Tensor Derivatives

Linear Algebra

Department of Computer Engineering

Sharif University of Technology

Hamid R. Rabiee rabiee@sharif.edu

Maryam Ramezani maryam.ramezani@sharif.edu

Introduction

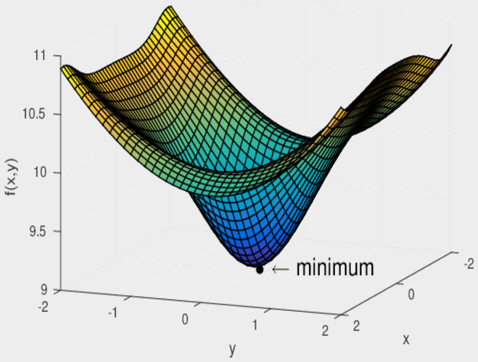
Overview

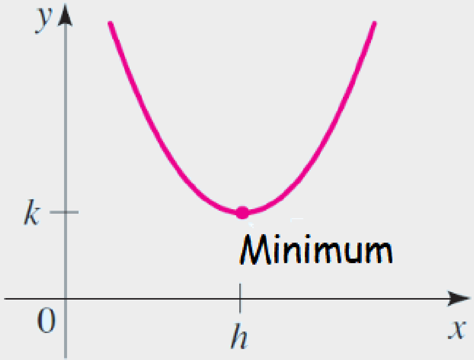
Types of matrix derivative

Types	Scalar	Vector	Matrix
Scalar	$\frac{\partial y}{\partial x}$	$\frac{\partial \mathbf{y}}{\partial x}$	$\frac{\partial \mathbf{Y}}{\partial x}$
Vector	$\frac{\partial y}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$	
Matrix	$\frac{\partial y}{\partial \mathbf{X}}$	Tensor! (Optional part of this course)	

Motivation

- Machine Learning training requires one to evaluate how one vector changes with respect to another
- How output changes with respect to parameters
- How do we find minimum of a scalar function?
- How do we find minimum of two variables?



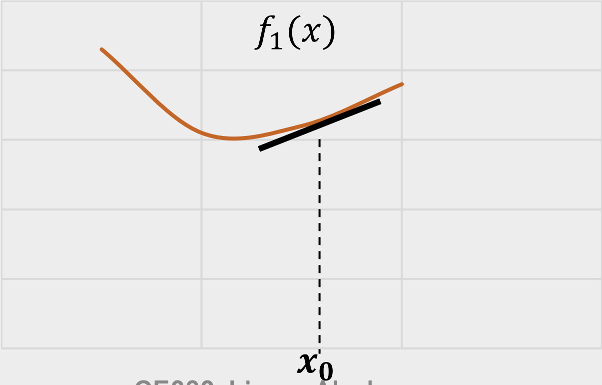


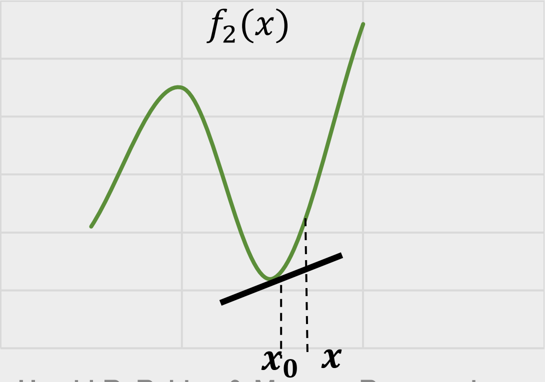
Vector-Valued Function

- Derivative of a vector-valued function $f: \mathbb{R} \rightarrow \mathbb{R}^n$ with respect to scalar $x \in \mathbb{R}$:

$$\frac{\partial f(x)}{\partial \boldsymbol{x}} \triangleq \begin{bmatrix} \frac{\partial f_1(x)}{\partial x} \\ \frac{\partial f_2(x)}{\partial x} \\ \vdots \\ \frac{\partial f_n(x)}{\partial x} \end{bmatrix}$$

$$f(x) \approx f(x_0) + m(x - x_0) \qquad m = \begin{bmatrix} f_1'(x_0) \\ .. \\ f_n'(x_0) \end{bmatrix}$$



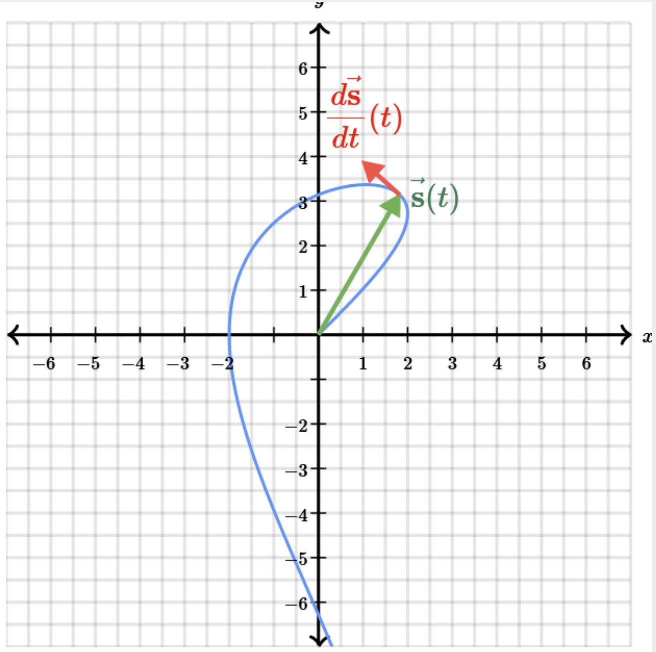


Example

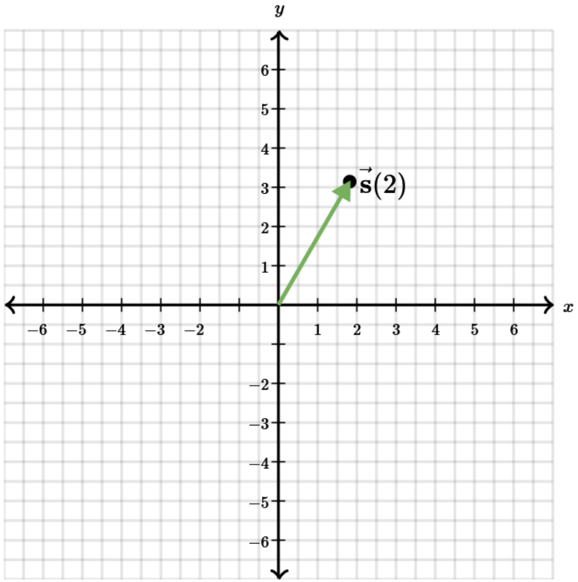
$$f(x) = \begin{bmatrix} \sin(x) \\ cos(x) \end{bmatrix}$$

Vector-Valued Function

$$\vec{\mathbf{s}}(t) = \begin{bmatrix} 2\sin(t) \\ 2\cos(t/3)t \end{bmatrix}$$



$$\vec{\mathbf{s}}(2) = \begin{bmatrix} 2\sin(2) \\ 2\cos(2/3)\cdot 2 \end{bmatrix} \approx \begin{bmatrix} 1.819 \\ 3.144 \end{bmatrix}$$



Vector-Valued Function

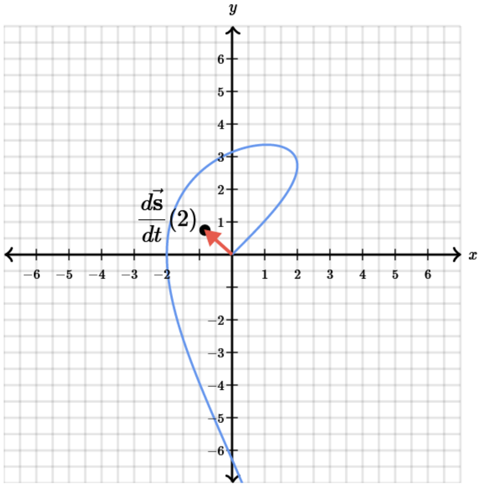
$$\frac{d\vec{s}}{dt}(t) = \begin{bmatrix} \frac{d}{dt}(2\sin(t)) \\ \\ \frac{d}{dt}(2\cos(t/3))t \end{bmatrix}$$

$$= \begin{bmatrix} 2\cos(t) \\ \\ 2\cos(t/3) - \frac{2}{3}\sin(t/3)t \end{bmatrix}$$

$$\frac{d\vec{s}}{dt}(2) = \begin{bmatrix} 2\cos(2) \\ \\ 2\cos(2/3) - \frac{2}{3}\sin(2/3)\cdot 2 \end{bmatrix}$$

$$\approx \begin{bmatrix} -0.832 \\ \\ 0.747 \end{bmatrix}$$

This is also some two-dimensional vector.



Matrix–Valued Function

- Derivative of a matrix–valued function $f: \mathbb{R} \to \mathbb{R}^{m\times n}$ with respect to scalar $x \in \mathbb{R}$:

$$\frac{\partial f(x)}{\partial x} \triangleq \begin{bmatrix} \frac{\partial f_{11}(x)}{\partial x} & \frac{\partial f_{12}(x)}{\partial x} & \cdots & \frac{\partial f_{1n}(x)}{\partial x} \\ \frac{\partial f_{21}(x)}{\partial x} & \frac{\partial f_{22}(x)}{\partial x} & \cdots & \frac{\partial f_{2n}(x)}{\partial x} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_{m1}(x)}{\partial x} & \frac{\partial f_{m2}(x)}{\partial x} & \cdots & \frac{\partial f_{mn}(x)}{\partial x} \end{bmatrix}$$

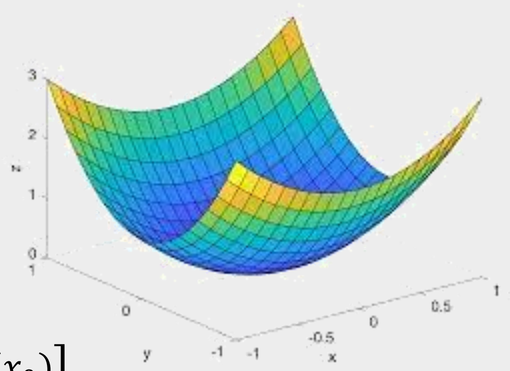
Example

- Rotation Matrix

Real-Valued Multivariant Function

- Derivative of a real-valued function $f: \mathbb{R}^n \to \mathbb{R}$ with respect to vector $\mathbf{x} \in \mathbb{R}^n$:

$$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \triangleq \begin{bmatrix} \dfrac{\partial f(x)}{\partial x_1} \\ \dfrac{\partial f(x)}{\partial x_2} \\ \vdots \\ \dfrac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

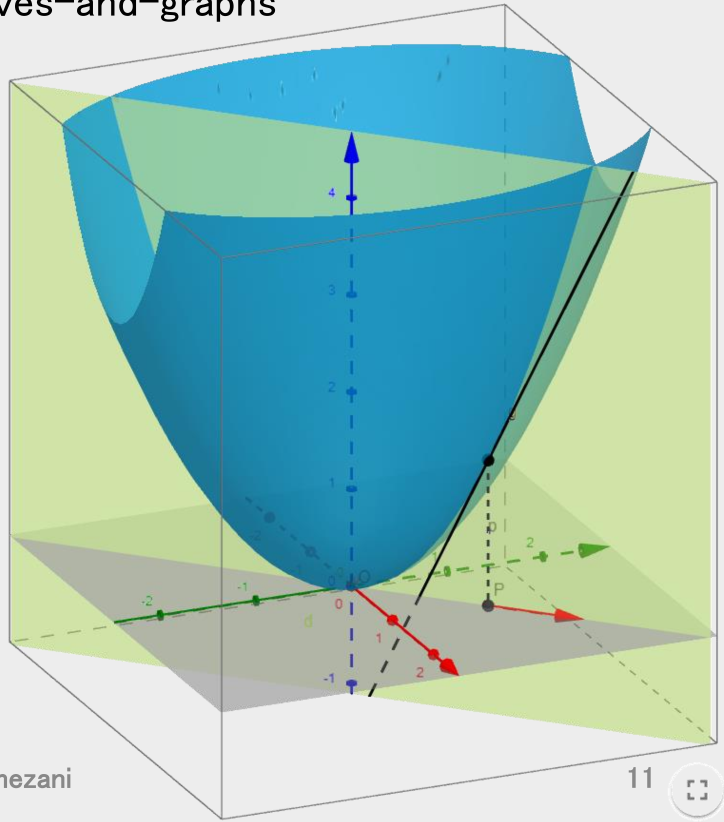


$$f(x) - f(x_0) = m^T(x - x_0) \qquad m = \begin{bmatrix} f_1'(x_0) \\ .. \\ f_n'(x_0) \end{bmatrix}$$

- Gradient

Directional Derivative

- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives/partial-derivatives/v/partial-derivatives-and-graphs
- https://www.geogebra.org/m/bxhwxr2x



Chain Rule

- $\dfrac{dY}{dx} = \dfrac{dY}{du}\dfrac{du}{dx}$ x, u:scalars Y: matrix
- $\dfrac{dy}{dX} = \dfrac{dy}{du}\dfrac{du}{dX}$ y, u:scalars X: matrix
- $\dfrac{dy}{dx} = \dfrac{dy}{du}\dfrac{du}{dx}$ x,y, u: vectors

Product Rule

- $(f(x)g(x))' = f'(x)g(x) + f(x)g'(x)$
- $\frac{\partial(AB)}{\partial\alpha} = A\frac{\partial(B)}{\partial\alpha} + \frac{\partial(A)}{\partial\alpha}B$ if A and B be matrices which elements are function of scalar α
- $\frac{\partial(x^T y)}{\partial z} = x^T\frac{\partial(y)}{\partial z} + y^T\frac{\partial(x)}{\partial z}$ if x and y be vectors which elements are function of vector z

Example

- $f, g: \mathbb{R} \to \mathbb{R}^n$ $\quad h(x) = f(x)^T g(x) \quad h'(x) = ?$
- $f: \mathbb{R}^n \to \mathbb{R} \quad g: \mathbb{R} \to \mathbb{R} \quad h(x) = f(x)g(x) \quad h'(x) = ?$
- $\mathrm{H}: \mathbb{R} \to \mathbb{R}^{m\times n}, \mathrm{F}: \mathbb{R} \to \mathbb{R}^{m\times p}, \mathrm{G}: \mathbb{R} \to \mathbb{R}^{p\times n} \quad H(x) = F(x)G(x)$

Motivation

- Derivative of a scalar function $f: \mathbb{R}^N \to \mathbb{R}$ with respect to vector $\mathbf{x} \in \mathbb{R}^N$:

- $$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \triangleq \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_N} \end{bmatrix}$$

- Derivative of a vector function $f: \mathbb{R}^N \to \mathbb{R}^M$ with respect to vector $\boldsymbol{x} \in \mathbb{R}^N$:

- $$\frac{\partial f(\boldsymbol{x})}{\partial \boldsymbol{x}} \triangleq \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_1(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_N} \\ \frac{\partial f_2(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_2(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_2(\boldsymbol{x})}{\partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_M(\boldsymbol{x})}{\partial x_1} & \frac{\partial f_M(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f_M(\boldsymbol{x})}{\partial x_N} \end{bmatrix}$$

Definitions

Definition

- Derivative of a scalar function $f: \mathbb{R}^{M\times N} \rightarrow \mathbb{R}$ with respect to matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$:

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} \triangleq \begin{bmatrix} \frac{\partial f(\boldsymbol{X})}{\partial X_{1,1}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{2,1}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{M,1}} \\ \frac{\partial f(\boldsymbol{X})}{\partial X_{1,2}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{2,2}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{M,2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f(\boldsymbol{X})}{\partial X_{1,N}} & \frac{\partial f(\boldsymbol{X})}{\partial X_{2,N}} & \cdots & \frac{\partial f(\boldsymbol{X})}{\partial X_{M,N}} \end{bmatrix}$$

- Using the above definitions, we can generalize the chain rule, Given $\boldsymbol{u} = \boldsymbol{h}(x)$ (i.e. $\boldsymbol{u}$ is a function of x) and $\boldsymbol{g}$ is a vector function of $\boldsymbol{u}$, the vector-by-vector chain rule states:

$$\frac{\partial \boldsymbol{g}(\boldsymbol{u})}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{g}(\boldsymbol{u})}{\partial \boldsymbol{u}}$$

- Board 😄

- Board 😄

Derivative Definition

DEFINITION

Suppose $z = f(x, y)$ is a function of two variables with a domain of D. Let $(a, b) \in D$ and define $\mathrm{u} = \cos\theta\, \mathbf{i} + \sin\theta\, \mathbf{j}$. Then the **directional derivative** of f in the direction of $\mathbf{u}$ is given by

$$D_{\mathbf{u}} f(a, b) = \lim_{h \to 0} \frac{f(a + h\cos\theta, b + h\sin\theta) - f(a, b)}{h},$$

provided the limit exists.

$$\nabla_{\vec{\mathbf{v}}} f(\mathbf{x}) = \lim_{h \to 0} \frac{f(\mathbf{x} + h\vec{\mathbf{v}}) - f(\mathbf{x})}{h||\vec{\mathbf{v}}||}$$

Conclusion

Try to proof the followings:

- ❑ $\frac{\partial(\boldsymbol{u}(\boldsymbol{x})+\boldsymbol{v}(\boldsymbol{x}))}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{u}(\boldsymbol{x})}{\partial \boldsymbol{x}} + \frac{\partial \boldsymbol{v}(\boldsymbol{x})}{\partial \boldsymbol{x}}$
- ❑ $\frac{\partial(\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{A}$
- ❑ $\frac{\partial(\boldsymbol{x}^T\boldsymbol{a})}{\partial \boldsymbol{x}} = \boldsymbol{a}^{\boldsymbol{T}}$
- ❑ $\frac{\partial(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = \boldsymbol{x}^{\boldsymbol{T}}(\boldsymbol{A} + \boldsymbol{A}^T)$
- ❑ $\frac{\partial(\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$ if $\boldsymbol{A}$ is symmetric

Hint!

$$A\vec{x} = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_1 x_1 + a_2 x_2 \\ a_3 x_1 + a_4 x_2 \end{bmatrix}$$

$$\frac{dA\vec{x}}{dx} = \begin{bmatrix} \frac{\partial(a_1 x_1 + a_2 x_2)}{\partial x_1} & \frac{\partial(a_1 x_1 + a_2 x_2)}{\partial x_2} \\ \frac{\partial(a_3 x_1 + a_4 x_2)}{\partial x_1} & \frac{\partial(a_3 x_1 + a_4 x_2)}{\partial x_2} \end{bmatrix}$$

$$= \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} = A$$

Conclusion

Important

1. Derivative of a linear function:

$$\frac{\partial}{\partial \vec{x}} \vec{a}.\vec{x} = \frac{\partial}{\partial \vec{x}} \vec{a}^T \vec{x} = \frac{\partial}{\partial \vec{x}} \vec{x}^T \vec{a} = \vec{a}^T$$

(If you think back to calculus, this is just like $\frac{\mathrm{d}}{\mathrm{dx}} \mathrm{ax} = \mathrm{a}$).

2. Derivative of a quadratic function:

$$\frac{\partial}{\partial x} \vec{x}^T A \vec{x} = 2A\vec{x}$$

(Again, if you think back to calculus, this is just like $\frac{\mathrm{d}}{\mathrm{dx}} \mathrm{ax}^2 = 2\mathrm{ax}$).

If you ever need it, the more general rule (for non-symmetric A) is:

$$\frac{\partial}{\partial x} \vec{x}^T A \vec{x} = x^T (A + A^T)$$

which of course is the same thing as $2A\vec{x}$ when A is symmetric.

Given $A = [a_{ij}]$, the (i,j)-cofactor of A is the number C_{ij} given by

$$C_{ij} = (-1)^{i+j} \det\left(A_{ij}\right)$$

Then

$$\det(A) \;= a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n}$$

Which is a cofactor expansion across the first row of A.

$$A^{-1} = \frac{1}{|A|}\begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix} = A^{-1} = \frac{1}{|A|}\, adj\ A$$

Adj A = C^T

The matrix of cofactors is called the adjugate (or classical adjoint) of A, denoted by $\operatorname{adj} A$.

Good Examples!!!

Proof the followings:

- ❑ $\dfrac{\partial \left(A(t)\right)^{-1}}{\partial t} = -A(t)^{-1}\,\dfrac{\partial \left(A(t)\right)}{\partial t}\,A(t)^{-1}$
- ❑ $\dfrac{\partial \det(A)}{\partial A} = \det(A)\,A^{-1}$
- ❑ $\dfrac{\partial \ln(\det(A))}{\partial A} = (A^{-1})^T$
- ❑ $\dfrac{\partial \det(A(t))}{\partial t} = \det(A)\,trace(A^{-1}\,\dfrac{\partial (A(t))}{\partial t})$
- ❑ $\dfrac{\partial trace(BA^{-1})}{\partial A} = -A^{-1}BA^{-1}$
- ❑ $\dfrac{\partial \left(y^T A x\right)}{\partial A} = yx^T$
- ❑ $\dfrac{\partial \left(x^T A x\right)}{\partial A} = xx^T$

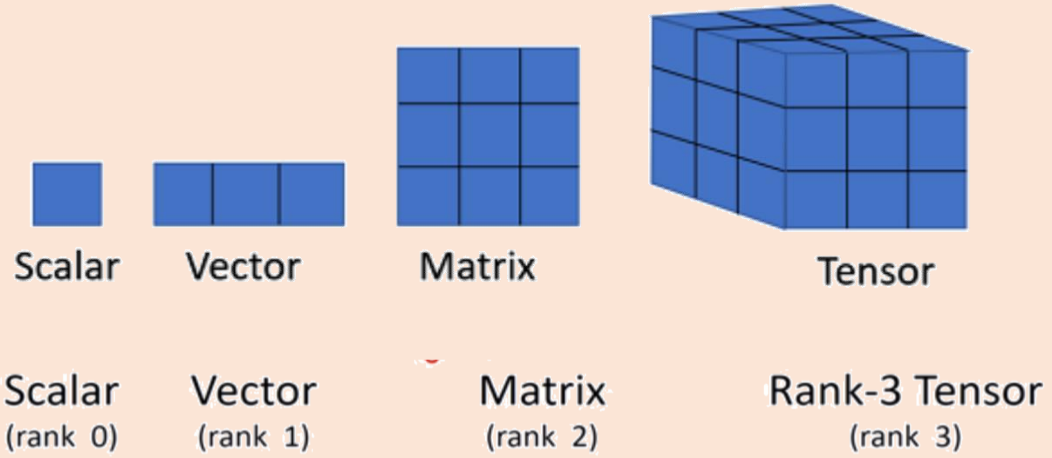
Tensor (Optional)

Tensor

Definition

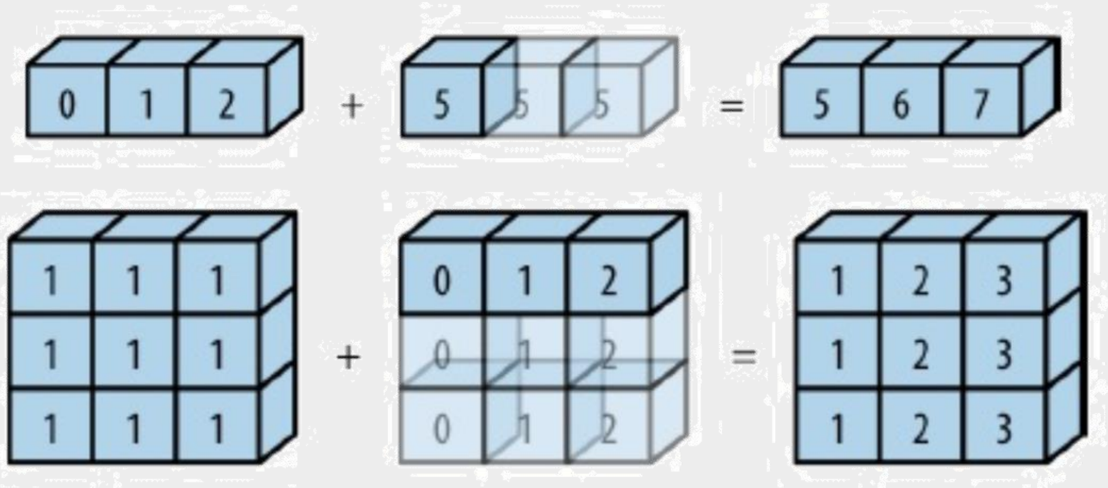
- Multi-dimensional array of numbers

```
w = torch.empty(3)
x = torch.empty(2, 3)
y = torch.empty(2, 3, 4)
z = torch.empty(2, 3, 2, 4)
```



Tensors Addition

- Adding tensors with same size
- Adding scalar to tensor
- Adding tensors with different size: if broacastable



Tensors Addition

- Two tensors are "broadcastable" if the following rules hold:
  - Each tensor has at least one dimension.
  - When iterating over the dimension sizes, starting at the trailing dimension, the dimension sizes must either be equal, one of them is 1, or one of them does not exist.
- Example
  - T1: (5,7,3) T2:(5,7,3)
  - T1: (5,3,4,1) T2:(3,1,1)

Notes

- Matrix Product on tensors

$$(m \times n) \cdot (n \times k) = (m \times k)$$

product is defined

Derivative of a vector with respect to a matrix

Derivative of a matrix with respect to a matrix

References

- Linear Algebra and Its Applications, David C. Lay
- Introduction to Applied Linear Algebra Vectors, Matrices, and Least Squares
- https://en.Wikipedia.org/wiki/matrix_calculus
- https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf
- https://www.kamperh.com/notes/kamper_matrixcalculus13.pdf