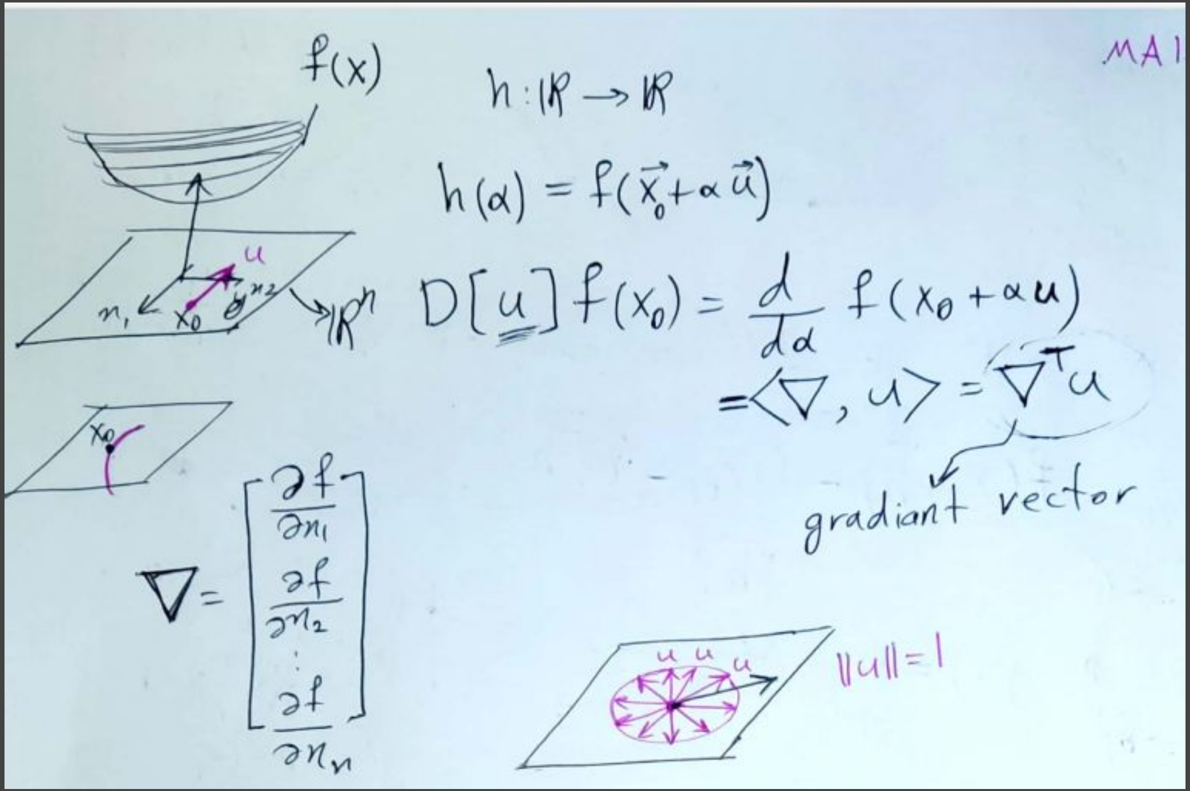
# Mathematics for AI

## Lecture 16

# The gradient vector

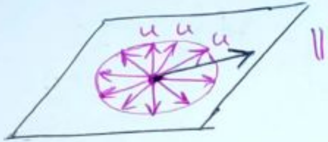
K. N. Toosi
University of Technology



# The gradient and steepest directions

gradiant vector

$$\nabla = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$




$\|u\|=1$

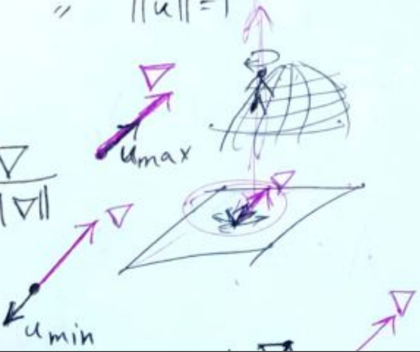
$$u_{max} = \arg\max_{u} D[u]f(x_0) \quad \text{subject to } \|u\|=1$$

$$u_{min} = \arg\min_{u} D[u]f(x_0) \quad \text{〃} \quad \text{〃} \quad \|u\|=1$$

$f$: diffrentar

$$u_{max} = \arg\max_{u} \langle u, \nabla \rangle = \frac{\nabla}{\|\nabla\|}$$

$$u_{min} = \arg\min_{u} \langle u, \nabla \rangle = -\frac{\nabla}{\|\nabla\|}$$



# Moving perpendicular to gradient



$$D[u]\,f(x_0) = 0 \Rightarrow \langle u, \nabla \rangle = 0 \Rightarrow u \perp \nabla$$

# Example

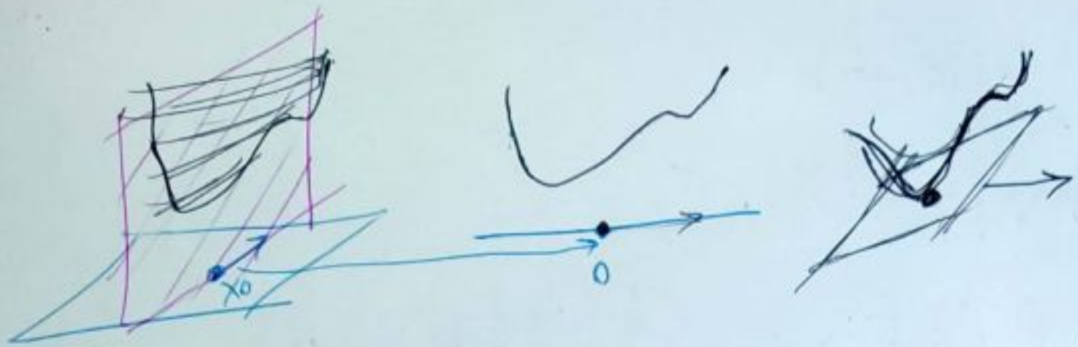
$$f(x) = f\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = x_1 x_2 + x_3 \sin x_2 + x_1 x_2 e^{x_3}$$

$$f : \mathbb{R}^3 \longrightarrow \mathbb{R}$$

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix} = \begin{bmatrix} x_2 + x_2 e^{x_3} \\ x_1 + x_3 \cos x_2 + x_1 e^{x_3} \\ \sin x_2 + x_1 x_2 e^{x_3} \end{bmatrix} \in \mathbb{R}^3$$

# Definition of differentiability

LA 16 (II)




$\to f(x_0)+\vec{m}^T(x-x_0)$

$f: \mathbb{R}^n \longrightarrow \mathbb{R}$ is differentiable at $x_0 \Rightarrow \exists\, \vec{m} \in \mathbb{R}^n$

$$\lim_{\vec{h}\to\vec{0}} \frac{f(\vec{x}_0+\vec{h}) - f(\vec{x}_0) - \vec{m}^T h}{\|\vec{h}\|} = 0$$

# Limits in higher dimensions

K. N. Toosi
University of Technology

~~lim g~~ $g: \mathbb{R}^m \longrightarrow \mathbb{R}^n$

$\lim_{\vec{x} \to \vec{x}_0} g(\vec{x}) = \vec{y}_0$ $\quad \forall \varepsilon \ \exists \delta \quad \|x - x_0\| < \delta \Rightarrow \|g(x) - \vec{y}_0\| < \varepsilon$

# Example



$x \to x_0$

$\cancel{A \in \mathbb{R}^{m \times n}}$ $\quad x \in \mathbb{R}^n \quad \mathrm{Diag}(x) = \begin{bmatrix} x_1 & & & \\ & x_2 & & \\ & & \ddots & \\ & & & x_n \end{bmatrix}$

$$\mathrm{Diag}(x) = \mathrm{Diag}\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}\right) = \begin{bmatrix} x_1 & 0 & 0 & 0 \\ 0 & x_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & x_n \end{bmatrix}$$

$\cancel{f(x)}$ $\quad f : \mathbb{R}^n \to \mathbb{R} \qquad f(x) = u^T \mathrm{Diag}(x)\, v$ for

constant vectors $u, v \in \mathbb{R}^n$

$$u^T \mathrm{Diag}(x)\, v = [u_1 \; u_2 \cdots u_n] \begin{bmatrix} x_1 & & \emptyset \\ & x_2 & \\ \emptyset & & \ddots \; x_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \sum_{i=1}^{n} x_i \cdot u_i \cdot v_i$$

$$\frac{\partial f}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} x_i u_i v_i = u_k v_k$$

$$\nabla = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} u_1 v_1 \\ u_2 v_2 \\ \vdots \\ u_n v_n \end{bmatrix} = \mathrm{Diag}(u)\, v = \mathrm{Diag}(v)\, u = u \odot v$$

# Hadamard Product

Hadamard product (element-wise product) LA16 (III)

$$u \odot v = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \odot \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} u_1 v_1 \\ u_2 v_2 \\ \vdots \\ u_n v_n \end{bmatrix} = \underbrace{\mathrm{Diag}(u)}_{n \times n} v = \mathrm{Diag}(v)\, u$$

K. N. Toosi University of Technology

# Example

$f: \mathbb{R}^n \to \mathbb{R}$ $\quad f(x) = x^T A x$ for a constant matrix $A \in \mathbb{R}^{n\times n}$.

$$f(x) = x^T A x = \underbrace{\begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}}_{1\times n} \underbrace{\overbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}}^{A}}_{n\times n} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{n\times 1} \in \mathbb{R}$$

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

$$\frac{\partial f}{\partial x_k}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j\right)$$

$$\frac{\partial}{\partial x_k}\left(a_{kk} \underbrace{x_k x_k}_{x_k^2} + \sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj} x_k x_j + \sum_{\substack{i=1\\ i\neq k}}^{n} a_{ik} x_i x_k + \overbrace{\cancel{\sum_{\substack{i=1\\ i\neq k}}^{n}\sum_{\substack{j=1\\ j\neq k}}^{n} a_{ij} x_i x_j}}^{0}\right)$$

$$= 2 a_{kk} x_k + \sum_{\substack{j=1\\ j\neq k}}^{n} a_{kj} x_j + \sum_{\substack{i=1\\ i\neq k}}^{n} a_{ik} x_i$$

$$= \sum_{j=1}^{n} a_{kj} x_j + \sum_{i=1}^{n} a_{ik} x_i = \begin{bmatrix} a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}}_{x} + \begin{bmatrix} a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$$\frac{\partial f}{\partial x_k} = \underbrace{a[k,:]}_{\text{k-th row of A}}^T x + \underbrace{a[:,k]}_{\text{k-th column of A}}^T x$$

$$\nabla = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix} = \begin{bmatrix} a_{1:}^T x + a_{:1}^T x \\ a_{2:}^T x + a_{:2}^T x \\ \vdots \\ a_{n:}^T x + a_{:n}^T x \end{bmatrix} = (A + A^T) x$$

# Example

K. N. Toosi
University of Technology

$$f : \mathbb{R}^n \longrightarrow \mathbb{R} \qquad x \longmapsto x^T A x$$

$$\nabla f = (\underbrace{A + A^T}_{n \times n})\underbrace{x}_{n \times 1} \in \mathbb{R}^n$$

$$A \text{ symmetric} \Longrightarrow \nabla f = 2Ax$$

# Example: Least Squares

least squares problem $Ax = b$, $A \in \mathbb{R}^{m \times n}$

$m \geqslant n$

$$\begin{bmatrix} A \end{bmatrix}\begin{bmatrix} x \end{bmatrix} = \begin{bmatrix} b \end{bmatrix}$$

A has full column rank

$\text{rank}(A) = n$

least squares problem geometric method

$$x^* = \underset{x}{\text{argmin}} \|Ax - b\|^2 \Longrightarrow x = (A^TA)^{-1}A^Tb$$

$$f(x) = \|Ax - b\|^2 \qquad \nabla f = 0 \Longrightarrow x = \checkmark$$

$$f(x) = \left\| \begin{bmatrix} a_{1:}^T x - b_1 \\ a_{2:}^T x - b_2 \\ \vdots \\ a_{n:}^T x - b_n \end{bmatrix} \right\|^2 = \sum_{i=1}^{n} (a_{i:}^T x - b_i)^2$$

$$= \sum_{i=1}^{n} (a_{i1} x_1 + a_{i2} x_2 + \cdots + a_{in} x_n - b_i)^2$$

# Example: Least Squares



$$= \sum_{i=1} ( \ldots$$

$$\frac{\partial f}{\partial x_k} = \frac{\partial}{\partial x_k} \sum_{i=1}^{n} \left( a_{i1} x_1 + a_{i2} x_2 + \cdots + a_{in} x_n - b_i \right)^2$$

$$= \sum_{i=1}^{n} 2\, a_{ik} \left( a_{i1} x_1 + a_{i2} x_2 + \cdots + a_{in} x_n - b_i \right)$$

$$= 2 \sum_{i=1}^{n} a_{ik} \left( a_{i:}^T x - b_i \right)$$

$$\frac{\partial f}{\partial x_k} = 2 \begin{bmatrix} a_{1k} & a_{2k} & \cdots & a_{nk} \end{bmatrix} \underbrace{\begin{bmatrix} a_{1:}^T x - b_1 \\ a_{2:}^T x - b_2 \\ \vdots \\ a_{n:}^T x - b_n \end{bmatrix}}_{Ax-b} = 2\, a_{:k}^T (Ax - b)$$

$$\nabla = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \vdots \\ \partial f / \partial x_n \end{bmatrix} = 2 \begin{bmatrix} a_{:1}^T \\ a_{:2}^T \\ \vdots \\ a_{:n}^T \end{bmatrix} (Ax - b) = 2 A^T (Ax - b)$$

# Example: Least Squares

$$\nabla f = \vec{0} \Rightarrow 2A^T(Ax - b) = \vec{0}$$

$$\Rightarrow \cancel{2}A^TAx - A^Tb = 0 \Rightarrow A^TAx = A^Tb$$

$$\boxed{x = (A^TA)^{-1}A^Tb}$$

least squares solution

K. N. Toosi
University of Technology

# Easier method of calculating gradient

K. N. Toosi
University of Technology

$f(x)$ $\qquad \underbrace{D[u]f} = \langle \nabla, u \rangle = \nabla^T u$

1- derive the directional derivative of $\underline{\underline{f}}$ for an arbitrary direction $u \in \mathbb{R}^n$.

2- write the solution in form of $\langle z, u \rangle$

3- $z$ is the gradient vector.

# Inner product



$\langle Ax, y \rangle$ $\quad x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$ $\quad \underline{\langle Ax, y \rangle} = \underline{\langle x, A^T y \rangle}$

product

# Inner product for matrices

K. N. Toosi
University of Technology

$A \in \mathbb{R}$

$\langle A, B \rangle \quad A, B \in \mathbb{R}^{m \times n}.$ Matrix ~~dot~~ inner product

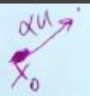
$$\langle A, B \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} B_{ij}$$

$$\langle A, B \rangle = \operatorname{trace}(A^T B) = \operatorname{trace}(A B^T)$$

$$\langle A, BC \rangle = \langle B^T A, C \rangle = \langle A C^T, B \rangle$$

# Compute gradient (easy way)

K. N. Toosi University of Technology




$$f(x) = x^T A x$$

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

$$\left[\frac{d}{d\alpha}(x+\alpha u)\right]^T A (x+\alpha u) + (x+\alpha u)^T A \left[\frac{d}{d\alpha}(x+\alpha u)\right]$$

$$u^T A (x+\alpha u) + (x+\alpha u)^T A u \Big|_{\alpha=0} \Rightarrow \underbrace{u^T A x}_{1\times 1} + \underbrace{x^T A u}_{1\times 1}$$

$$= x^T A^T u + x^T A u = (x^T A^T + x^T A) u = (Ax + A^T x)^T u =$$

$$= (Ax + A^T x)^T u \Rightarrow \langle \underbrace{Ax + A^T x}_{\nabla}, u \rangle$$

$$\Rightarrow \nabla f = Ax + A^T x = (A + A^T) x$$

# Compute Gradient (easy way) least squares



$$f(x) = \|Ax-b\|^2 = (Ax-b)^T(Ax-b)$$

$$\frac{d}{d\alpha} f(x+\alpha u) = \frac{d}{d\alpha} (A(x+\alpha u)-b)^T (A(x+\alpha u)-b) \Big|_{\alpha=0}$$

$$= (Au)^T (A(x+\alpha u)-b) + (A(x+\alpha u)-b)^T (Au) \Big|_{\alpha=0}$$

$$= (Au)^T(Ax-b) + (Ax-b)^T Au$$

$$= \underbrace{2(Ax-b)^T A}\, u = (2A^T(Ax-b))^T u$$

$$= \langle \underbrace{2A^T(Ax-b)}_{\nabla}, u \rangle$$

$$\nabla = 2A^T(Ax-b)$$