ for all $u \neq 0$

$\Longrightarrow$ H is positive definite

When is the stationary point the minimum?

K. N. Toosi
University of Technology

$p(x) = x^T A x + b^T x + c \Rightarrow$ 2A is positive definite

A is positiv definite

When do we have a maximum?

 $\Rightarrow$ A is negative-definite

Quadratic Approximation - Multivariate case

$$f(x_t) = P(x_t)$$

$$\nabla f(x_t) = \nabla p = 2Ax_t + b$$

$$H_f(x_t) = \cancel{\nabla} H_p(x_t) = 2A \Rightarrow 2A = H_f(x)$$

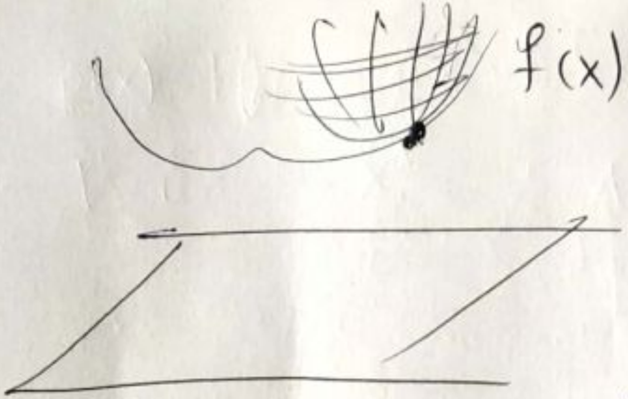
$$b = \nabla f(x_t) - 2Ax_t = \nabla_f(x_t) - H_f(x_t)\, x_t$$

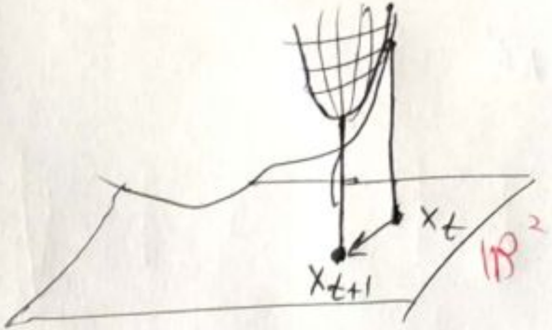
$H_f(x_t)$ should be positive definite

$$x_{t+1} = -\frac{1}{2} A^{-1} b = -(2A)^{-1} b = -H^{-1}(\nabla - H)x_t = x_t - H^{-1}\nabla$$

Newton's method - Multivariate case

K. N. Toosi
University of Technology





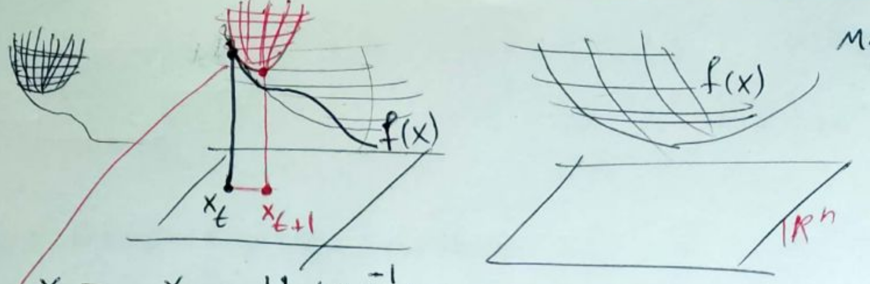
$$x_{t+1} = x_t - \underbrace{H_f^{-1}(x_t)}_{\mathbb{R}^{n\times n}} \underbrace{\nabla_f(x_t)}_{\in \mathbb{R}^n} \qquad x \in \mathbb{R}^n$$

H_f positive-definite

Quadratic Approximation - Taylor series perspective

K. N. Toosi
University of Technology

MA 29(



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

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

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

$$\nabla + H(x - x_t) = 0$$

$$\Rightarrow x - x_t = -H^{-1}\nabla$$

$$x^* = x_t - H^{-1}\nabla$$

$O(n^3)$