Vector Calculus II

Jo Ciucă
Australian National University
comp36706670@anu.edu.au

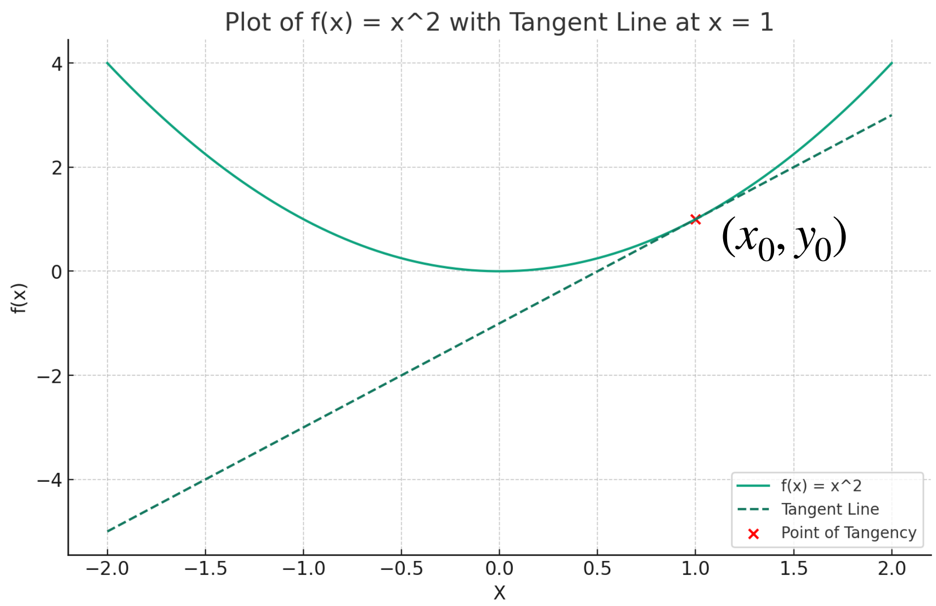
Outline

- The Gradient and the Chain Rule
- The Hessian
- Gradients of **Vector-Valued** Functions
- The Jacobian
- Exercises (inc. Gradients of Matrices, iPad session)
- Useful identities for computing gradients

The Derivative: geometrical perspective

- Consider univariate functions $f(x) = y$
- Derivative at a point x_0 is the **slope** of the tangent line at x_0
  - Negative slope: f is decreasing
  - Positive slope: f is increasing

- The **steeper** the slope (i.e. the larger the absolute value of the slope), the **larger the rate of change of f.**



The Gradient of multivariate functions

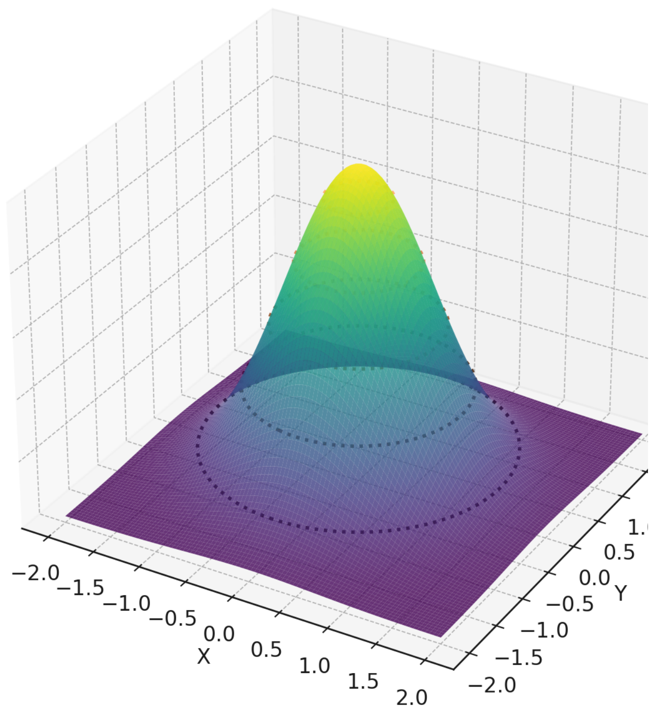
- $f: \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$
- We find the gradient of the function f with respect to $\boldsymbol{x}$ by
  - varying one variable at a time and keeping the others constant.
  - The gradient is the collection of the partial derivatives.

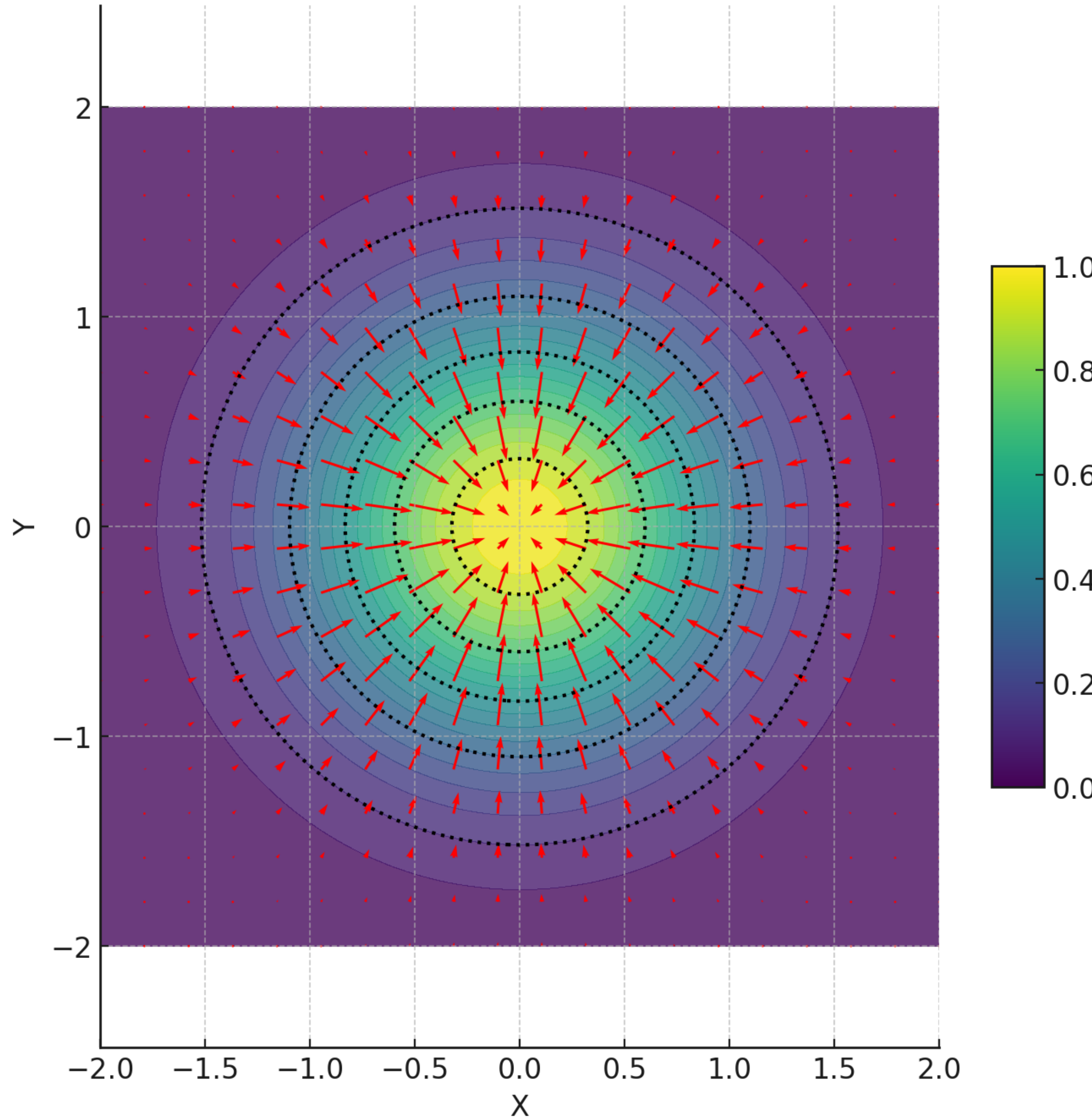
- We collect the partial derivatives in the row vector

$$\nabla_x f = \operatorname{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \boxed{\begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\boxed{\partial x_1}} & \frac{\partial f(\boldsymbol{x})}{\boxed{\partial x_2}} & \cdots & \frac{\partial f(\boldsymbol{x})}{\boxed{\partial x_n}} \end{bmatrix}} \in \mathbb{R}^{1 \times n}$$

The Gradient: geometrical perspective

$$f(x, y) = e^{-(x^2+y^2)}$$





Gradient perpendicular to
level curve (isocurve).

The Gradient: why we like it

- Encodes how our function responds to changes in the input at a specific point.
- In other words, how “sensitive” our function is to changes in input.
- **Direction**: The gradient points in the direction of the steepest ascent of the function.
- **Magnitude**: The length of the gradient represents the rate of change of the function at that point.

- It likes to be dotted.

$$[\frac{\partial f}{\partial x_1} \dots \frac{\partial f}{\partial x_n}] \begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \\ \vdots \\ \frac{dx_n(t)}{dt} \end{bmatrix}$$

Chain Rule

Chain Rule

- Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables x_1 and x_2.
- $x_1(t)$ and $x_2(t)$ are themselves functions of t.

Approximation $$\Delta f \approx f_{x_1} \Delta x_1 + f_{x_2} \Delta x_2$$

$$\frac{\Delta f}{\Delta t} \approx f_{x1} \frac{\Delta_{x_1}}{\Delta_t} + f_{x_2} \frac{\Delta_{x_2}}{\Delta_t}$$

when Δt goes to 0 : $$\frac{df}{dt} = f_{x_1} \frac{dx_1}{dt} + f_{x_2} \frac{dx_2}{dt}$$

- Using the chain rule:

$$\frac{df}{dt} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{d\boldsymbol{x}}{dt} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{dx_1(t)}{dt} \\ \frac{dx_2(t)}{dt} \end{bmatrix} = \frac{\partial f}{\partial x_1} \frac{dx_1}{dt} + \frac{\partial f}{\partial x_2} \frac{dx_2}{dt}$$

Chain Rule

- If $f(x_1, x_2)$ is a function of x_1 and x_2, where $f: \mathbb{R}^2 \to \mathbb{R}$, $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the **partial derivatives:**

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}$$

- The **gradient** can be obtained by matrix multiplication.

$$\frac{df}{d(s,t)} = \frac{\partial f}{\partial \mathbf{x}}\frac{\partial \mathbf{x}}{\partial (s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{= \frac{\partial f}{\partial \mathbf{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{= \frac{\partial \mathbf{x}}{\partial (s,t)}} = \begin{bmatrix} \frac{\partial f}{\partial s} & \frac{\partial f}{\partial t} \end{bmatrix}$$

The gradient likes to be dotted.

The Hessian Matrix

- Consider a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ of two variables x_1 and x_2.
- We consider the second-order partial derivatives, for which:

$$\frac{\partial^2 f}{\partial x_1 \partial x_2} = \frac{\partial^2 f}{\partial x_2 \partial x_1}$$

- The Hessian matrix is the collection of these second-order partial derivatives.

$$\boldsymbol{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} \end{bmatrix}$$

- The Hessian measures the **local curvature** at some point (x, y).
- The gradient tells us about the local slope, i.e. steepness of function.
- The Hessian tells us how the slope is changing, so in a sense is the "derivative of the slope."

The Hessian Matrix

- For $f\colon \mathbb{R}^n \to \mathbb{R}, \boldsymbol{x} \mapsto f(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$.

$$\boldsymbol{H}(f) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

The Hessian is symmetric.

Remember for a square matrix $\boldsymbol{A}$:

PD: $x^{\mathsf{T}} A x > 0$ PSD: $x^{\mathsf{T}} A x \geq 0$

ND: $x^{\mathsf{T}} A x < 0$ NSD: $x^{\mathsf{T}} A x \leq 0$

- The interplay between the Hessian matrix and its **definitiveness** properties is profound.
- If Hessian is positive definite (PD) at a point, the function is locally convex. If critical point, then the point is local minimum.
- If negative definite (ND), then the function is locally concave. If critical point, then is local maximum.

Gradients of Vector-Valued Functions

- We discussed partial derivatives and gradients of function $f: \mathbb{R}^n \to \mathbb{R}$
- We will generalize the concept of the gradient to **vector-valued functions** (vector fields) $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$, where $n \geq 1$ and $m > 1$.
- For a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m$$

- Writing the vector-valued function in this way allows us to view a vector-valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ as a vector of functions $[f_1, \ldots, f_m]^{\mathrm{T}}$, $f_i: \mathbb{R}^n \to \mathbb{R}$ that map onto $\mathbb{R}$.

- The differentiation rules for every f_i are exactly the ones we discussed before.

Gradients of Vector-Valued Functions

- The partial derivative of a vector-valued function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}$, $i = 1, \ldots n$, is given as the vector

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \ldots, x_{i-1}, x_i + h, x_{i+1}, \ldots, x_n) - f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m$$

- In above, every partial derivative $\dfrac{\partial \boldsymbol{f}}{\partial x_i}$ is a column vector.

- To obtain the gradient of $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$ we collect these partial derivatives:

$$\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m \times n}$$

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m$$

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m$$

$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \dfrac{\partial f_1}{\partial x_i} \\ \vdots \\ \dfrac{\partial f_m}{\partial x_i} \end{bmatrix}$$

$$\nabla_{\mathbf{x}} f_i = \left[\frac{\partial f_i}{\partial x_1} \;\; \frac{\partial f_i}{\partial x_2} \;\; \cdots \;\; \frac{\partial f_i}{\partial x_n} \right] \in \mathbb{R}^{1 \times n}$$

$$\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \left[\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} \;\; \cdots \;\; \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \right] = \begin{bmatrix} \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \dfrac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \dfrac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{bmatrix}$$

The Jacobian

The Jacobian

- The collection of all first-order partial derivatives of a **vector-valued** function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^m$ is called the Jacobian. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

$$\boldsymbol{J} = \nabla_{\boldsymbol{x}} \boldsymbol{f} = \frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

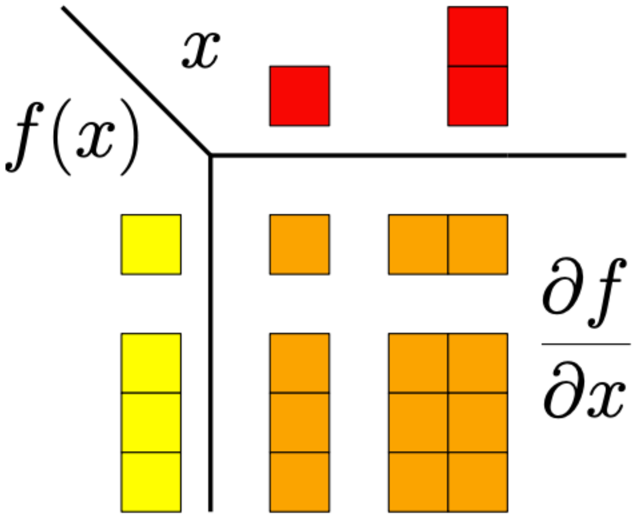
$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad \boldsymbol{J}(i,j) = \frac{\partial f_i}{\partial x_j}$$

- The elements of $\boldsymbol{f}$ define the rows and the elements of $\boldsymbol{x}$ define the columns of the corresponding Jacobian
- Special case: for a function $\boldsymbol{f}: \mathbb{R}^n \to \mathbb{R}^1$ which maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a scalar, i.e., $m = 1$, the Jacobian is a row vector of dimension $1 \times n$.

To note

- If $f\colon \mathbb{R} \to \mathbb{R}$, the gradient is a scalar
- If $f\colon \mathbb{R}^D \to \mathbb{R}$, the gradient is a $1 \times D$ row vector
- If $\boldsymbol{f}\colon \mathbb{R} \to \mathbb{R}^E$, the gradient is a $E \times 1$ column vector
- If $\boldsymbol{f}\colon \mathbb{R}^D \to \mathbb{R}^E$, the gradient is an $E \times D$ matrix



Example - Gradient of a Vector-Valued Function

- We are given $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^M, \quad \boldsymbol{A} \in \mathbb{R}^{M\times N}, \quad \boldsymbol{x} \in \mathbb{R}^N$.
- To compute the gradient $d\boldsymbol{f}/d\boldsymbol{x}$ we first determine the dimension of $d\boldsymbol{f}/d\boldsymbol{x}$: Since $\boldsymbol{f}: \mathbb{R}^N \to \mathbb{R}^M$, it follows that $d\boldsymbol{f}/d\boldsymbol{x} \in \mathbb{R}^{M\times N}$.
- Then, we determine the partial derivatives of $\boldsymbol{f}$ with respect to every x_j:

$$f_i(\boldsymbol{x}) = \sum_{j=1}^{N} A_{ij}x_j \Rightarrow \frac{\partial f_i}{\partial x_j} = A_{ij}$$

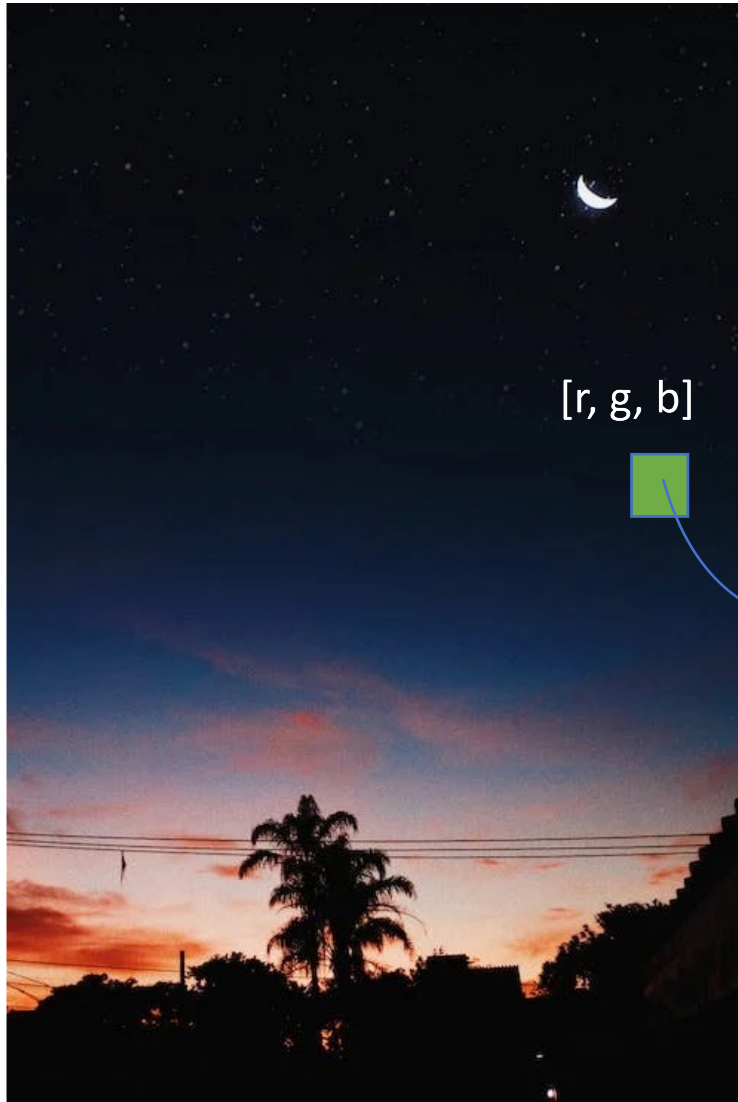
- We collect the partial derivatives in the Jacobian and obtain the gradient

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_M}{\partial x_1} & \cdots & \frac{\partial f_M}{\partial x_N} \end{bmatrix} = \begin{bmatrix} A_{11} & \cdots & A_{1N} \\ \vdots & & \vdots \\ A_{M1} & \cdots & A_{MN} \end{bmatrix} = \boldsymbol{A} \in \mathbb{R}^{M\times N}$$

What even is a tensor? The machine learning answer.

Height

# of rows



Tensor of dimension 3:
Height x Width x 3

Width # of columns

Example #1 - Chain Rule

- Consider the function $h: \mathbb{R} \to \mathbb{R}$, $h(t) = (f \circ g)(t)$ with

$f: \mathbb{R}^2 \to \mathbb{R}$

$g: \mathbb{R} \to \mathbb{R}^2$

$f(\boldsymbol{x}) = \exp\left(x_1 x_2^2\right)$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = g(t) = \begin{bmatrix} t\cos t \\ t\sin t \end{bmatrix}$$

- We compute the gradient of h with respect to t. Since $f: \mathbb{R}^2 \to \mathbb{R}$ and $g: \mathbb{R} \to \mathbb{R}^2$ we note that

$\frac{\partial f}{\partial \boldsymbol{x}} \in \mathbb{R}^{1\times 2}, \quad \frac{\partial g}{\partial t} \in \mathbb{R}^{2\times 1}$

- The desired gradient is computed by applying the chain rule:

$$\frac{dh}{dt} = \frac{\partial f}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{x}}{\partial t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial t} \end{bmatrix}$$

$$= \begin{bmatrix} \exp\left(x_1 x_2^2\right) x_2^2 & 2\exp\left(x_1 x_2^2\right) x_1 x_2 \end{bmatrix} \begin{bmatrix} \cos t - t\sin t \\ \sin t + t\cos t \end{bmatrix}$$

$$= \exp\left(x_1 x_2^2\right) \left(x_2^2 \left(\cos t - t \sin t\right) + 2 x_1 x_2 \left(\sin t + t\cos t\right) \right)$$

where $x_1 = t\cos t$ and $x_2 = t \sin t$

Example #2 - Gradient of a Least-Squares Loss in a Linear Model

- Let us consider the linear model

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} \in \mathbb{R}^{D}$ is a parameter vector, $\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$ are input features and $\boldsymbol{y} \in \mathbb{R}^{N}$ are the corresponding observations. We define the functions

$$L(\boldsymbol{e}) := \|\boldsymbol{e}\|^{2},$$

$$\boldsymbol{e}(\boldsymbol{\theta}) := \boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}$$

- We seek $\frac{\partial L}{\partial \boldsymbol{\theta}}$, and we will use the chain rule for this purpose. L is called a least-squares loss function.
- First, we determine the dimensionality of the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} \in \mathbb{R}^{1 \times D}$$

- The chain rule allows us to compute the gradient as

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \frac{\partial L}{\partial \boldsymbol{e}} \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}}$$

Example #2 - Gradient of a Least-Squares Loss in a Linear Model

- We know that $||\boldsymbol{e}||^2 = \boldsymbol{e}^{\mathrm{T}}\boldsymbol{e}$ and determine

$$\frac{\partial L}{\partial \boldsymbol{e}} = 2\boldsymbol{e}^{\mathrm{T}} \in \mathbb{R}^{1\times N}$$

- Further, we obtain

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\theta}} = -\boldsymbol{\Phi} \in \mathbb{R}^{N\times D}$$

- Our desired derivative is

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = -2\boldsymbol{e}^{\mathrm{T}}\boldsymbol{\Phi} = -\underbrace{2\left(\boldsymbol{y}^{\mathrm{T}} - \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\right)}_{1\times \mathrm{N}}\underbrace{\boldsymbol{\Phi}}_{\mathrm{N}\times \mathrm{D}} \in \mathbb{R}^{1\times D}$$

Example #3: Gradients of Matrices

- Consider the following:

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f} \in \mathbb{R}^{M}, \quad \boldsymbol{A} \in \mathbb{R}^{M \times N}, \quad \boldsymbol{x} \in \mathbb{R}^{N}$$

- We seek the gradient $\dfrac{d\boldsymbol{f}}{d\boldsymbol{A}}$
- First, we determine the dimension of the gradient

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} \in \mathbb{R}^{M \times (M \times N)}$$

- By definition, the gradient is the collection of the partial derivatives:

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \dfrac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \dfrac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1 \times (M \times N)}$$

- To compute the partial derivatives, we explicitly write out the matrix vector multiplication

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \cdots, M,$$

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \cdots, M$$

- The partial derivatives are then given as

$$\frac{\partial f_i}{\partial A_{iq}} = x_q$$

- Partial derivatives of f_i with respect to a row of $\boldsymbol{A}$ are given as

$$\frac{\partial f_i}{\partial A_{i,:}} = \boldsymbol{x}^{\mathrm{T}} \in \mathbb{R}^{1\times 1\times N}, \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = \mathbf{0}^{\mathrm{T}} \in \mathbb{R}^{1\times 1\times N}$$

- Since f_i maps onto $\mathbb{R}$ and each row of $\boldsymbol{A}$ is of size $1 \times N$, we obtain a $1 \times 1 \times N$ sized tensor as the partial derivative of f_i with respect to a row of $\boldsymbol{A}$.

- We stack the partial derivatives and get the desired gradient

$$\frac{\partial f_i}{\partial \boldsymbol{A}} = \begin{bmatrix} \mathbf{0}^{\mathrm{T}} \\ \vdots \\ \mathbf{0}^{\mathrm{T}} \\ \boldsymbol{x}^{\mathrm{T}} \\ \mathbf{0}^{\mathrm{T}} \\ \vdots \\ \mathbf{0}^{\mathrm{T}} \end{bmatrix} \in \mathbb{R}^{1\times(M\times N)}$$

Example #4: Gradient of Matrices with Respect to Matrices

- Consider a matrix $\boldsymbol{R} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{f}: \mathbb{R}^{M\times N} \to \mathbb{R}^{N\times N}$ with

$$\boldsymbol{f}(\boldsymbol{R}) = \boldsymbol{R}^{\mathrm{T}}\,\boldsymbol{R} =: \boldsymbol{K} \in \mathbb{R}^{N\times N}$$

- We seek the gradient $\dfrac{d\boldsymbol{K}}{d\boldsymbol{R}}$
- First, the dimension of the gradient is given as

$$\frac{d\boldsymbol{K}}{d\boldsymbol{R}} \in \mathbb{R}^{(N\times N)\times(M\times N)}$$

$$\frac{dK_{pq}}{d\boldsymbol{R}} \in \mathbb{R}^{1\times M\times N}$$

for $p,\ q = 1,\ldots,N$, where K_{pq} is the pqth entry of $\boldsymbol{K} = \boldsymbol{f}(\boldsymbol{R})$.

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^{\mathrm{T}}\boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp}R_{mq}$$

Example #4: Gradient of Matrices with Respect to Matrices

- Denoting the ith column of $\boldsymbol{R}$ by $\boldsymbol{r}_i$, every entry of $\boldsymbol{K}$ is given by the dot product of two columns of $\boldsymbol{R}$, i.e.,

$$K_{pq} = \boldsymbol{r}_p^{\mathrm{T}} \boldsymbol{r}_q = \sum_{m=1}^{M} R_{mp} R_{mq}$$

- We now compute the partial derivative $\frac{\partial K_{pq}}{\partial R_{ij}}$, we obtain

$$\frac{\partial K_{pq}}{\partial R_{ij}} = \sum_{m=1}^{M} \frac{\partial}{\partial R_{ij}} R_{mp} R_{mq} = \partial_{pqij}$$

$$\partial_{pqij} = \begin{cases} R_{iq} & if \;\; j = p, p \neq q \\ R_{ip} & if \;\; j = q, p \neq q \\ 2R_{iq} & if \;\; j = p, p = q \\ 0 & \text{otherwise} \end{cases}$$

- The desired gradient has the dimension $(N \times N) \times (M \times N)$, and every single entry of this tensor is given by ∂_{pqij}, where $p,\; q,\; j = 1, \ldots, N$ and $i = 1, \ldots, M$

Useful Identities for Computing Gradients

- Some useful gradients that are frequently required in machine learning.

Note that the trace of a square matrix, tr$(A) = \sum_{i=1}^{n} A_{ii}$.

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{a}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\mathrm{T}}$$

You should be able to calculate these gradients.

$$\frac{\partial \boldsymbol{a}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{b}}{\partial \boldsymbol{X}} = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$$

$$\frac{\partial \boldsymbol{x}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\mathrm{T}}(\boldsymbol{B} + \boldsymbol{B}^{\mathrm{T}})$$

$$\frac{\partial}{\partial \boldsymbol{s}}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s}) = -2(\boldsymbol{x} - \boldsymbol{A}\boldsymbol{s})^{\mathrm{T}}\boldsymbol{W}\boldsymbol{A} \quad \text{for symmetric } \boldsymbol{W}$$