OUTLINE

I. Introduction

- Artificial Intelligence
- Machine Learning
- Neural Network and Deep Learning
- Applications

II. Background

III. Fitting a Model

IV. Supervised Learning

V. Unsupervised Learning

VI. Fantastic DNNs: How to choose them, how to train them

VII. Machine Learning in Robot Audition

OUTLINE

~~I. Introduction~~

II. Background

III. Fitting a Model

IV. Supervised Learning

V. Unsupervised Learning

VI. Fantastic DNNs: How to choose them, how to train them

VII. Machine Learning in Robot Audition

OUTLINE

~~I. Introduction~~

II. Background

- Multivalued Multivariate Functions
- Tensors
- Differential Calculus
- Exercises

III. Fitting a Model

IV. Supervised Learning

V. Unsupervised Learning

VI. Fantastic DNNs: How to choose them, how to train them

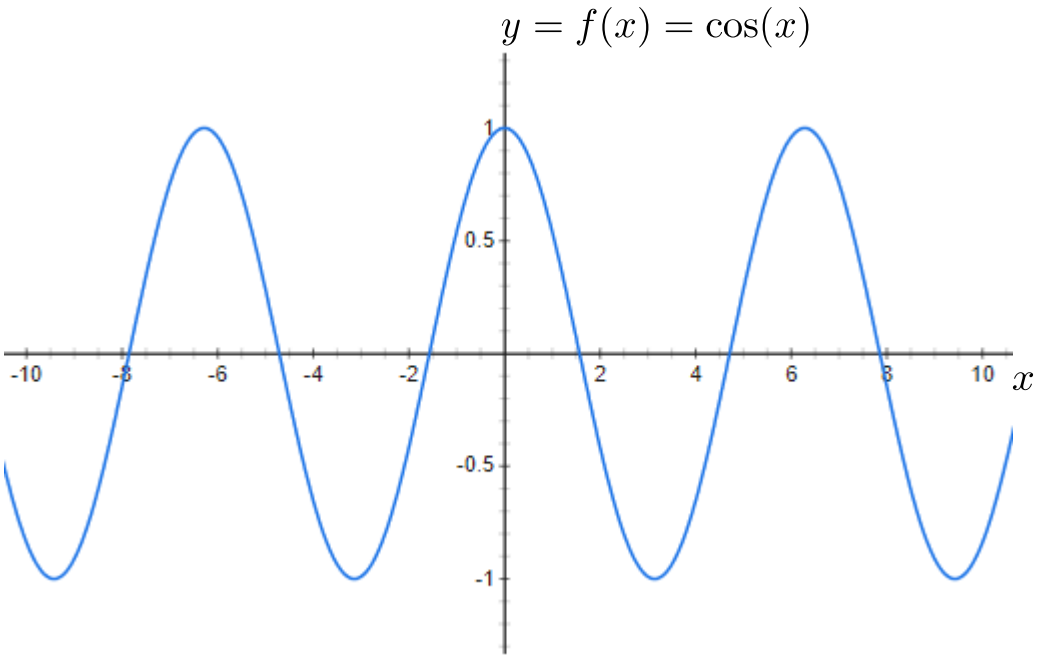
VII. Machine Learning in Robot Audition

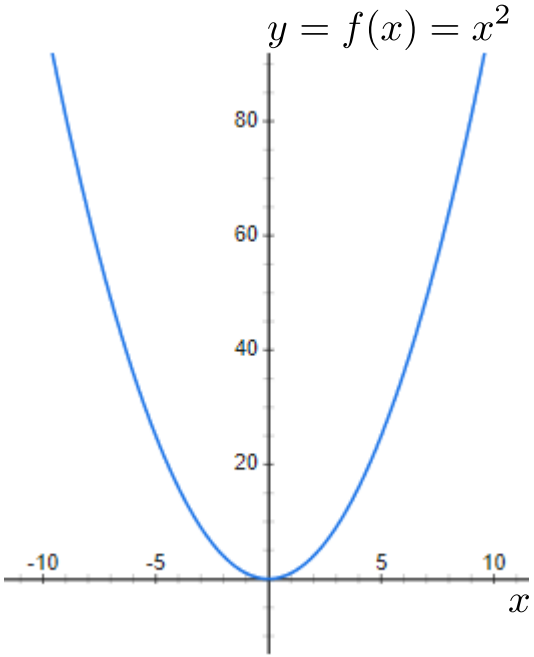
Simple functions

- In high-school calculus, we study functions from reals to reals, denoted $f : \mathbb{R} \to \mathbb{R}$.

Simple functions

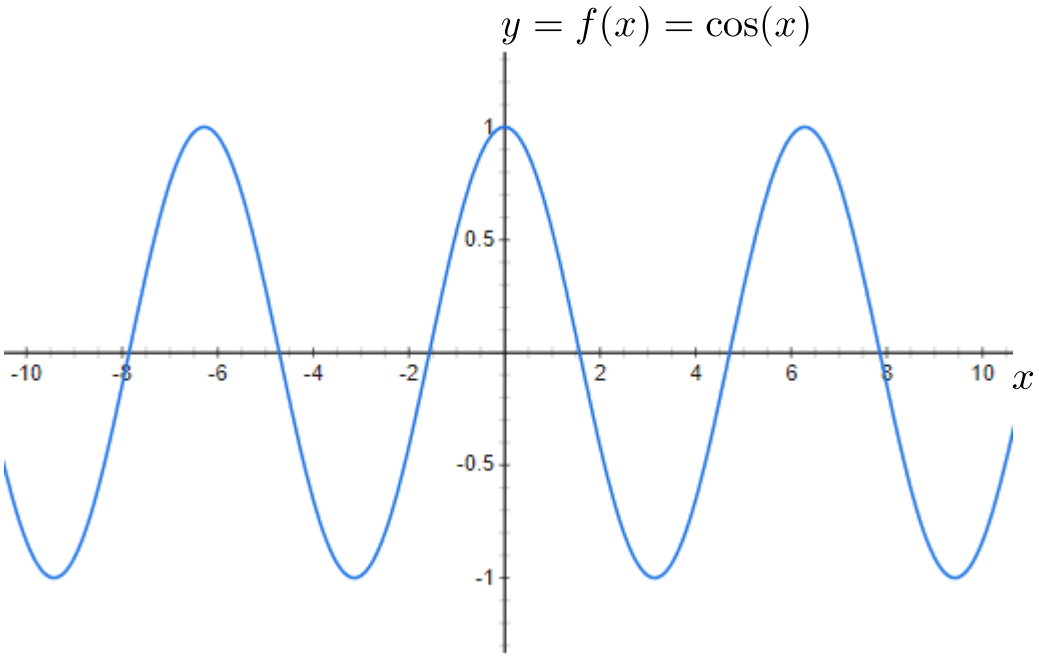
- In high-school calculus, we study functions from reals to reals, denoted $f : \mathbb{R} \to \mathbb{R}$.
- Here are some ***graphs*** of functions:

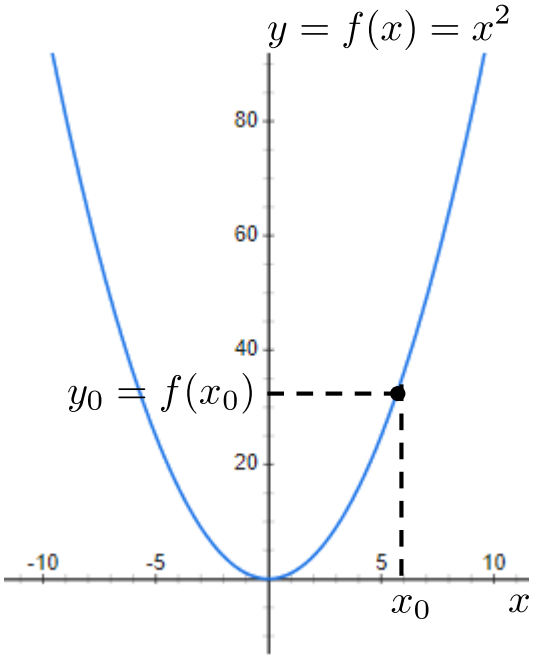




Simple functions

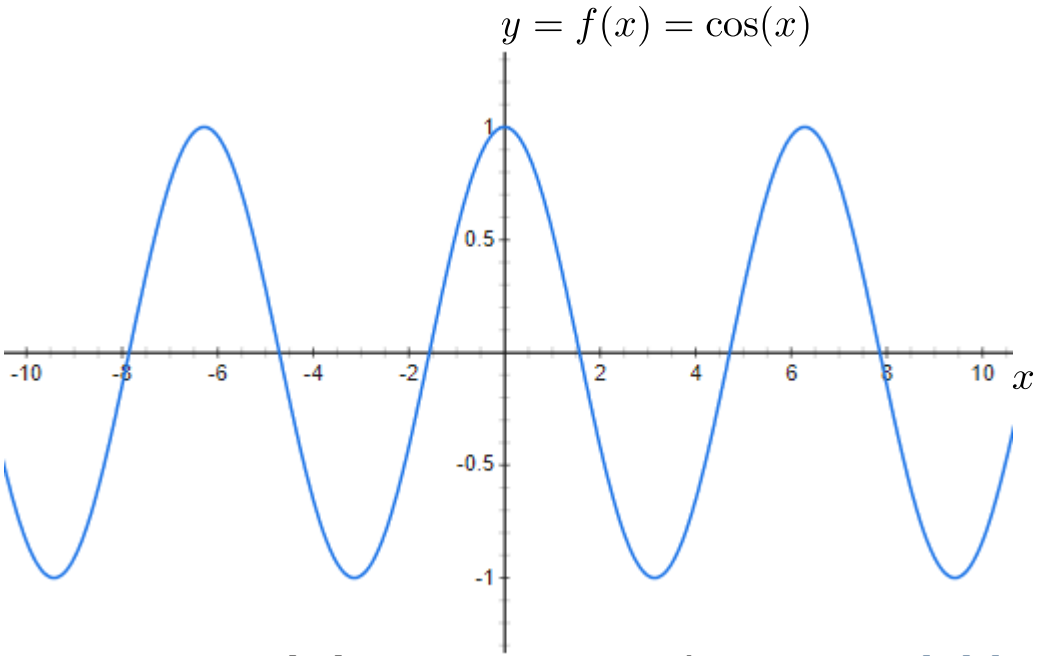
- In high-school calculus, we study functions from reals to reals, denoted $f : \mathbb{R} \to \mathbb{R}$.
- Here are some ***graphs*** of functions:

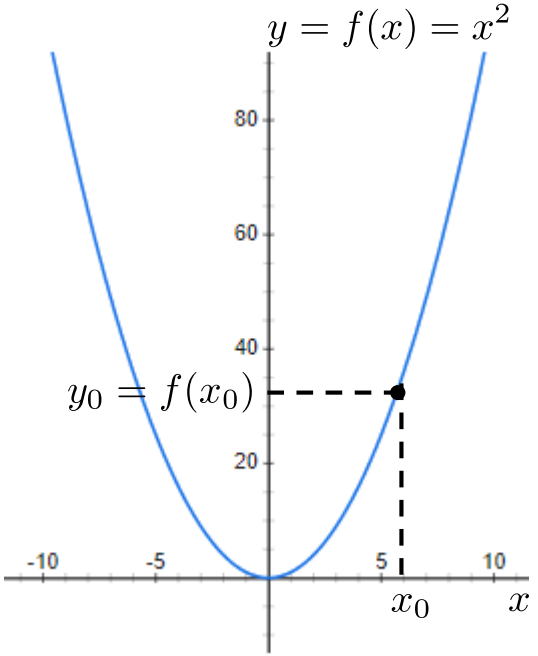




Simple functions

- In high-school calculus, we study functions from reals to reals, denoted $f : \mathbb{R} \to \mathbb{R}$.
- Here are some ***graphs*** of functions:





- **Important subtlety:** Here, x and y are ***variables*** that ***depend*** on each other, x_0 and y_0 are ***constants***, and f is a ***function*** that can be ***applied*** to a variable or a constant.

Multivalued functions

- This straightforwardly generalizes to **multivalued** functions $f : \mathbb{R} \to \mathbb{R}^N$:

$$\boldsymbol{y} = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_N(x) \end{bmatrix} \in \mathbb{R}^N$$

Multivalued functions

- This straightforwardly generalizes to **multivalued** functions $f : \mathbb{R} \to \mathbb{R}^N$:

$$\boldsymbol{y} = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_N(x) \end{bmatrix} \in \mathbb{R}^N$$

Example:

$$f : \begin{cases} [0, 2\pi] & \to & \mathbb{R}^2 \\ \theta & \mapsto & f(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \end{cases}$$

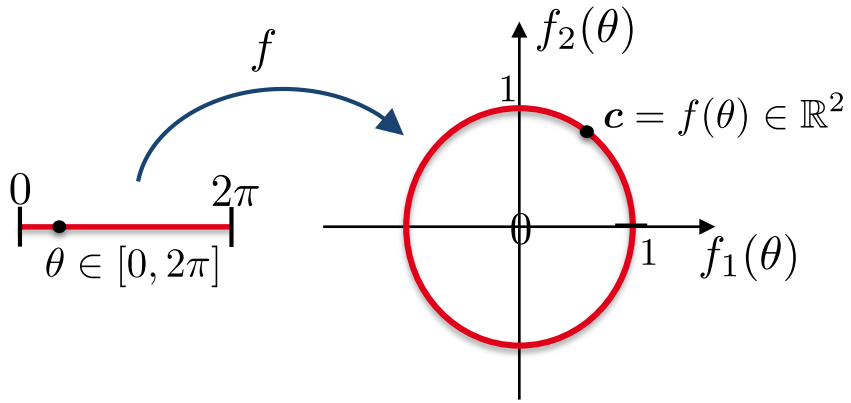
Multivalued functions

- This straightforwardly generalizes to **multivalued** functions $f : \mathbb{R} \to \mathbb{R}^N$:

$$\boldsymbol{y} = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_N(x) \end{bmatrix} \in \mathbb{R}^N$$

Example:

$$f : \begin{cases} [0, 2\pi] & \to & \mathbb{R}^2 \\ \theta & \mapsto & f(\theta) = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \end{cases}$$

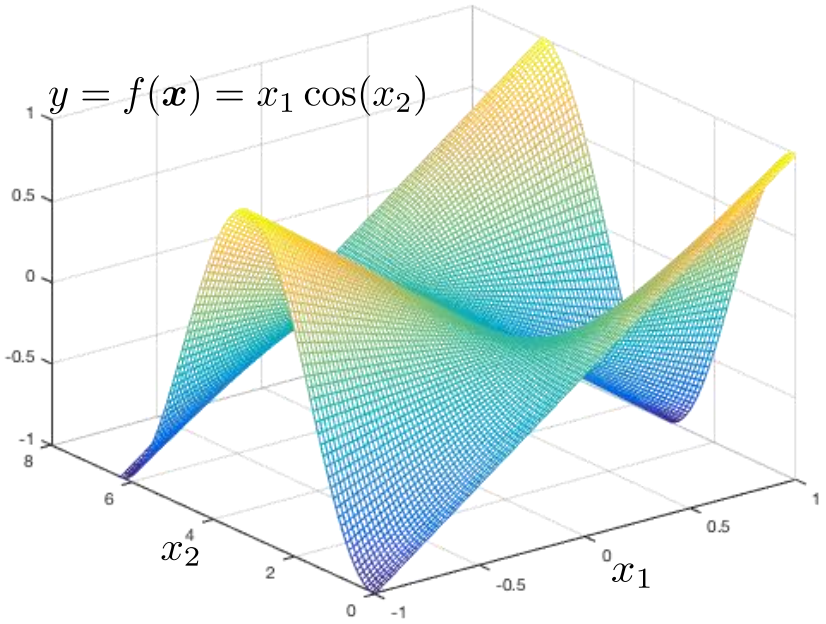


Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}$

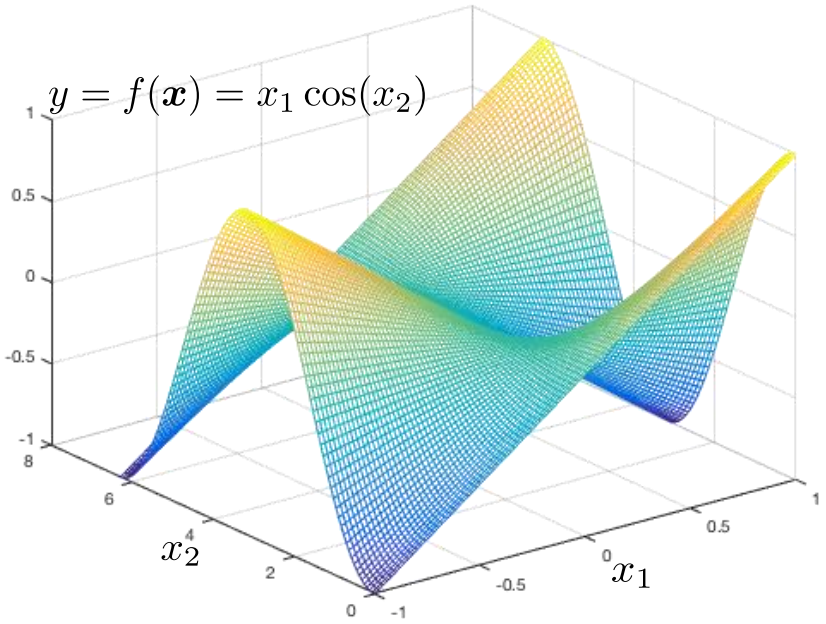
Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}$



Multivariate functions

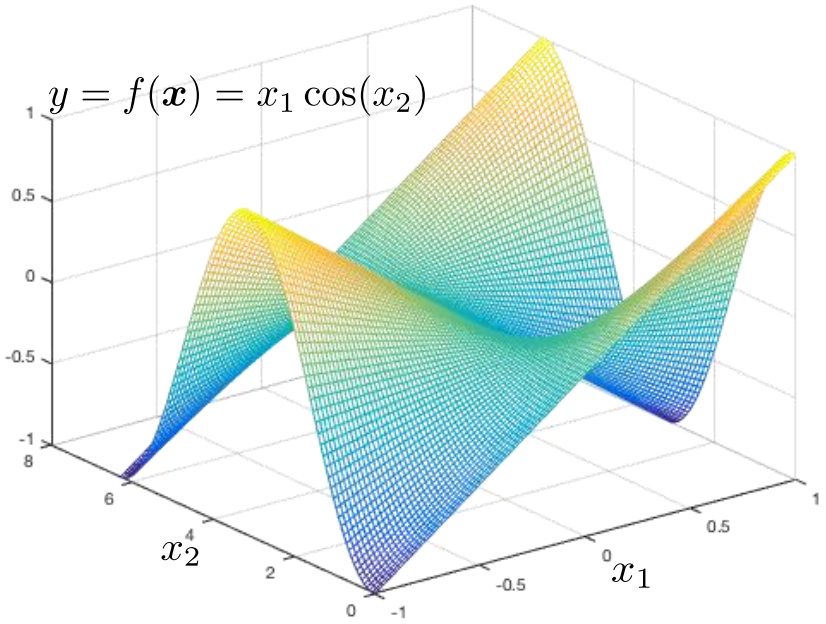
- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}$



Important examples:

Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}$

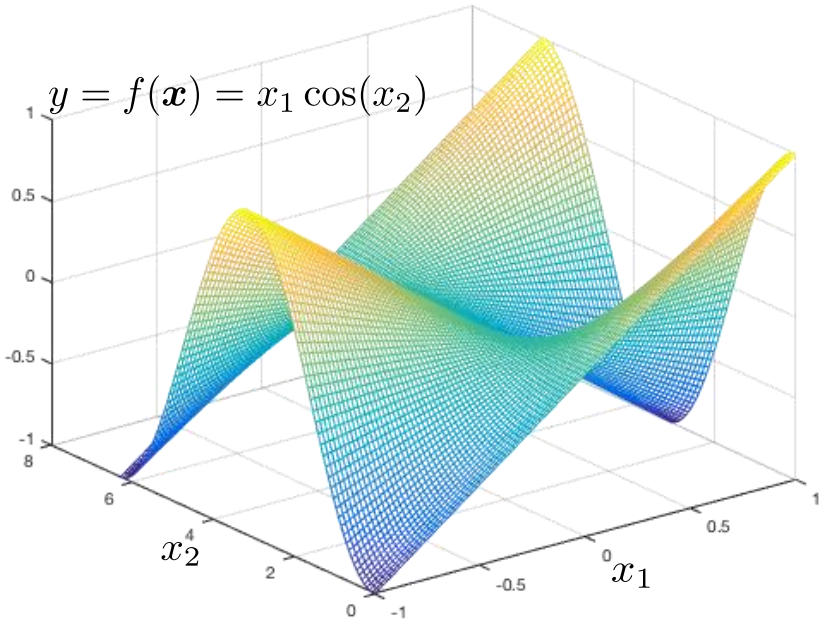


Important examples:

- ***Linear forms***: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \sum_{d=1}^{D} w_d x_d$

Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}$

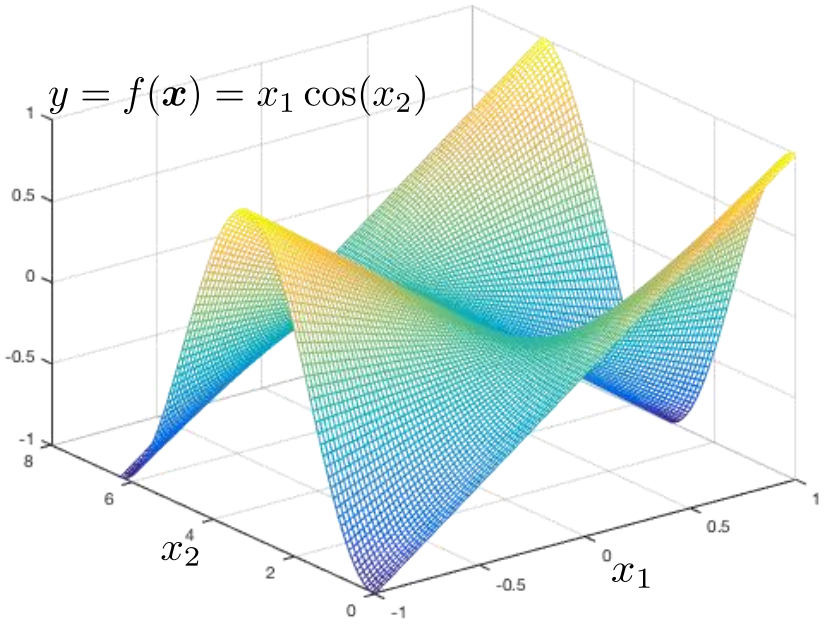


Important examples:

- ***Linear forms***: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \sum_{d=1}^{D} w_d x_d$ → *represented* by a **row vector** : $\boldsymbol{w}^\top \in \mathbb{R}^{1 \times D}$

Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}$



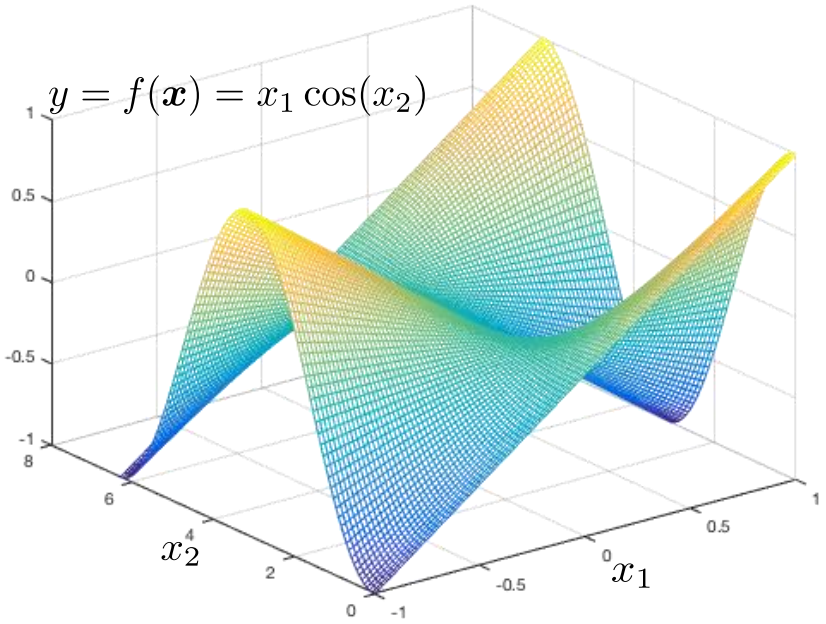
Important examples:

- ***Linear forms***: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \sum_{d=1}^{D} w_d x_d$ → *represented* by a **row vector** : $\boldsymbol{w}^\top \in \mathbb{R}^{1 \times D}$

- ***Affine forms:*** $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$

Multivariate functions

- Another generalization are **real-valued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}$



Important examples:

- ***Linear forms***: $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = \langle \boldsymbol{w}, \boldsymbol{x} \rangle = \sum_{d=1}^{D} w_d x_d$ → *represented* by a **row vector** : $\boldsymbol{w}^\top \in \mathbb{R}^{1 \times D}$

- ***Affine forms:*** $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + b$

- The ***Euclidean norm***: $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^\top \boldsymbol{x} = \sum_{d=1}^{D} (x_d)^2$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D \times N}$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^{D} a_{1,d} x_d \\ \sum_{d=1}^{D} a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^{D} a_{N,d} x_d \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^D a_{1,d} x_d \\ \sum_{d=1}^D a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^D a_{N,d} x_d \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^D a_{1,d} x_d \\ \sum_{d=1}^D a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^D a_{N,d} x_d \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^D a_{1,d} x_d \\ \sum_{d=1}^D a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^D a_{N,d} x_d \end{bmatrix}$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \to \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^D a_{1,d} x_d \\ \sum_{d=1}^D a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^D a_{N,d} x_d \end{bmatrix}$$

$$= [\boldsymbol{a}_1, \dots, \boldsymbol{a}_D] \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \sum_{d=1}^D \boldsymbol{a}_d x_d$$

Multivalued Multivariate functions

- Finally, combining the two, we get **multivalued multivariate functions**, i.e., $f : \mathbb{R}^D \rightarrow \mathbb{R}^N$.

Important examples:

- ***Linear maps***: $f(\boldsymbol{x}+\boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$

  They can be *represented* by **matrices** : $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x},\ \mathbf{A} \in \mathbb{R}^{D\times N}$

$$f(\boldsymbol{x}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,D} \\ a_{2,1} & a_{2,2} & \dots & a_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N,1} & a_{L,2} & \dots & a_{N,D} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \begin{bmatrix} \hat{\boldsymbol{a}}_1^\top \\ \hat{\boldsymbol{a}}_2^\top \\ \vdots \\ \hat{\boldsymbol{a}}_N^\top \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} \langle \hat{\boldsymbol{a}}_1, \boldsymbol{x} \rangle \\ \langle \hat{\boldsymbol{a}}, \boldsymbol{x} \rangle \\ \vdots \\ \langle \hat{\boldsymbol{a}}_N, \boldsymbol{x} \rangle \end{bmatrix} = \begin{bmatrix} \sum_{d=1}^D a_{1,d} x_d \\ \sum_{d=1}^D a_{2,d} x_d \\ \vdots \\ \sum_{d=1}^D a_{N,d} x_d \end{bmatrix}$$

$$= [\boldsymbol{a}_1, \dots, \boldsymbol{a}_D] \times \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_D \end{bmatrix} = \sum_{d=1}^D \boldsymbol{a}_d x_d$$

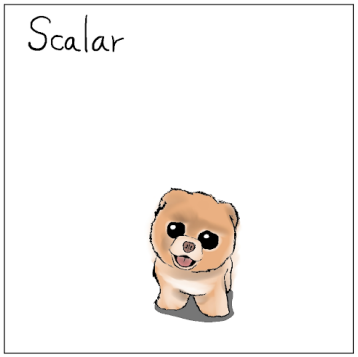
- ***Affine maps:*** $f(\boldsymbol{x}) = \mathbf{A}\boldsymbol{x} + \boldsymbol{b},\ \mathbf{A} \in \mathbb{R}^{N\times D},\ \boldsymbol{b} = \mathbb{R}^N$

Tensors

- Matrices and vectors can be generalized to **Tensors**

Tensors

- Matrices and vectors can be generalized to **Tensors**

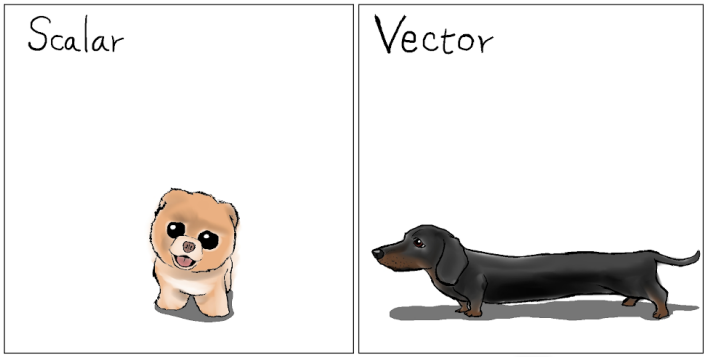


$a \in \mathbb{R}$

(**0**-way tensor)

Tensors

- Matrices and vectors can be generalized to **Tensors**



$a \in \mathbb{R}$

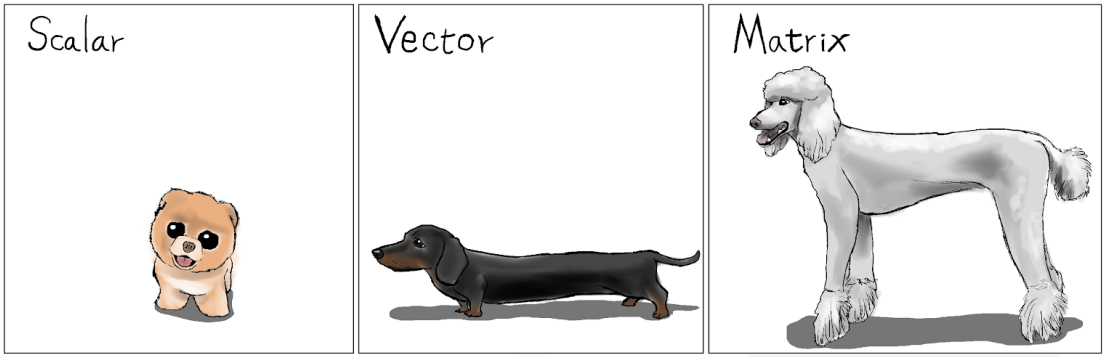
(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$

(**1**-way tensor)

Tensors

- Matrices and vectors can be generalized to **Tensors**

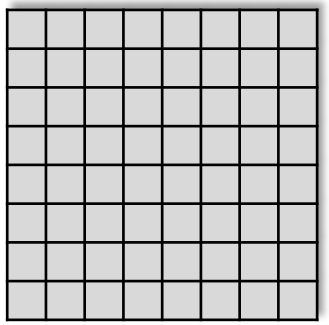


$a \in \mathbb{R}$

(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$

(**1**-way tensor)

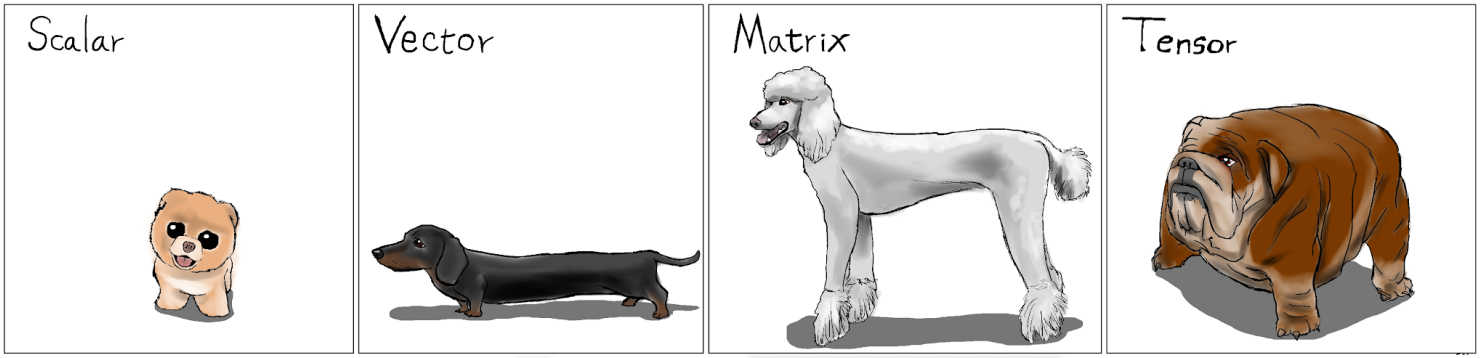


$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$

(**2**-way tensor)

Tensors

- Matrices and vectors can be generalized to **Tensors**



$a \in \mathbb{R}$
(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$
(**1**-way tensor)

$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$
(**2**-way tensor)

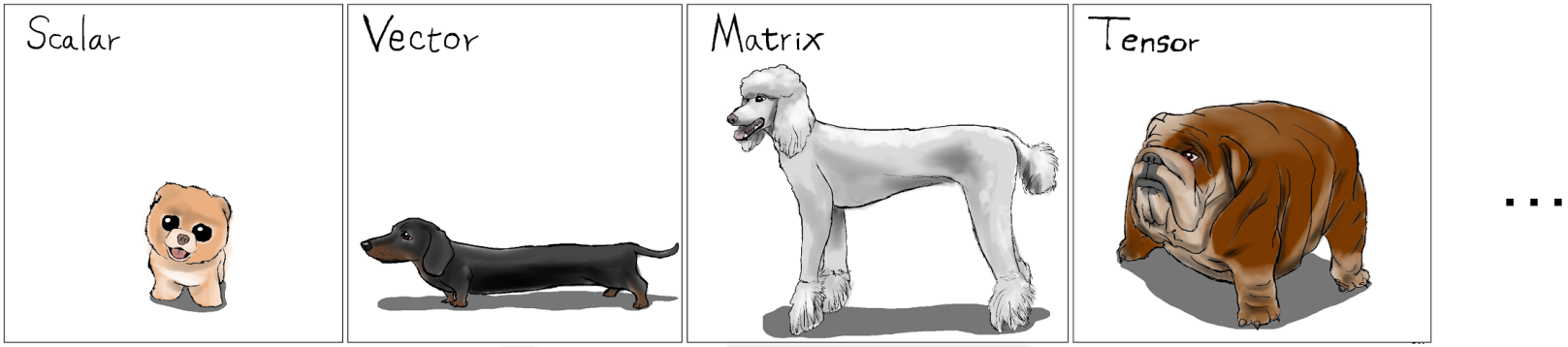
$[a_{ijk}]_{ijk} \in \mathbb{R}^{I \times J \times K}$
(**3**-way tensor)

...

...

Tensors

- Matrices and vectors can be generalized to **Tensors**



$a \in \mathbb{R}$
(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$
(**1**-way tensor)

rows

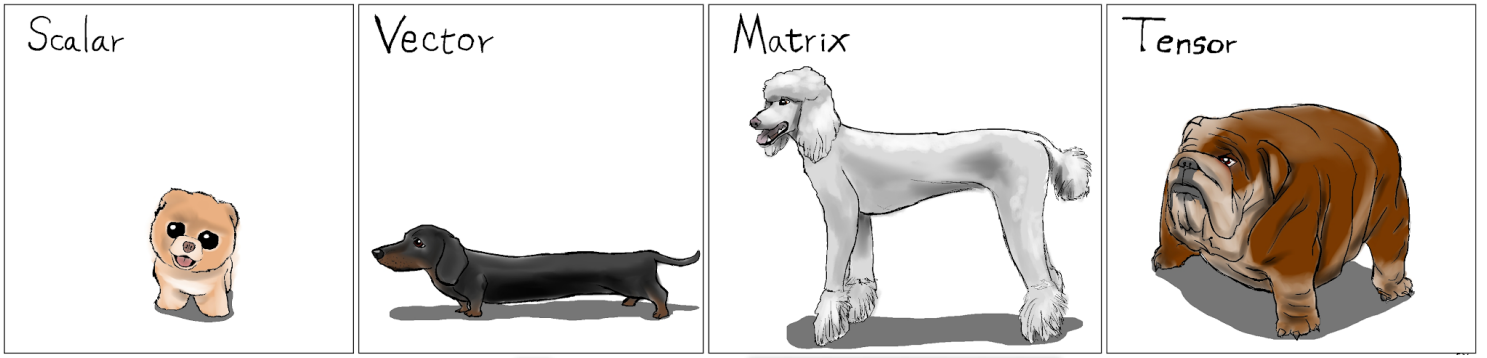
$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$
(**2**-way tensor)

...

$[a_{ijk}]_{ijk} \in \mathbb{R}^{I \times J \times K}$
(**3**-way tensor)

Tensors

- Matrices and vectors can be generalized to **Tensors**



$a \in \mathbb{R}$
(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$
(**1**-way tensor)

rows
columns

$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$
(**2**-way tensor)

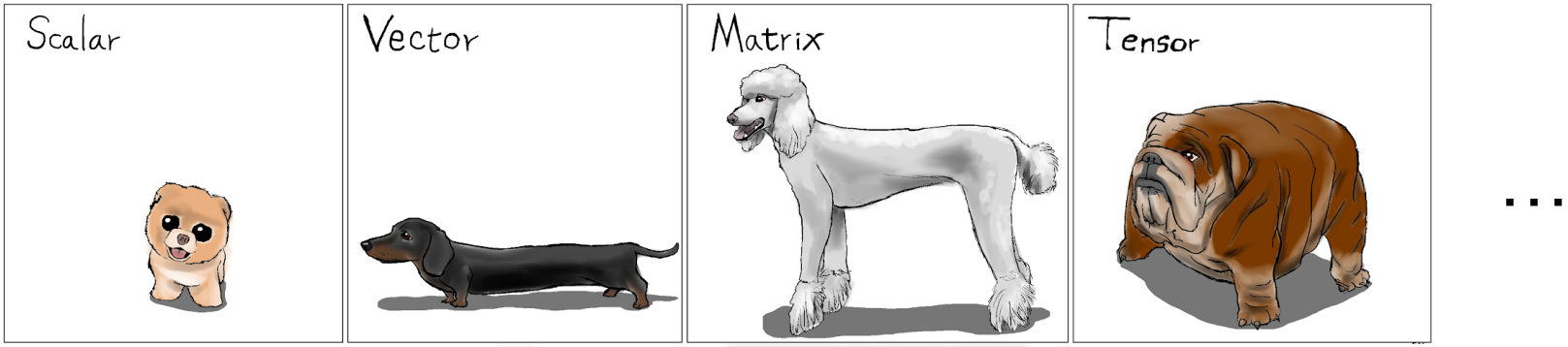
$[a_{ijk}]_{ijk} \in \mathbb{R}^{I \times J \times K}$
(**3**-way tensor)

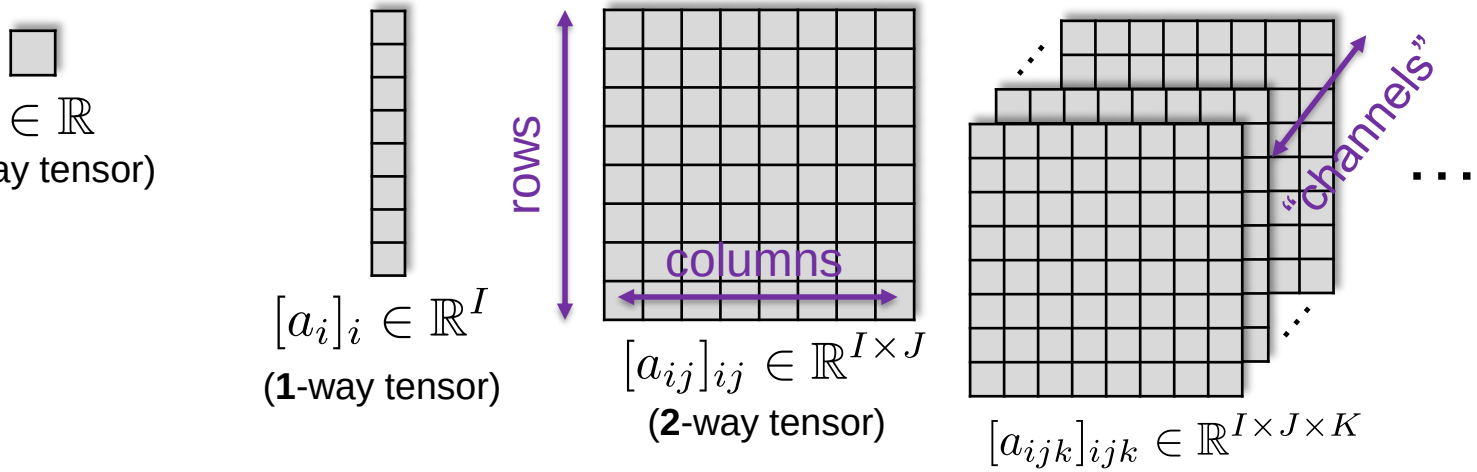
...

...

Tensors

- Matrices and vectors can be generalized to **Tensors**





$a \in \mathbb{R}$
(**0**-way tensor)

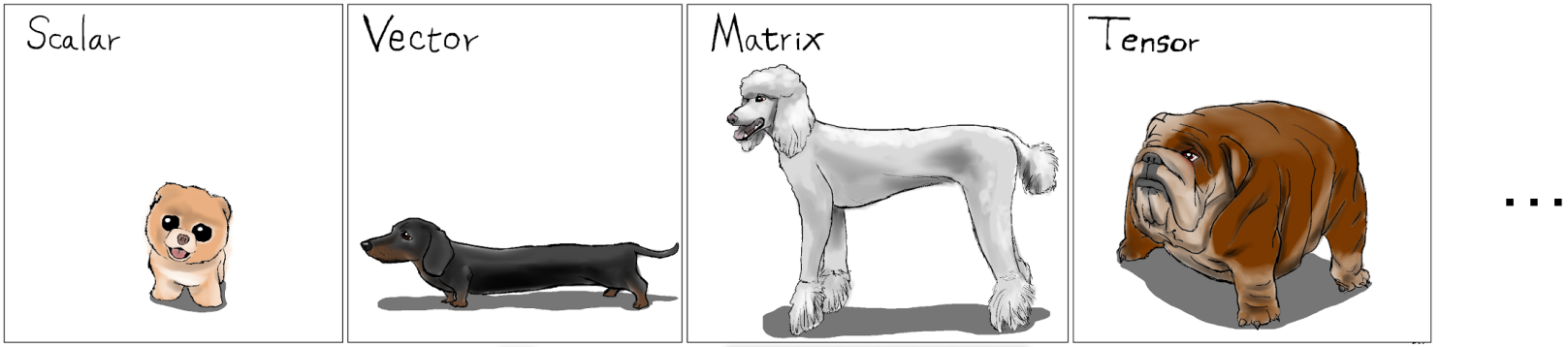
$[a_i]_i \in \mathbb{R}^I$
(**1**-way tensor)

$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$
(**2**-way tensor)

$[a_{ijk}]_{ijk} \in \mathbb{R}^{I \times J \times K}$
(**3**-way tensor)

Tensors

- Matrices and vectors can be generalized to **Tensors**



$a \in \mathbb{R}$
(**0**-way tensor)

$[a_i]_i \in \mathbb{R}^I$
(**1**-way tensor)

rows
columns

$[a_{ij}]_{ij} \in \mathbb{R}^{I \times J}$
(**2**-way tensor)

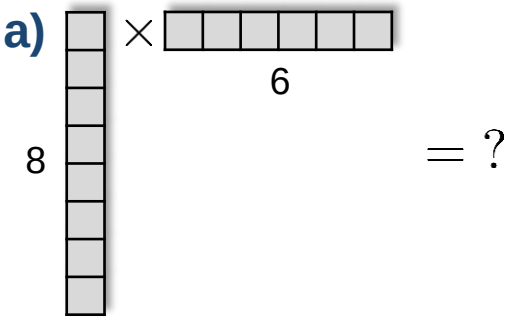
"channels"
...

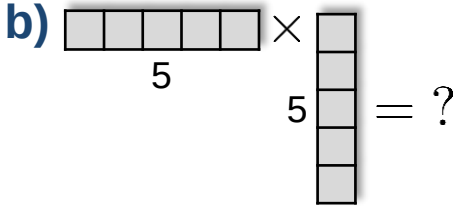
$[a_{ijk}]_{ijk} \in \mathbb{R}^{I \times J \times K}$
(**3**-way tensor)

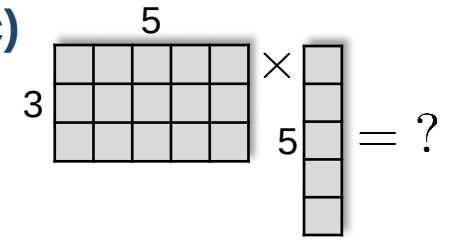
Tensors are the ubiquitous way to represent data in modern deep learning frameworks

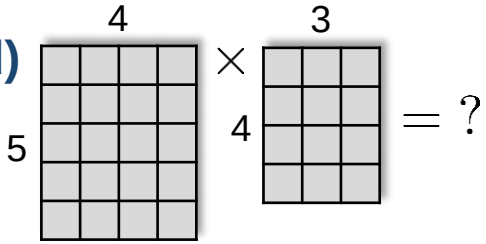
Tensors

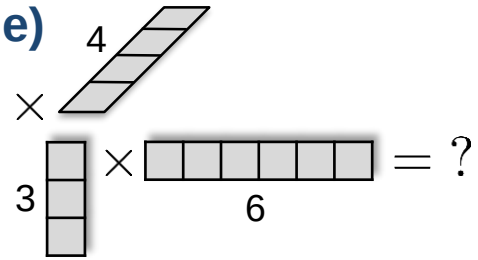
- Some vector/matrix/tensor operations:

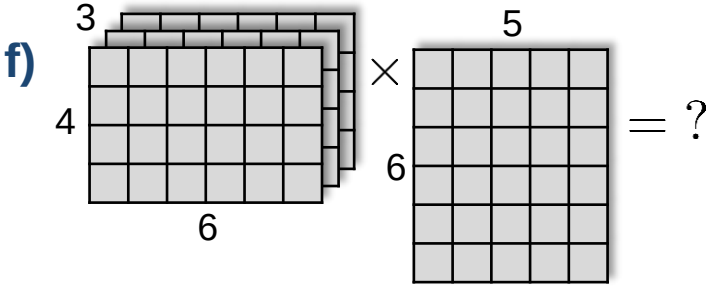
a) 

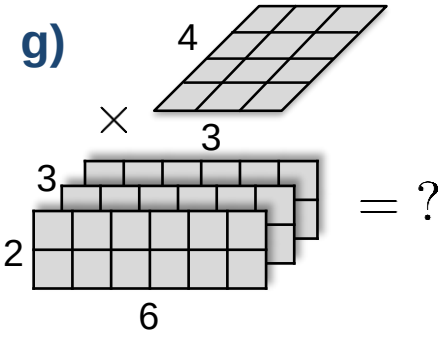
b) 

c) 

d) 

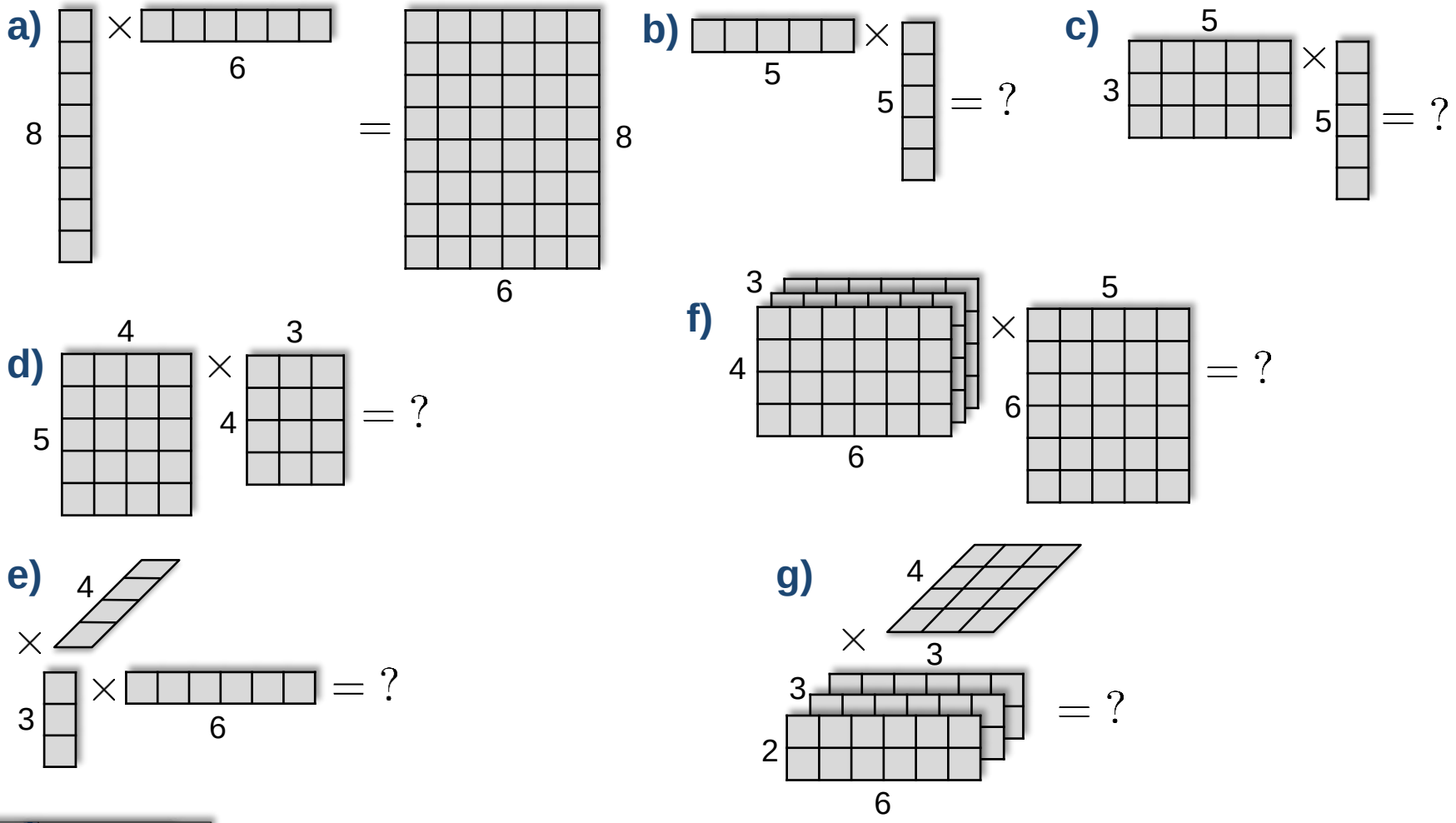
e) 

f) 

g) 

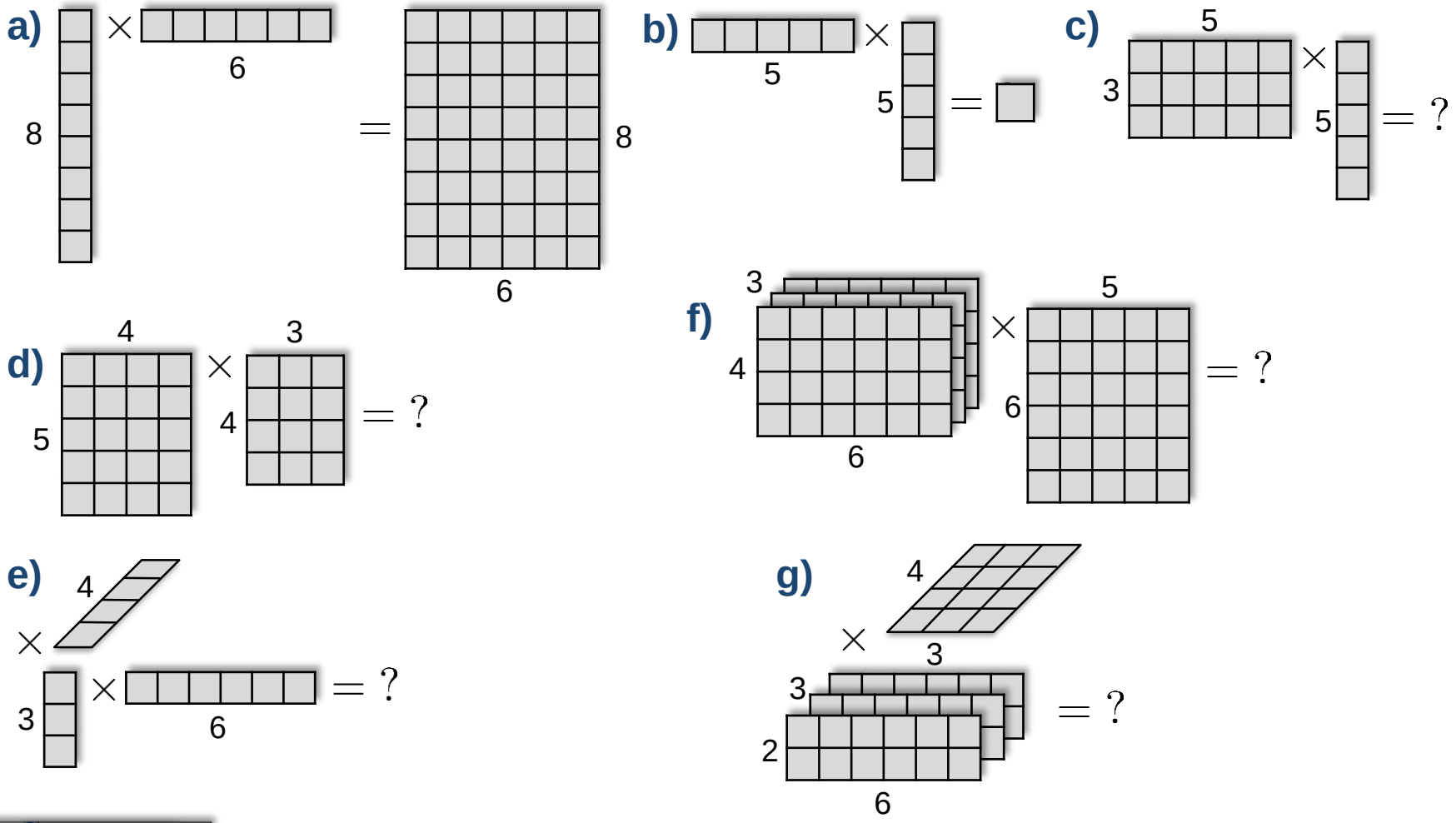
Tensors

- Some vector/matrix/tensor operations:



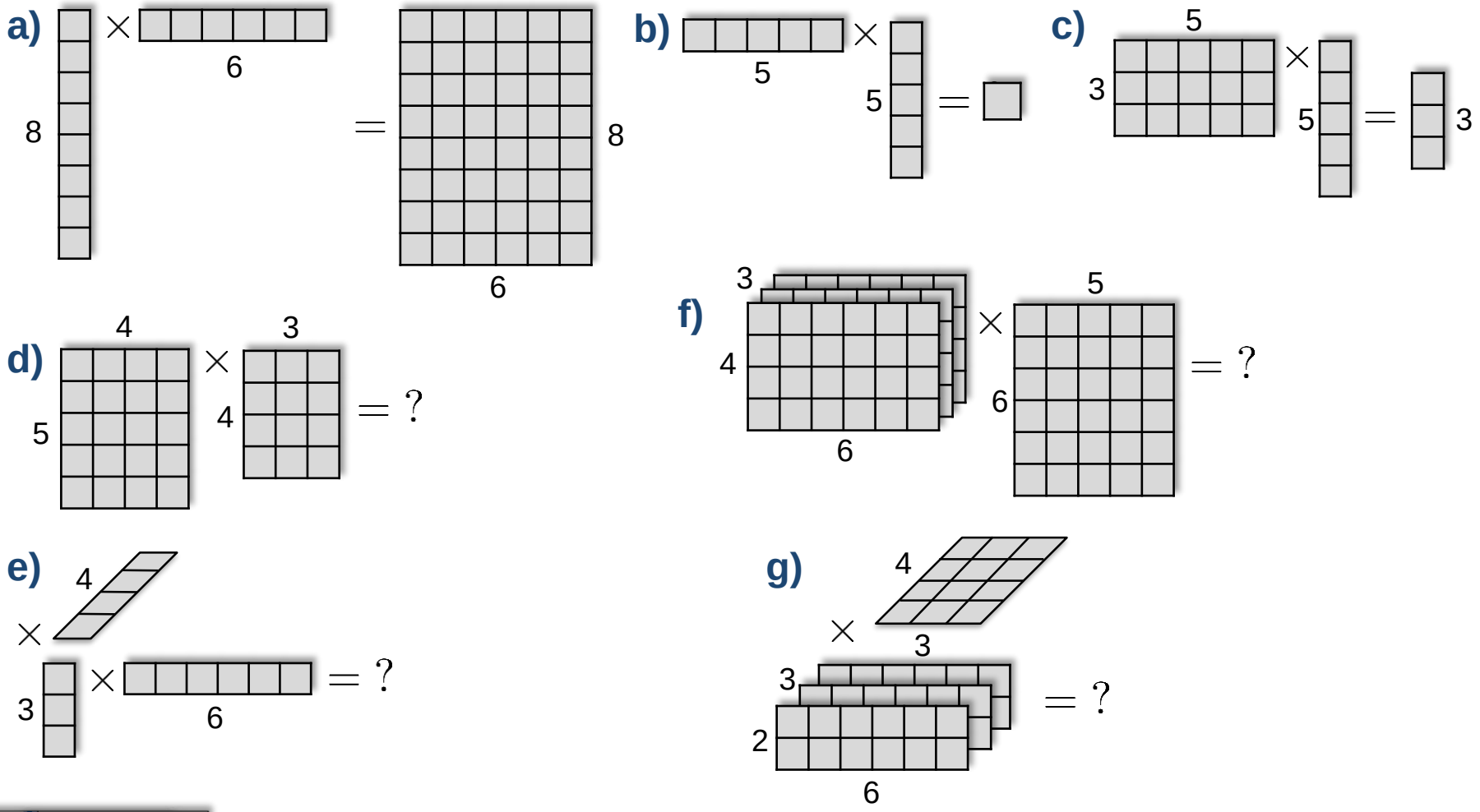
Tensors

- Some vector/matrix/tensor operations:



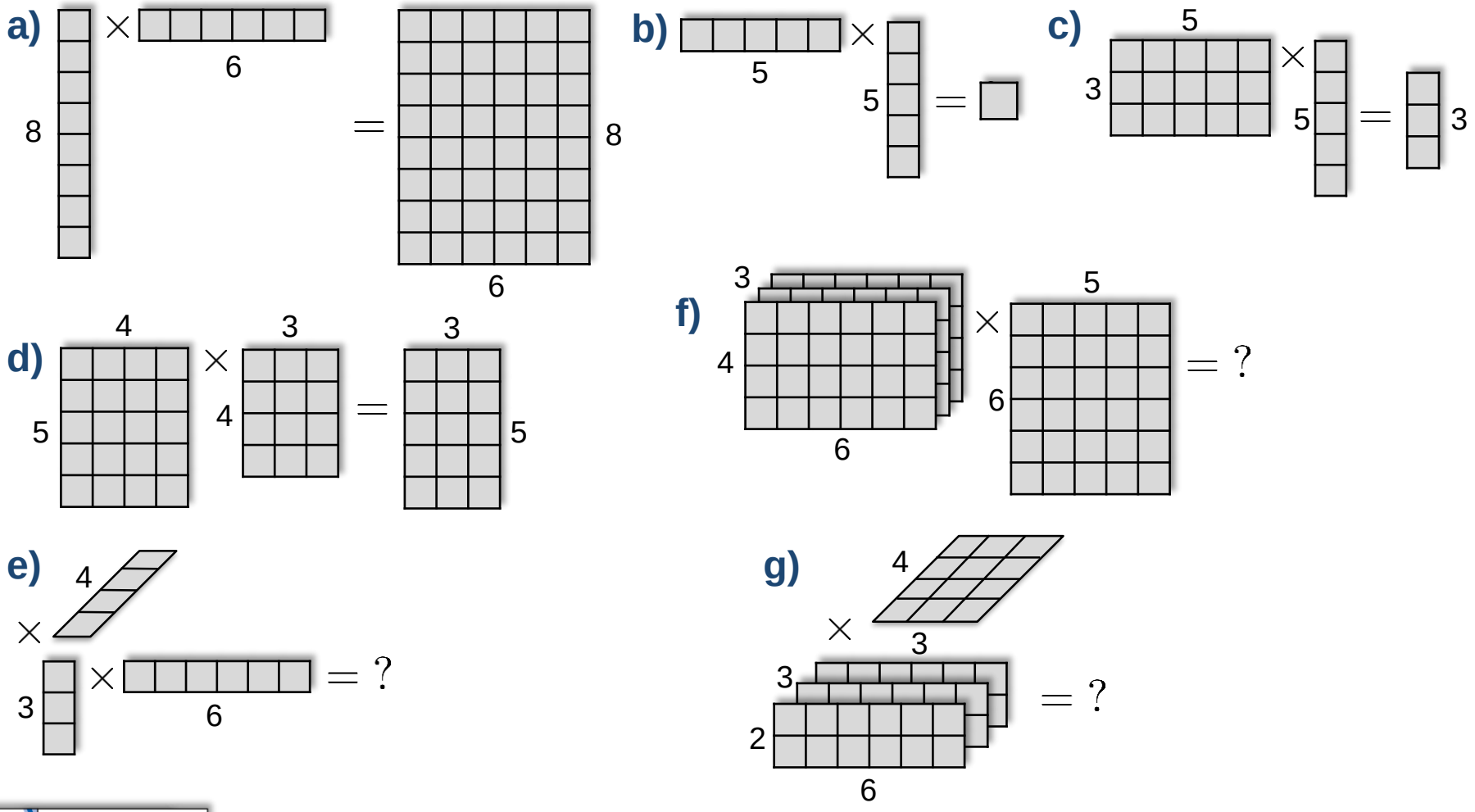
Tensors

- Some vector/matrix/tensor operations:



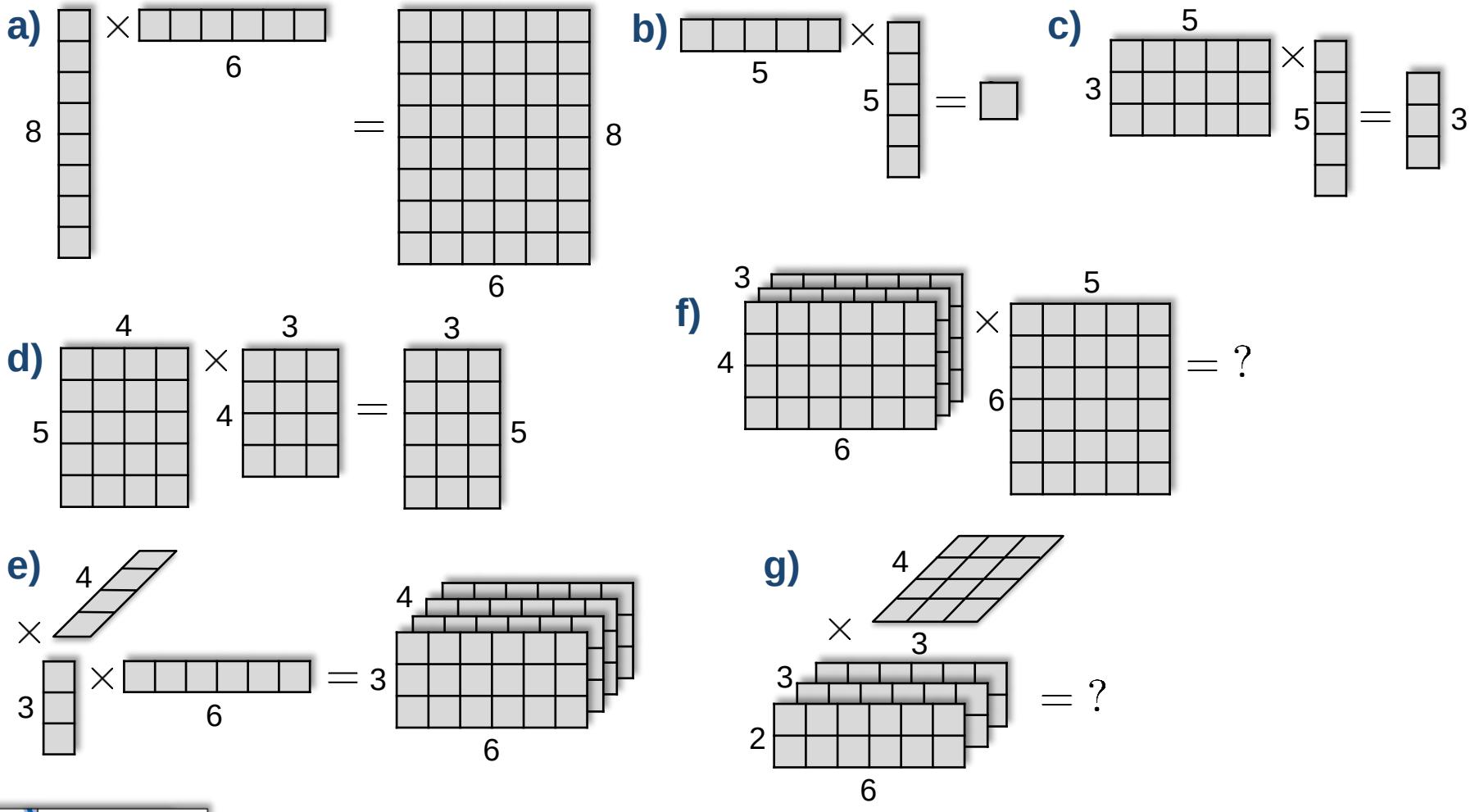
Tensors

- Some vector/matrix/tensor operations:



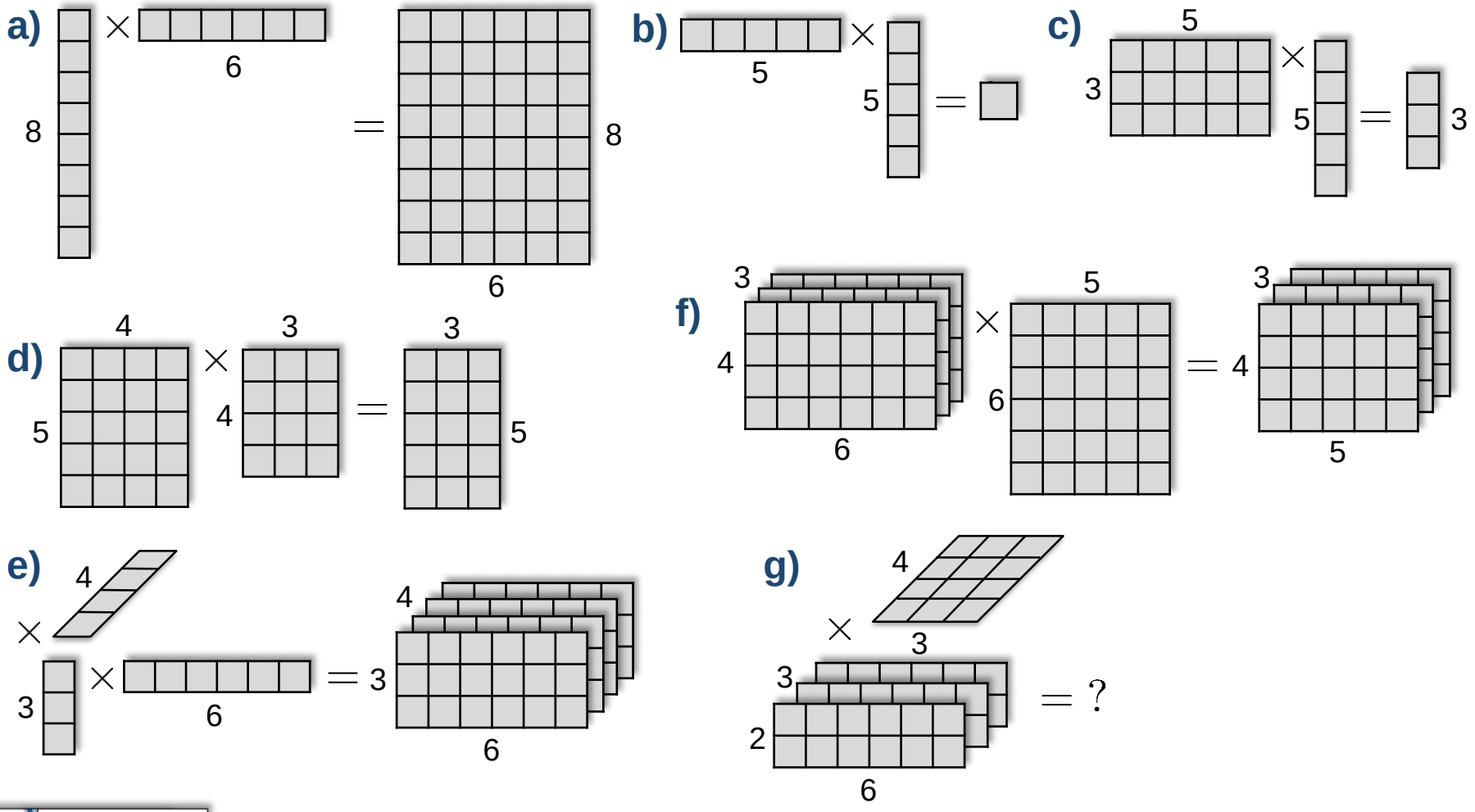
Tensors

- Some vector/matrix/tensor operations:



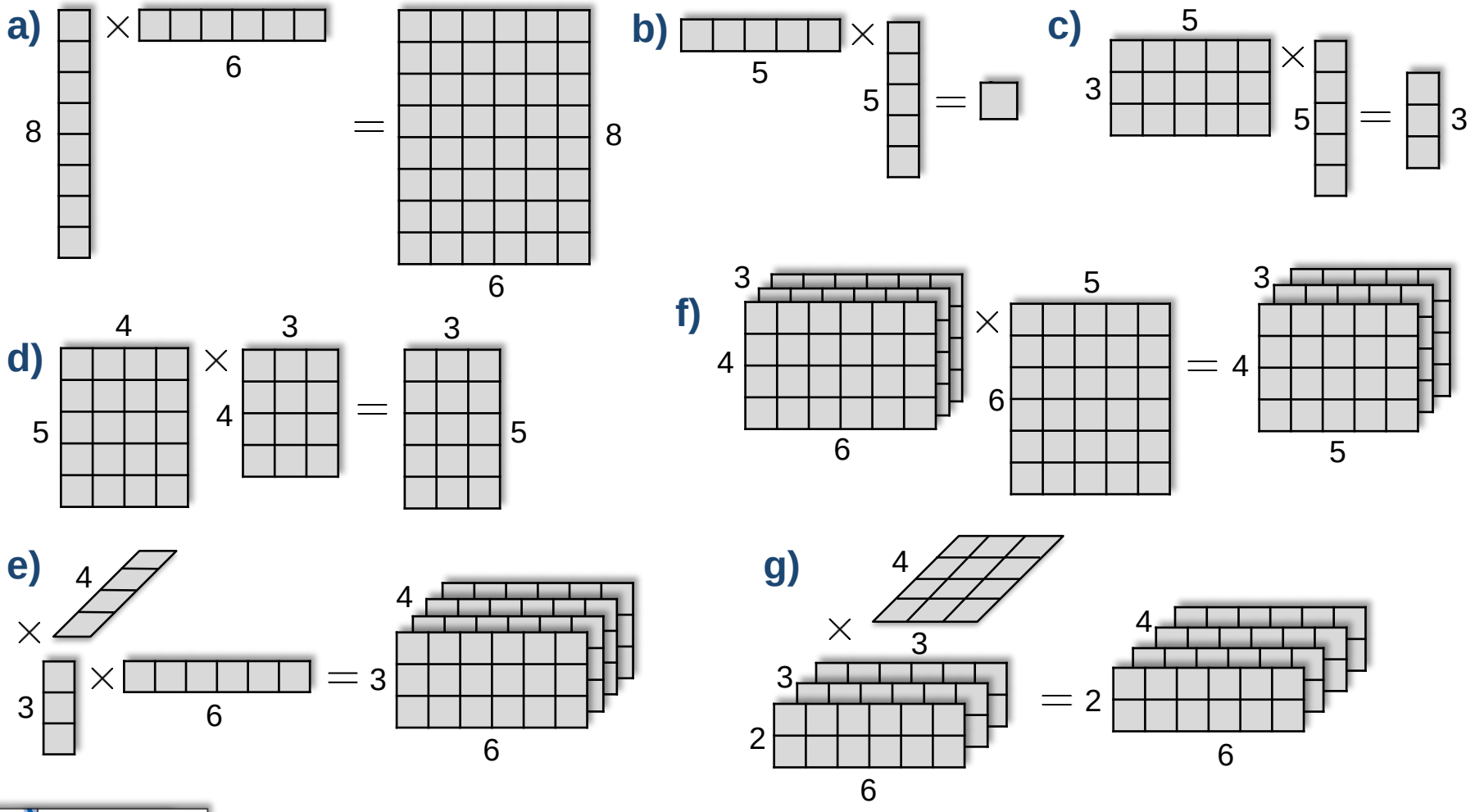
Tensors

- Some vector/matrix/tensor operations:



Tensors

- Some vector/matrix/tensor operations:



Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \mathbb{R}^D, \quad \boldsymbol{y} \in \mathbb{R}^N.$ *How can we **define** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$?*

Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^D,\ \boldsymbol{y} \in \mathbb{R}^N.$ *How can we* ***define*** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$ *?*
- Let $\delta_{\boldsymbol{x}_0} : \left\{ \begin{array}{ccc} \mathbb{R}^D & \rightarrow & \mathbb{R}^N \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{array} \right.$

"the variation of f when moving by a step $\boldsymbol{h}$ from $\boldsymbol{x}_0$."

Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^D,\ \boldsymbol{y} \in \mathbb{R}^N$. *How can we **define** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$?*
- Let $\delta_{\boldsymbol{x}_0} : \left\{ \begin{array}{ccc} \mathbb{R}^D & \rightarrow & \mathbb{R}^N \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{array} \right.$ *"the variation of f when moving by a step $\boldsymbol{h}$ from $\boldsymbol{x}_0$."*
- $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is the **linear approx.** of $\delta_{\boldsymbol{x}_0}$ for $\boldsymbol{h}$ ***infinitesimally*** *small*.

Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^D,\ \boldsymbol{y} \in \mathbb{R}^N$. *How can we* ***define*** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$ *?*
- Let $\delta_{\boldsymbol{x}_0} : \left\{ \begin{array}{ccc} \mathbb{R}^D & \rightarrow & \mathbb{R}^N \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{array} \right.$ *"the variation of f when moving by a step $\boldsymbol{h}$ from $\boldsymbol{x}_0$."*
- $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is the **linear approx.** of $\delta_{\boldsymbol{x}_0}$ for $\boldsymbol{h}$ ***infinitesimally*** *small*.
- We call this the ***total derivative*** of f, or of $\boldsymbol{y}$, with resp. to $\boldsymbol{x}$, at $\boldsymbol{x}_0$.

Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^D,\ \boldsymbol{y} \in \mathbb{R}^N$. *How can we* ***define*** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$ *?*
- Let $\delta_{\boldsymbol{x}_0} : \left\{ \begin{array}{ccc} \mathbb{R}^D & \rightarrow & \mathbb{R}^N \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{array} \right.$ *"the variation of f when moving by a step $\boldsymbol{h}$ from $\boldsymbol{x}_0$."*
- $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is the **linear approx.** of $\delta_{\boldsymbol{x}_0}$ for $\boldsymbol{h}$ ***infinitesimally*** *small*.
- We call this the ***total derivative*** of f, or of $\boldsymbol{y}$, with resp. to $\boldsymbol{x}$, at $\boldsymbol{x}_0$.
- Since $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}$ is a **linear map** from $\mathbb{R}^D$ to $\mathbb{R}^N$, it can be *represented* by a **matrix**, $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \mathbf{J}_{\boldsymbol{x}}[f](\boldsymbol{x}_0) \in \mathbb{R}^{N \times D}$, called the ***Jacobian***.

Generalizing the derivative

- Let $\boldsymbol{y} = f(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^D,\ \boldsymbol{y} \in \mathbb{R}^N$. *How can we* ***define*** $\frac{d\boldsymbol{y}}{d\boldsymbol{x}}$ *?*
- Let $\delta_{\boldsymbol{x}_0} : \left\{ \begin{array}{ccc} \mathbb{R}^D & \to & \mathbb{R}^N \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{array} \right.$ *"the variation of f when moving by a step $\boldsymbol{h}$ from $\boldsymbol{x}_0$."*
- $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}=\boldsymbol{x}_0}$ is the **linear approx.** of $\delta_{\boldsymbol{x}_0}$ for $\boldsymbol{h}$ ***infinitesimally*** *small*.
- We call this the ***total derivative*** of f, or of $\boldsymbol{y}$, with resp. to $\boldsymbol{x}$, at $\boldsymbol{x}_0$.
- Since $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}$ is a **linear map** from $\mathbb{R}^D$ to $\mathbb{R}^N$, it can be *represented* by a **matrix**, $\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \mathbf{J}_{\boldsymbol{x}}[f](\boldsymbol{x}_0) \in \mathbb{R}^{N \times D}$, called the ***Jacobian***.

Formulas:

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1(x_{1,0})}{\partial x_1} & \frac{\partial f_1(x_{2,0})}{\partial x_2} & \cdots & \frac{\partial f_1(x_{D,0})}{\partial D_1} \\ \frac{\partial f_2(x_{1,0})}{\partial x_1} & \frac{\partial f_2(x_{2,0})}{\partial x_2} & \cdots & \frac{\partial f_2(x_{D,0})}{\partial D_1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_N(x_{1,0})}{\partial x_1} & \frac{\partial f_N(x_{2,0})}{\partial x_2} & \cdots & \frac{\partial f_N(x_{D,0})}{\partial D_1} \end{bmatrix} = \mathbf{J}_{\boldsymbol{x}}[f](\boldsymbol{x}_0)$$

Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \rightarrow \mathbb{R}^1$?

Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

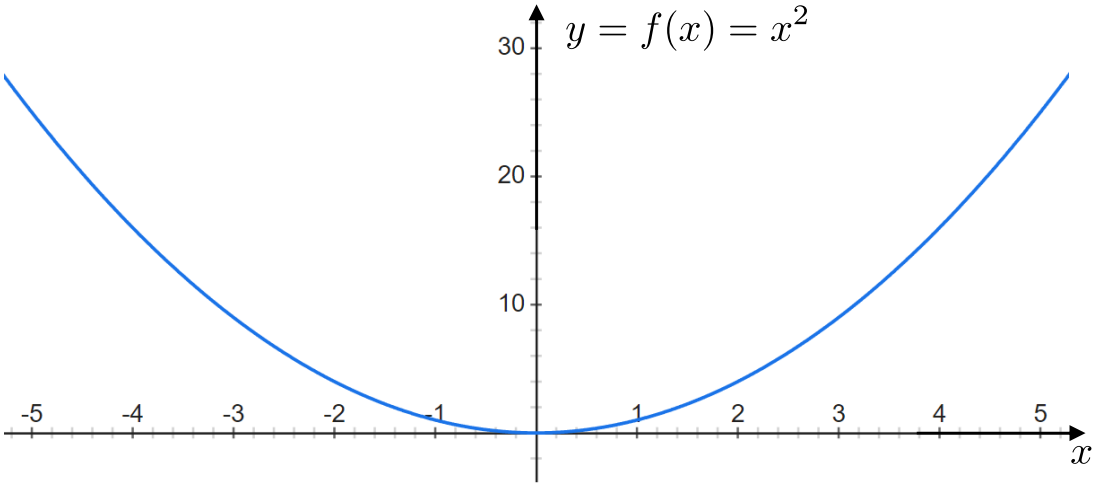
Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$



Special Case 1

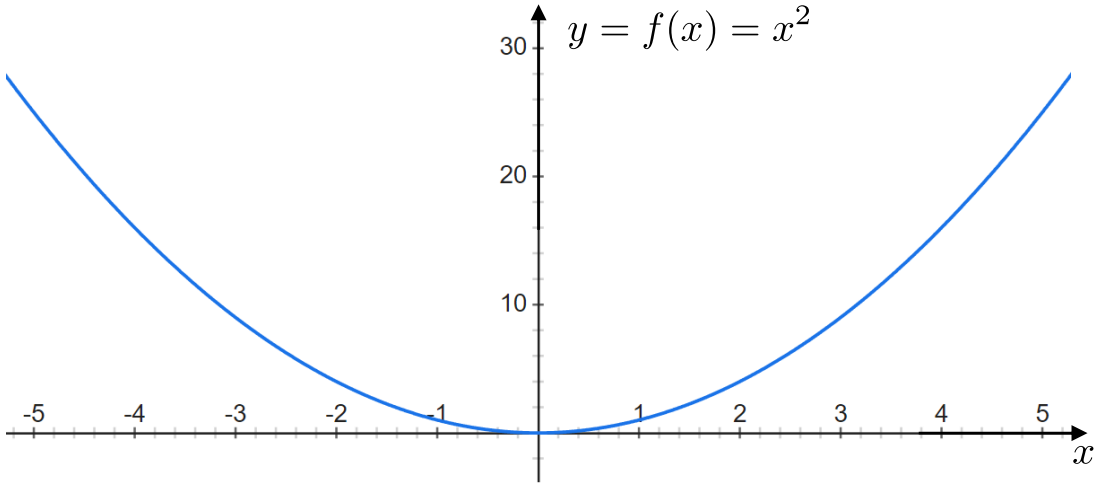
What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left.\frac{dy}{dx}\right|_{x_0}$$



Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

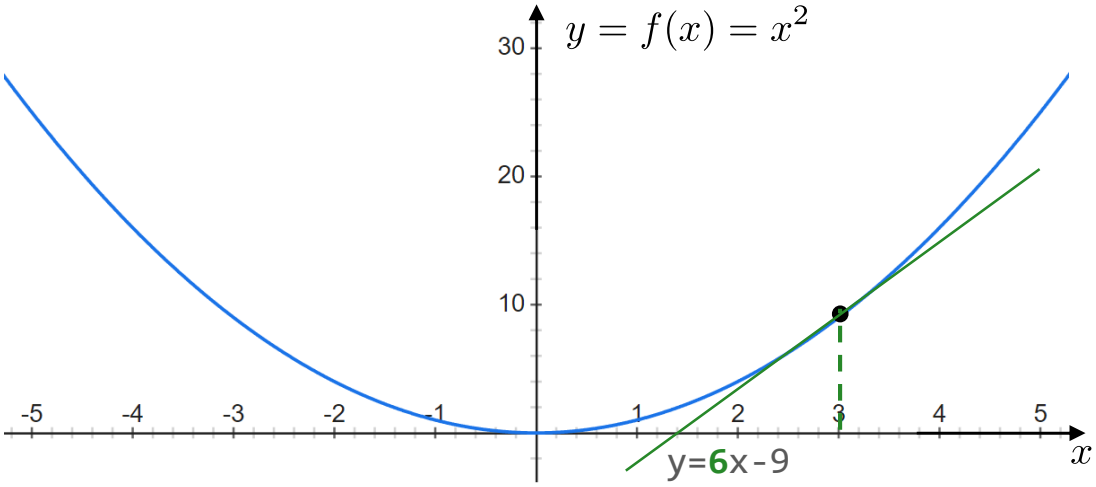
- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left. \frac{dy}{dx} \right|_{x_0}$$

$$f'(3) = 6$$



Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

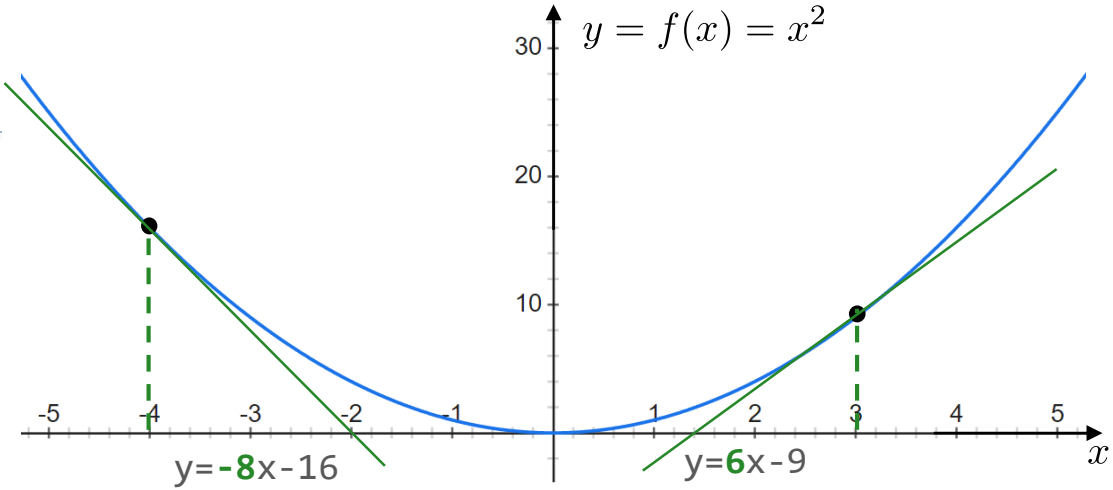
- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left. \frac{dy}{dx} \right|_{x_0}$$

$$f'(3) = 6, \; f'(-4) = -8$$



Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

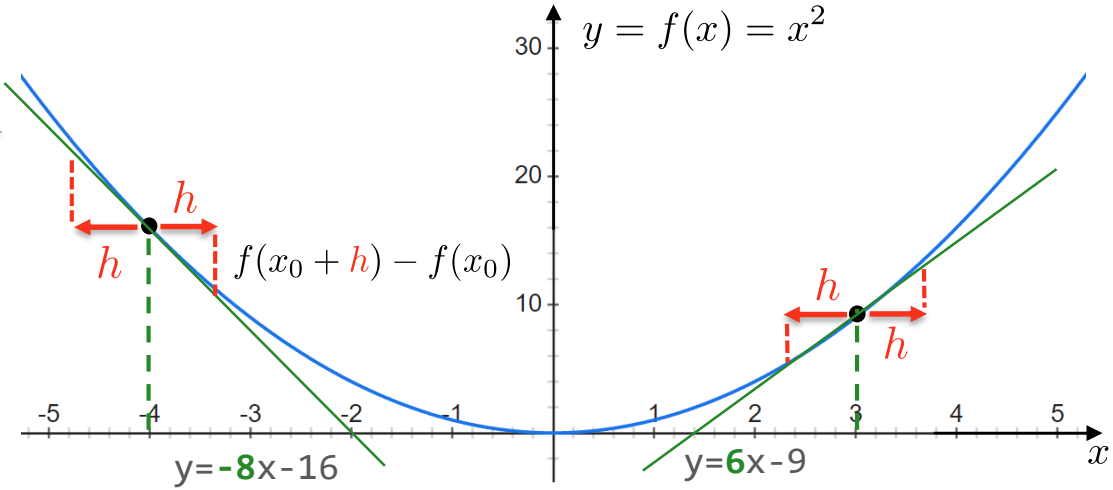
- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left.\frac{dy}{dx}\right|_{x_0}$$

$$f'(3) = 6, \; f'(-4) = -8$$



- The derivative of f at x_0 may be viewed as the **linear map** that approximates

$$\delta_{x_0} : \begin{cases} \mathbb{R}^1 & \to & \mathbb{R}^1 \\ h & \mapsto & f(x_0 + h) - f(x_0) \end{cases}$$, for h ***infinitesimally*** small.

"the variation of f when moving to the left or to the right from x_0."

Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

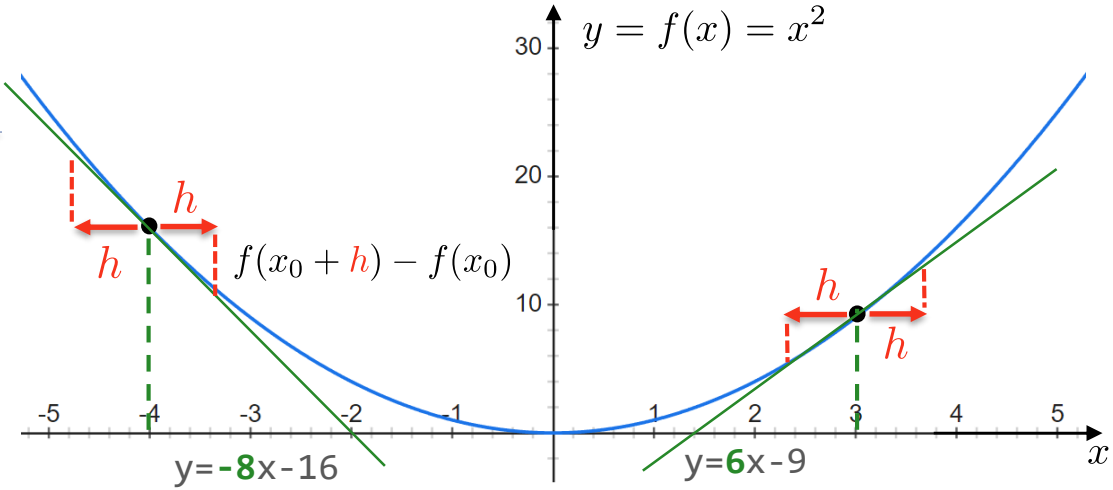
- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left.\frac{dy}{dx}\right|_{x_0}$$

$$f'(3) = 6, \; f'(-4) = -8$$



- The derivative of f at x_0 may be viewed as the **linear map** that approximates

$$\delta_{x_0} : \begin{cases} \mathbb{R}^1 & \to & \mathbb{R}^1 \\ h & \mapsto & f(x_0 + h) - f(x_0) \end{cases}$$, for h ***infinitesimally*** small.

"the variation of f when moving to the left or to the right from x_0."

Indeed: $\lim_{h \to \infty} f(x_0 + h) - f(x_0) = f'(x_0)h$!

Special Case 1

What does it give for a simple function $f : \mathbb{R}^1 \to \mathbb{R}^1$?

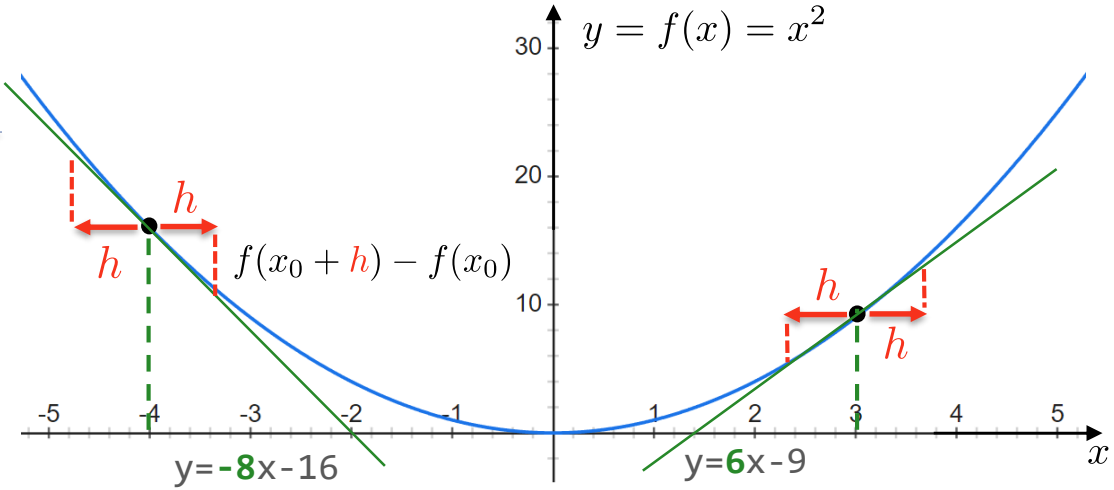
- In high-school we learn:

$$f'(x_0) = \lim_{h \to \infty} \frac{f(x_0 + h) - f(x_0)}{h}$$

- **Ex.** $y = f(x) = x^2$

$$f'(x_0) = 2x_0 = \left.\frac{dy}{dx}\right|_{x_0}$$

$$f'(3) = 6, \; f'(-4) = -8$$



- The derivative of f at x_0 may be viewed as the **linear map** that approximates

$$\delta_{x_0} : \begin{cases} \mathbb{R}^1 & \to & \mathbb{R}^1 \\ h & \mapsto & f(x_0 + h) - f(x_0) \end{cases}$$, for h ***infinitesimally*** small.

"the variation of f when moving to the left or to the right from x_0."

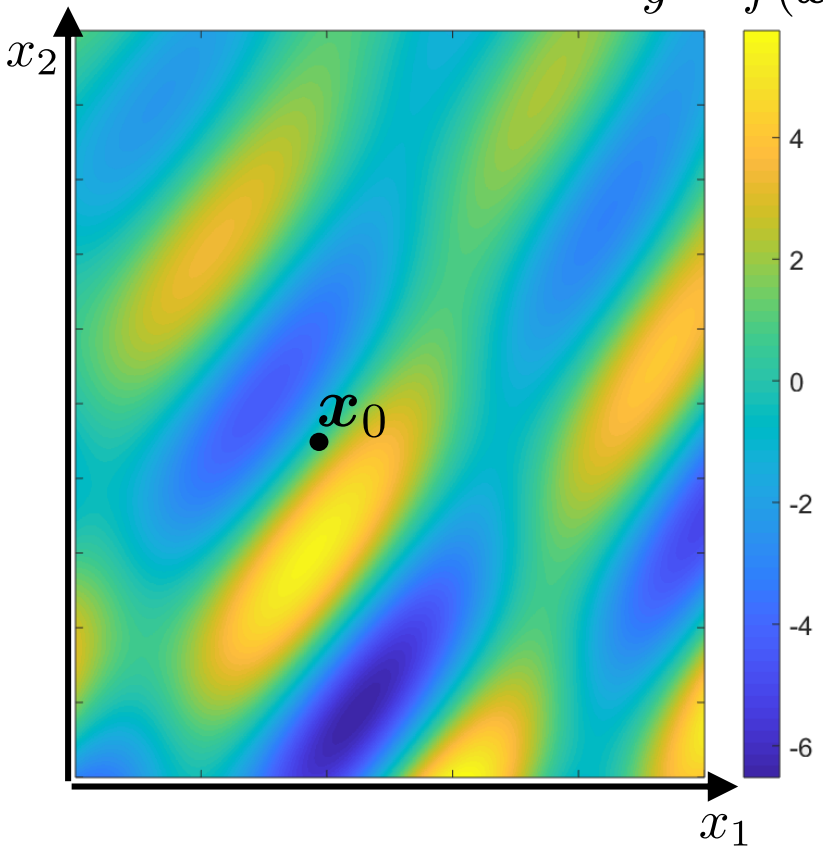
Indeed: $\lim_{h \to \infty} f(x_0 + h) - f(x_0) = f'(x_0)h$!

- As a **linear map**, $f'(x_0)$ can be seen as a **0-way tensor**, in $\mathbb{R}^{1 \times 1}$!

Special Case 2

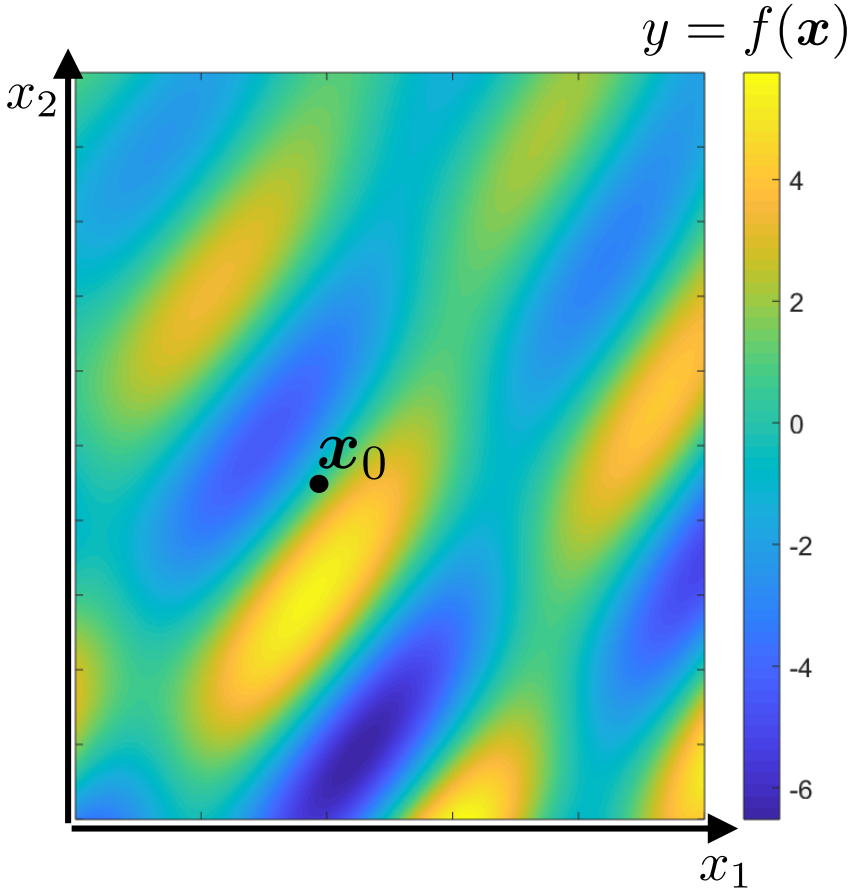
What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?

$$y = f(\boldsymbol{x})$$



Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



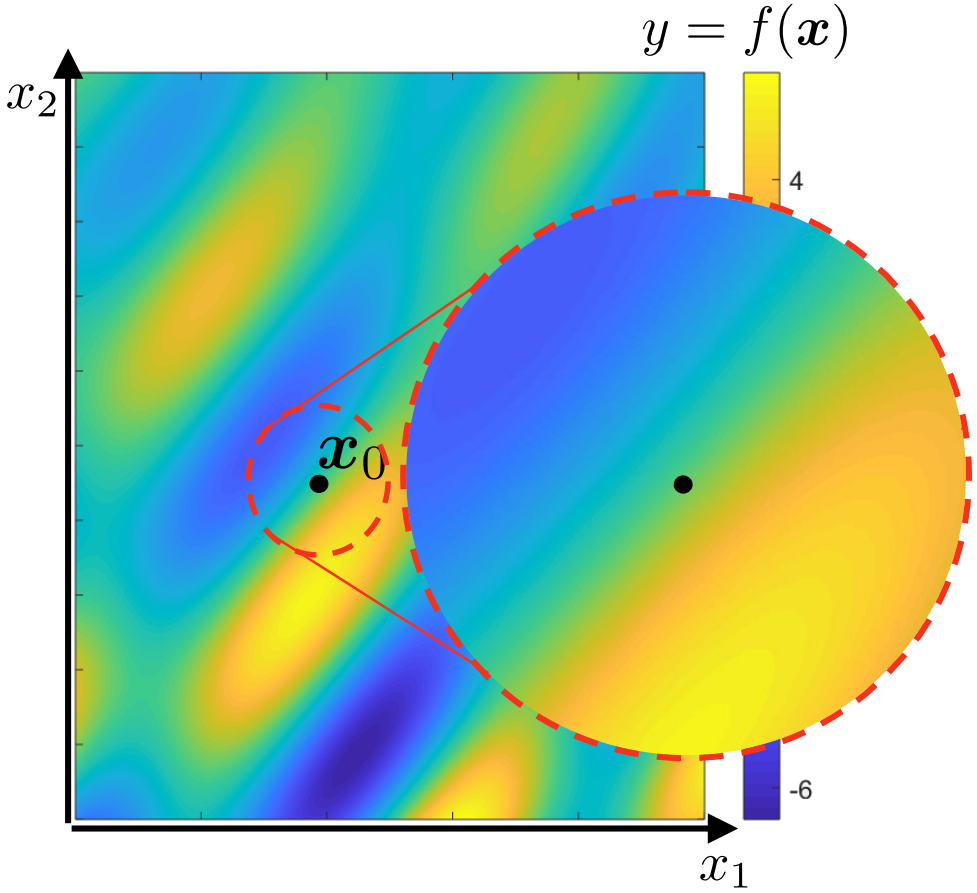
- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



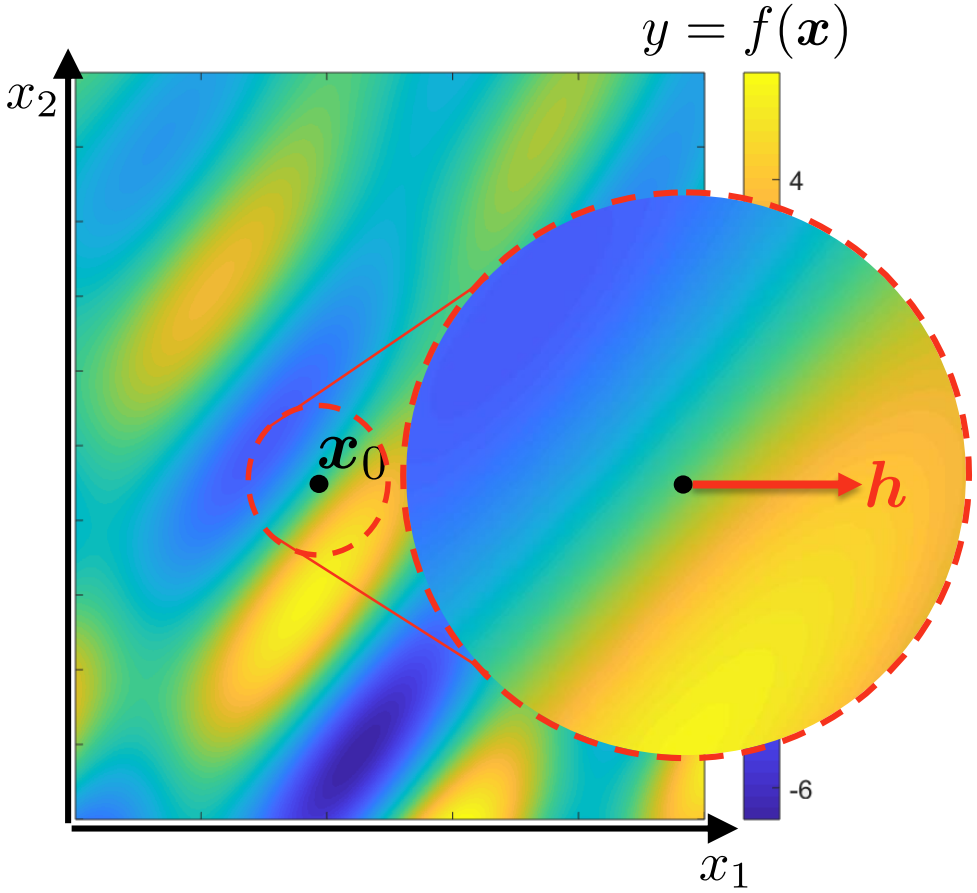
- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \to \mathbb{R}^1$?



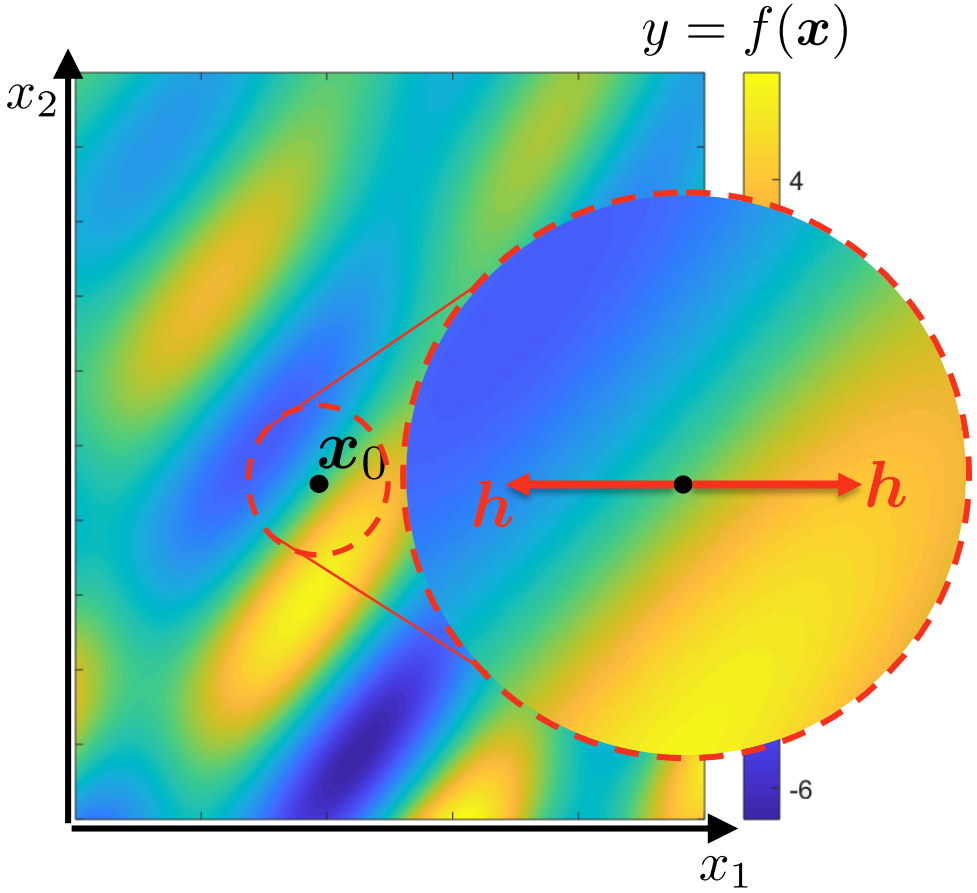
- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \to & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



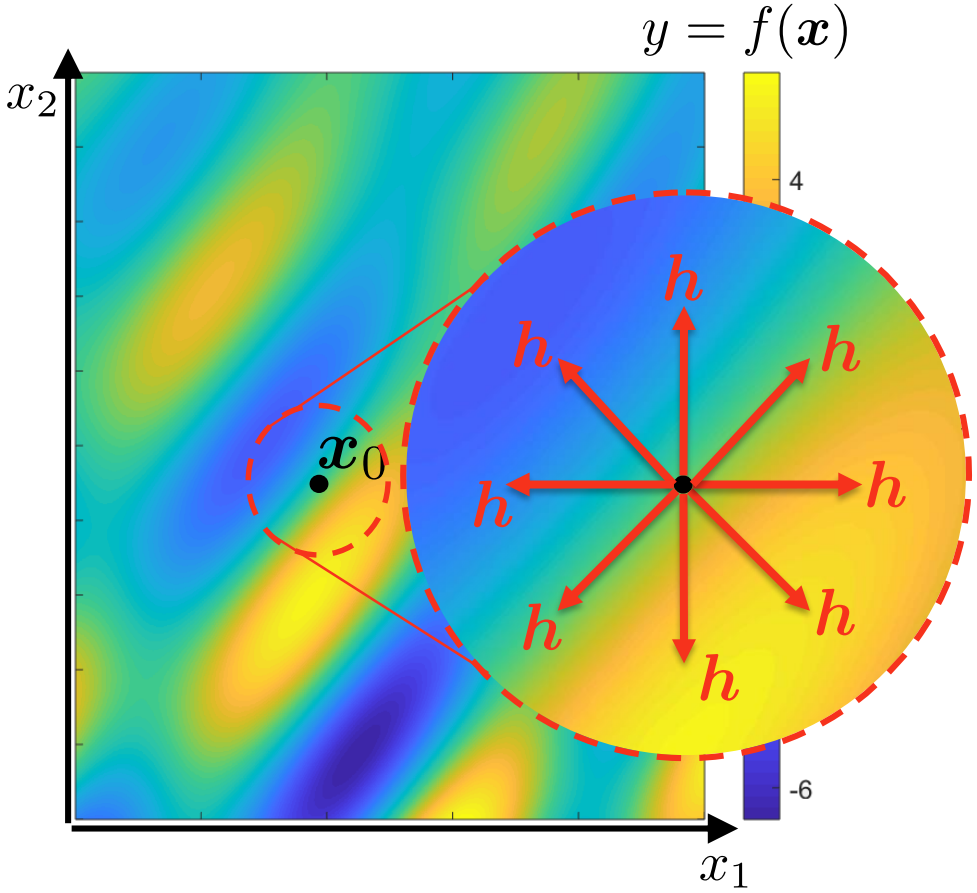
- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



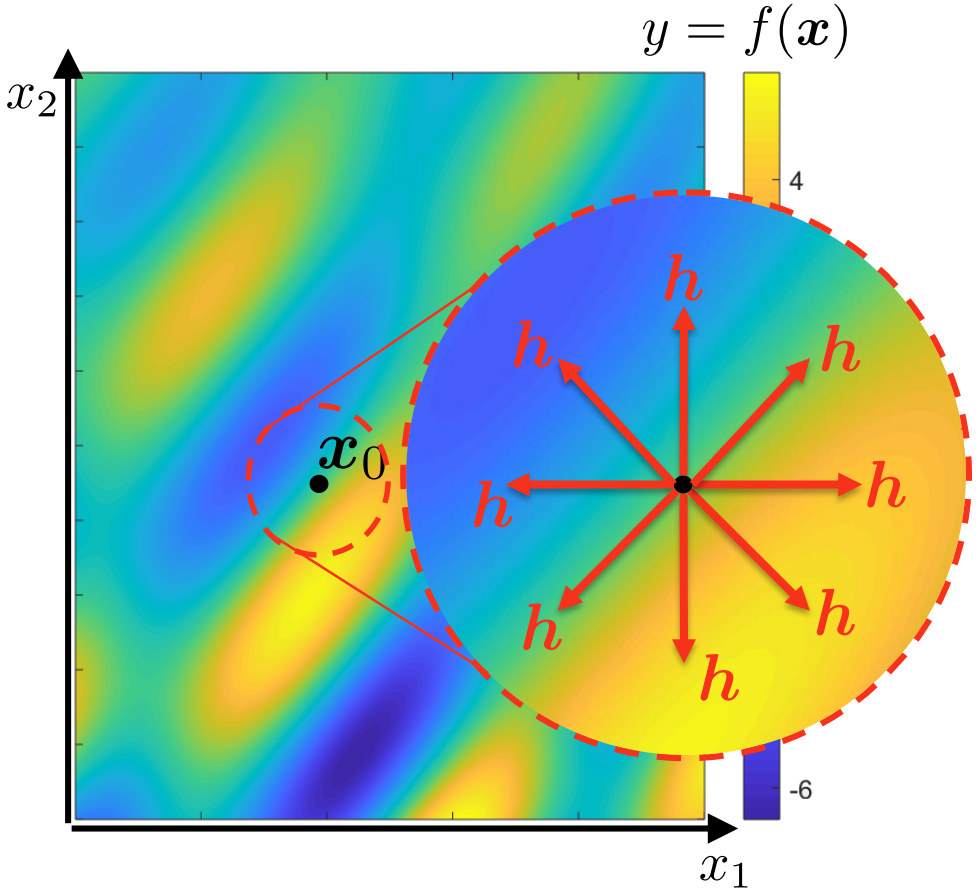
- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \to \mathbb{R}^1$?



- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

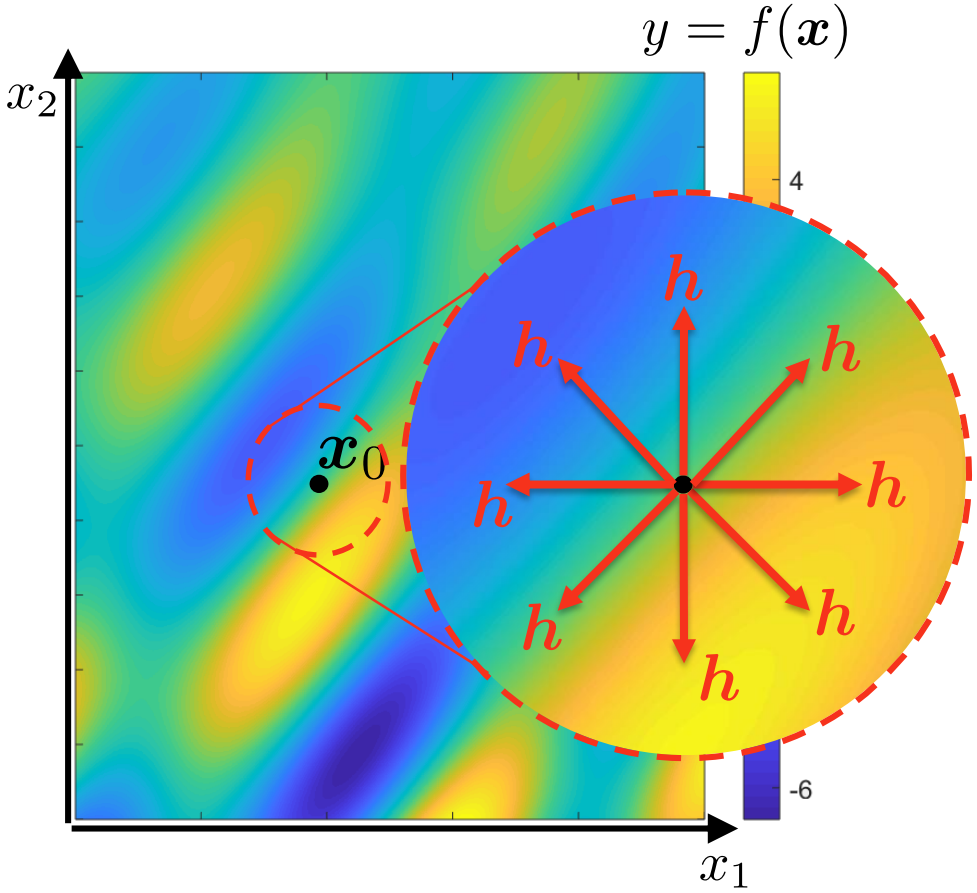
$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \to & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

  for $\boldsymbol{h}$ infinitesimally small.

- As a linear form, it can be represented by a **row vector**: $\delta_{\boldsymbol{x}_0}(\boldsymbol{h}) \approx \boldsymbol{w}^\top \boldsymbol{h}$

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

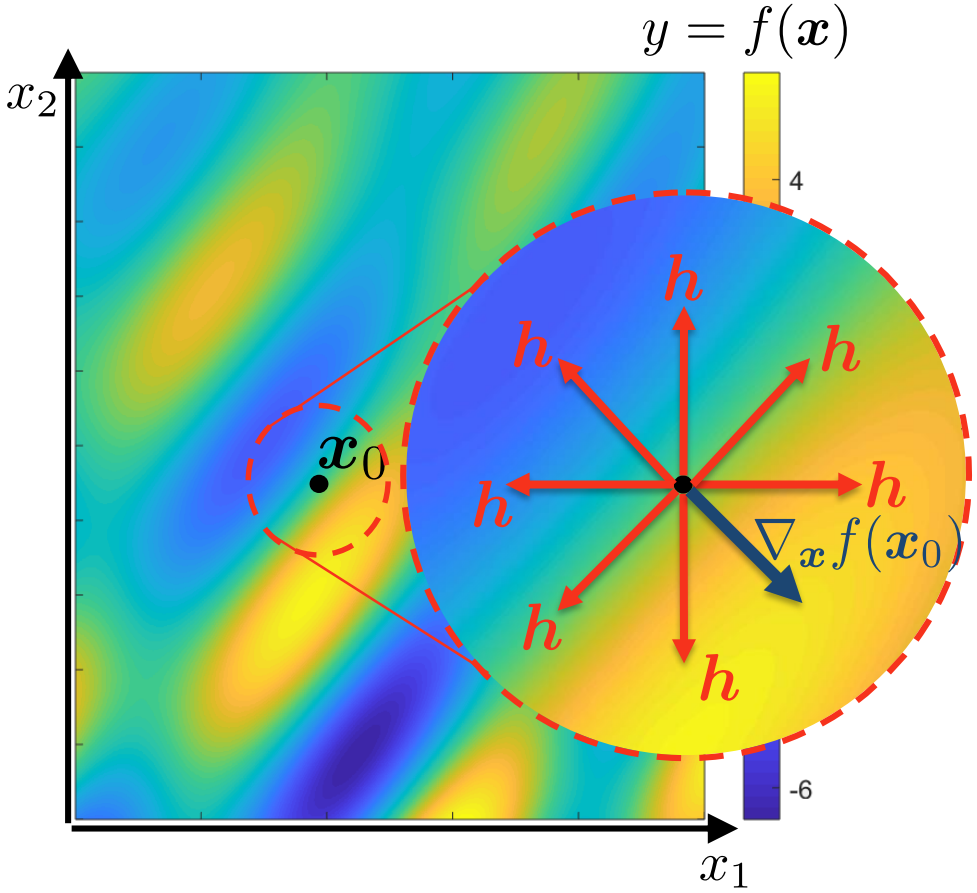
  for $\boldsymbol{h}$ infinitesimally small.

- As a linear form, it can be represented by a **row vector**: $\delta_{\boldsymbol{x}_0}(\boldsymbol{h}) \approx \boldsymbol{w}^\top \boldsymbol{h}$
- The **transpose** of this vector is called the ***gradient*** of f at $\boldsymbol{x}_0$ denoted:

$$\boldsymbol{w} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) \in \mathbb{R}^D$$

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

  for $\boldsymbol{h}$ infinitesimally small.

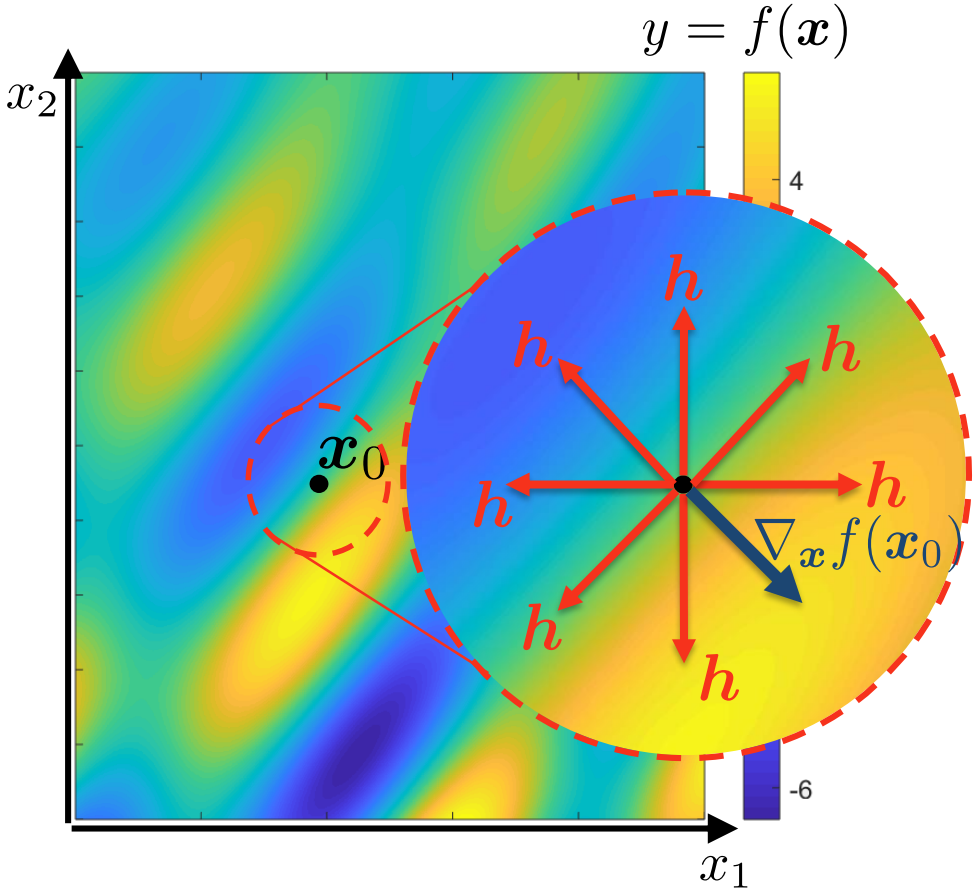
- As a linear form, it can be represented by a **row vector**: $\delta_{\boldsymbol{x}_0}(\boldsymbol{h}) \approx \boldsymbol{w}^\top \boldsymbol{h}$
- The **transpose** of this vector is called the ***gradient*** of f at $\boldsymbol{x}_0$ denoted:

$$\boldsymbol{w} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) \in \mathbb{R}^D$$

- It can be interpreted as the ***direction and rate of fastest increase*** of f at $\boldsymbol{x}_0$

Special Case 2

What about a **real-valued** multivariate function $f : \mathbb{R}^D \rightarrow \mathbb{R}^1$?



- The ***total derivative*** of f at $\boldsymbol{x}_0$ is the **linear form** that approximates

$$\delta_{\boldsymbol{x}_0} : \begin{cases} \mathbb{R}^D & \rightarrow & \mathbb{R}^1 \\ \boldsymbol{h} & \mapsto & f(\boldsymbol{x}_0 + \boldsymbol{h}) - f(\boldsymbol{x}_0) \end{cases}$$

for $\boldsymbol{h}$ infinitesimally small.

- As a linear form, it can be represented by a **row vector**: $\delta_{\boldsymbol{x}_0}(\boldsymbol{h}) \approx \boldsymbol{w}^\top \boldsymbol{h}$
- The **transpose** of this vector is called the ***gradient*** of f at $\boldsymbol{x}_0$ denoted:

$$\boldsymbol{w} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) \in \mathbb{R}^D$$

- It can be interpreted as the ***direction and rate of fastest increase*** of f at $\boldsymbol{x}_0$

- The formula is $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = \left[\left. \frac{\partial y_1}{\partial x_1} \right|_{x_{1,0}}, \ldots, \left. \frac{\partial y_1}{\partial x_D} \right|_{x_{D,0}} \right]^\top = \left. \frac{dy}{d\boldsymbol{x}} \right|_{\boldsymbol{x}_0}^\top$

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \ldots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \dots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$
$\mathbb{R}^D$	$\mathbb{R}$	

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \dots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$
$\mathbb{R}^D$	$\mathbb{R}$	$\left.\frac{dy}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top \in \mathbb{R}^{1\times D}$

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \dots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$
$\mathbb{R}^D$	$\mathbb{R}$	$\left.\frac{dy}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top \in \mathbb{R}^{1\times D}$
$\mathbb{R}^D$	$\mathbb{R}^N$	

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \dots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$
$\mathbb{R}^D$	$\mathbb{R}$	$\left.\frac{dy}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top \in \mathbb{R}^{1\times D}$
$\mathbb{R}^D$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \mathbf{J}_{\boldsymbol{x}}[f](\boldsymbol{x}_0) \in \mathbb{R}^{N\times D}$

Summary (for $y = f(x)$)

Domain of x	Domain of y	Total Derivative at x_0
$\mathbb{R}$	$\mathbb{R}$	$\left.\frac{dy}{dx}\right\|_{x_0} = f'(x_0) \in \mathbb{R}^{1\times 1}$
$\mathbb{R}$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{dx}\right\|_{x_0} = [f'_1(x_0), \dots, f'_N(x_0)]^\top \in \mathbb{R}^{N\times 1}$
$\mathbb{R}^D$	$\mathbb{R}$	$\left.\frac{dy}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top \in \mathbb{R}^{1\times D}$
$\mathbb{R}^D$	$\mathbb{R}^N$	$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right\|_{\boldsymbol{x}_0} = \mathbf{J}_{\boldsymbol{x}}[f](\boldsymbol{x}_0) \in \mathbb{R}^{N\times D}$

Note: All of this generalizes naturally to **matrix** or **tensor variables**, yielding **tensor total derivatives**.

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

Let:
- $\boldsymbol{x} \in \mathbb{R}^D$
- $y = f(\boldsymbol{x}) \in \mathbb{R}$
- $\boldsymbol{z} = g(y) = g(f(\boldsymbol{x})) \in \mathbb{R}^N$
- $\boldsymbol{w} = h(\boldsymbol{z}) = h(g(f(\boldsymbol{x}))) \in \mathbb{R}^M$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

Let:
- $\boldsymbol{x} \in \mathbb{R}^D$
- $y = f(\boldsymbol{x}) \in \mathbb{R}$
- $\boldsymbol{z} = g(y) = g(f(\boldsymbol{x})) \in \mathbb{R}^N$
- $\boldsymbol{w} = h(\boldsymbol{z}) = h(g(f(\boldsymbol{x}))) \in \mathbb{R}^M$

$$\underset{\mathbb{R}^{M\times D}}{\left.\frac{d\boldsymbol{w}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}} = \underset{\mathbb{R}^{M\times N}}{\left.\frac{d\boldsymbol{w}}{d\boldsymbol{z}}\right|_{\boldsymbol{z}_0}} \times \underset{\mathbb{R}^{N\times 1}}{\left.\frac{d\boldsymbol{z}}{dy}\right|_{y_0}} \times \underset{\mathbb{R}^{1\times D}}{\left.\frac{dy}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}}$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

Let:
- $\boldsymbol{x} \in \mathbb{R}^D$
- $y = f(\boldsymbol{x}) \in \mathbb{R}$
- $\boldsymbol{z} = g(y) = g(f(\boldsymbol{x})) \in \mathbb{R}^N$
- $\boldsymbol{w} = h(\boldsymbol{z}) = h(g(f(\boldsymbol{x}))) \in \mathbb{R}^M$

$$\underset{\mathbb{R}^{M\times D}}{\left.\frac{d\boldsymbol{w}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}} = \underset{\mathbb{R}^{M\times N}}{\left.\frac{d\boldsymbol{w}}{d\boldsymbol{z}}\right|_{\boldsymbol{z}_0}} \times \underset{\mathbb{R}^{N\times 1}}{\left.\frac{d\boldsymbol{z}}{dy}\right|_{y_0}} \times \underset{\mathbb{R}^{1\times D}}{\left.\frac{dy}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}}$$

- In functional notations:

$$\mathbf{J}_{\boldsymbol{x}}[h \circ g \circ f](\boldsymbol{x}_0) = \mathbf{J}_{\boldsymbol{z}}[h](g \circ f(\boldsymbol{x}_0)) \times \begin{bmatrix} g_1' \circ f(\boldsymbol{x}_0) \\ \vdots \\ g_N' \circ f(\boldsymbol{x}_0) \end{bmatrix} \times \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top$$

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

Let:
- $\boldsymbol{x} \in \mathbb{R}^D$
- $y = f(\boldsymbol{x}) \in \mathbb{R}$
- $\boldsymbol{z} = g(y) = g(f(\boldsymbol{x})) \in \mathbb{R}^N$
- $\boldsymbol{w} = h(\boldsymbol{z}) = h(g(f(\boldsymbol{x}))) \in \mathbb{R}^M$

$$\left.\frac{d\boldsymbol{w}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{d\boldsymbol{w}}{d\boldsymbol{z}}\right|_{\boldsymbol{z}_0} \times \left.\frac{d\boldsymbol{z}}{dy}\right|_{y_0} \times \left.\frac{dy}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

$\mathbb{R}^{M \times D}$ $\mathbb{R}^{M \times N}$ $\mathbb{R}^{N \times 1}$ $\mathbb{R}^{1 \times D}$

- In functional notations:

$$\mathbf{J}_{\boldsymbol{x}}[h \circ g \circ f](\boldsymbol{x}_0) = \mathbf{J}_{\boldsymbol{z}}[h](g \circ f(\boldsymbol{x}_0)) \times \begin{bmatrix} g_1' \circ f(\boldsymbol{x}_0) \\ \vdots \\ g_N' \circ f(\boldsymbol{x}_0) \end{bmatrix} \times \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top$$

- All of this has natural generalizations to **tensors**

The Chain Rule

- High school review: $(g \circ f)'(x_0) = (g' \circ f)(x_0) \cdot f'(x_0)$
- By letting $y = f(x)$ and $z = g(y) = g(f(x))$ this writes:

$$\left.\frac{dz}{dx}\right|_{x_0} = \left.\frac{dz}{dy}\right|_{f(x_0)} \cdot \left.\frac{dy}{dx}\right|_{x_0}$$

"the constant x_0 flows' through the variables"

- This generalizes immediately to total derivatives!

Let:
- $\boldsymbol{x} \in \mathbb{R}^D$
- $y = f(\boldsymbol{x}) \in \mathbb{R}$
- $\boldsymbol{z} = g(y) = g(f(\boldsymbol{x})) \in \mathbb{R}^N$
- $\boldsymbol{w} = h(\boldsymbol{z}) = h(g(f(\boldsymbol{x}))) \in \mathbb{R}^M$

$$\left.\frac{d\boldsymbol{w}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{d\boldsymbol{w}}{d\boldsymbol{z}}\right|_{\boldsymbol{z}_0} \times \left.\frac{d\boldsymbol{z}}{dy}\right|_{y_0} \times \left.\frac{dy}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

with $\mathbb{R}^{M\times D}$, $\mathbb{R}^{M\times N}$, $\mathbb{R}^{N\times 1}$, $\mathbb{R}^{1\times D}$ respectively.

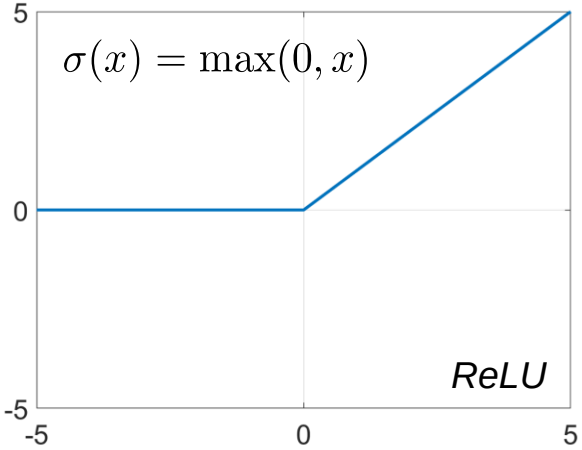
- In functional notations:

$$\mathbf{J}_{\boldsymbol{x}}[h \circ g \circ f](\boldsymbol{x}_0) = \mathbf{J}_{\boldsymbol{z}}[h](g \circ f(\boldsymbol{x}_0)) \times \begin{bmatrix} g_1' \circ f(\boldsymbol{x}_0) \\ \vdots \\ g_N' \circ f(\boldsymbol{x}_0) \end{bmatrix} \times \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)^\top$$

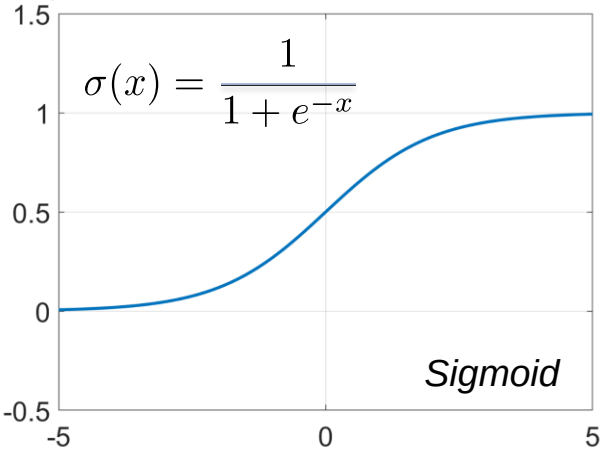
- All of this has natural generalizations to **tensors**
- All of this generalizes to ***partial derivatives***, e.g., $\left.\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{y}}\right|_{\boldsymbol{y}_0} \times \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$

Exercises: *Simple Derivatives*

- Calculate the derivatives of the following functions:



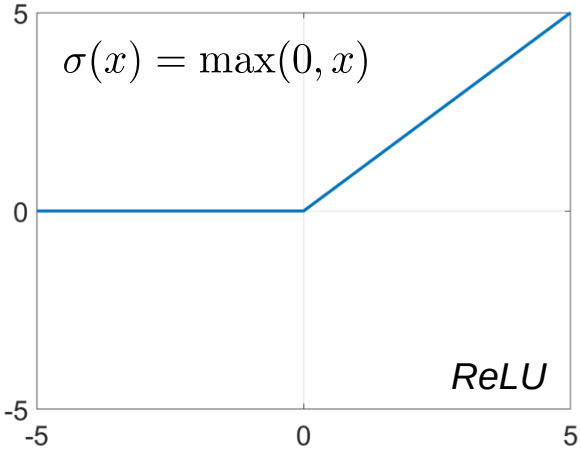
$\sigma'(x) = ?$



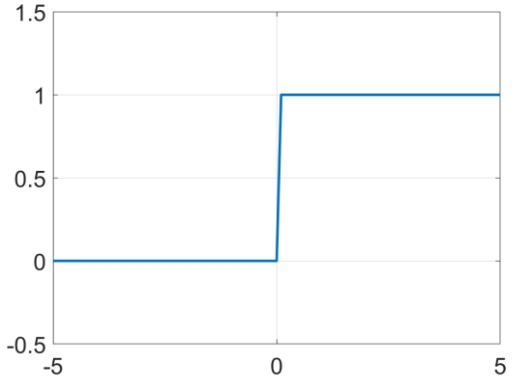
$\sigma'(x) = ?$

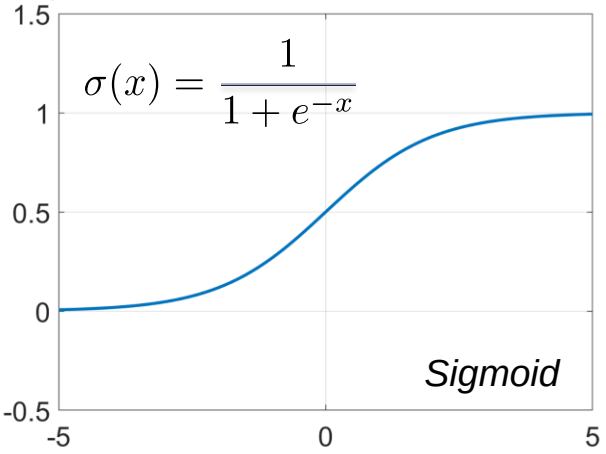
Exercises: *Simple Derivatives*

- Calculate the derivatives of the following functions:



$$\sigma'(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x < 1 \end{cases} = H(x)$$

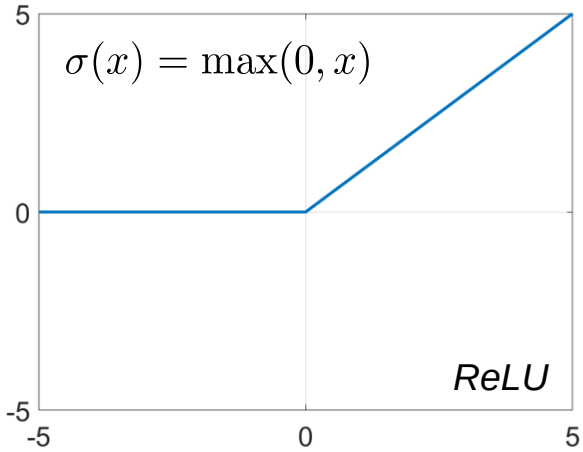


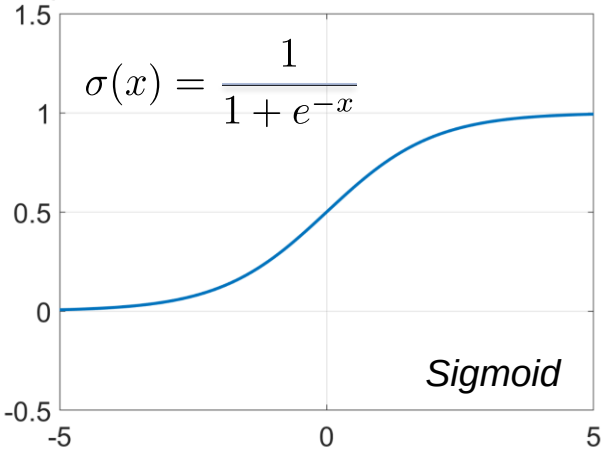


$$\sigma'(x) = ?$$

Exercises: *Simple Derivatives*

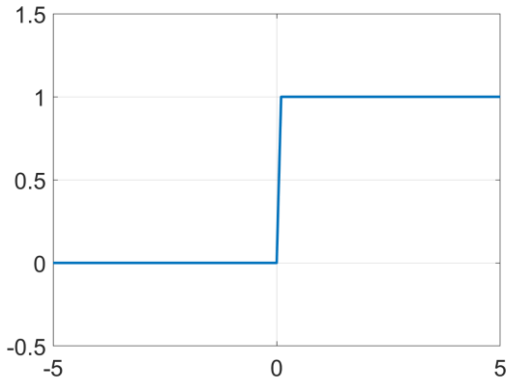
- Calculate the derivatives of the following functions:

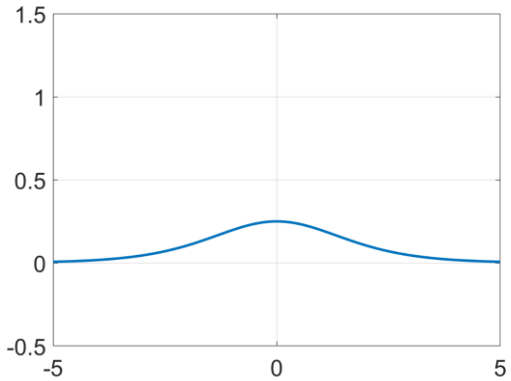




$$\sigma'(x) = \begin{cases} 0 \text{ for } x < 0 \\ 1 \text{ for } x < 1 \end{cases} = H(x)$$

$$\sigma'(x) = \frac{e^{-x}}{(1+e^{-x})^2} = \sigma(x)(1-\sigma(x))$$





Exercises: *Gradients*

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = \left[\frac{\partial f}{\partial x_1}(x_{0,1}), \ldots, \frac{\partial f}{\partial x_D}(x_{0,D}) \right]^\top$$

- Calculate the gradient of the following functions at $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x} \qquad \Rightarrow \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = ?$$

$$g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 \qquad \Rightarrow \nabla_{\boldsymbol{x}} g(\boldsymbol{x}_0) = ?$$

Exercises: *Gradients*

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = \left[\frac{\partial f}{\partial x_1}(x_{0,1}), \ldots, \frac{\partial f}{\partial x_D}(x_{0,D}) \right]^\top$$

- Calculate the gradient of the following functions at $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x} = b + \sum_{d=1}^{D} w_d x_d \quad \Rightarrow \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = ?$$

$$g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \sum_{d=1}^{D} (x_d)^2 \quad \Rightarrow \nabla_{\boldsymbol{x}} g(\boldsymbol{x}_0) = ?$$

Exercises: *Gradients*

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = \left[\frac{\partial f}{\partial x_1}(x_{0,1}), \ldots, \frac{\partial f}{\partial x_D}(x_{0,D})\right]^\top$$

- Calculate the gradient of the following functions at $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x} = b + \sum_{d=1}^{D} w_d x_d \quad \Rightarrow \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = [w_1, \ldots, w_D]^\top = \boldsymbol{w}$$

$$g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \sum_{d=1}^{D} (x_d)^2 \quad \Rightarrow \nabla_{\boldsymbol{x}} g(\boldsymbol{x}_0) = ?$$

Exercises: *Gradients*

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = \left[\frac{\partial f}{\partial x_1}(x_{0,1}), \ldots, \frac{\partial f}{\partial x_D}(x_{0,D}) \right]^\top$$

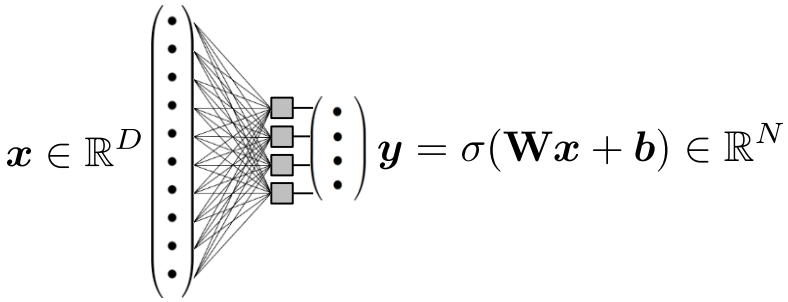
- Calculate the gradient of the following functions at $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) = b + \boldsymbol{w}^\top \boldsymbol{x} = b + \sum_{d=1}^{D} w_d x_d \quad \Rightarrow \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0) = [w_1, \ldots, w_D]^\top = \boldsymbol{w}$$

$$g(\boldsymbol{x}) = \|\boldsymbol{x}\|_2^2 = \sum_{d=1}^{D} (x_d)^2 \quad \Rightarrow \nabla_{\boldsymbol{x}} g(\boldsymbol{x}_0) = [2x_{0,1}, \ldots, 2x_{0,D}]^\top = 2\boldsymbol{x}_0$$

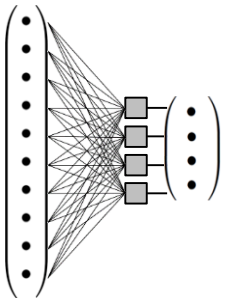
Exercises: *Jacobians*

Remember: *the* ***simple perceptron***



Exercises: *Jacobians*

Remember: *the* ***simple perceptron***

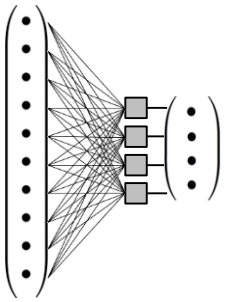


$\boldsymbol{x} \in \mathbb{R}^D$

$$\boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$
$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b}$$

Exercises: *Jacobians*

Remember: *the* ***simple perceptron***

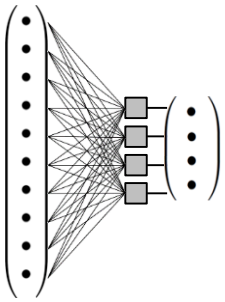


$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b}$$

Exercises: *Jacobians*

Remember: *the* ***simple perceptron***



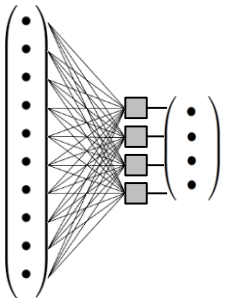
$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

Exercises: *Jacobians*

Remember: *the* ***simple perceptron***

$$\boldsymbol{x} \in \mathbb{R}^D$$



$$\boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$
$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

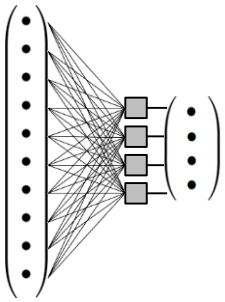
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

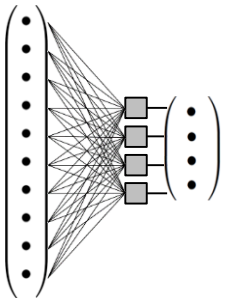
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

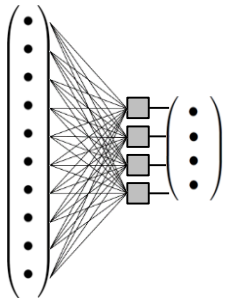
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

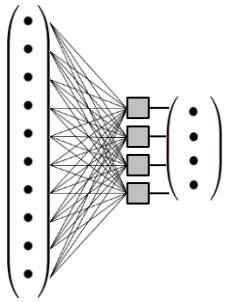
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

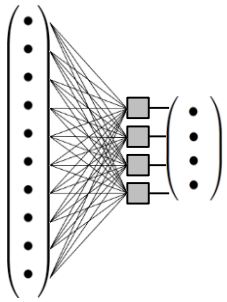
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$
$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

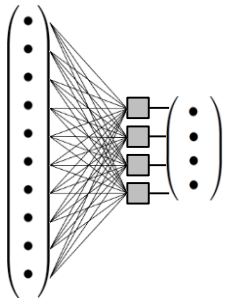
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = x_d \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

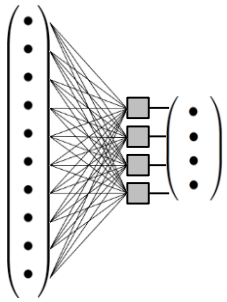
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \cdots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = x_d \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} = \begin{bmatrix} \sigma'(a_{0,1}) & 0 & \cdots & 0 \\ 0 & \sigma'(a_{0,2}) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma'(a_{0,N}) \end{bmatrix} = \text{diag}[\sigma'(\boldsymbol{a}_0)]$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} =$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

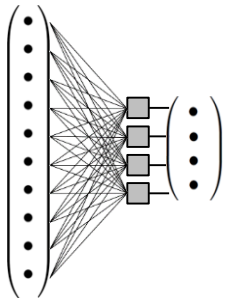
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = x_d \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} = \begin{bmatrix} \sigma'(a_{0,1}) & 0 & \dots & 0 \\ 0 & \sigma'(a_{0,2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma'(a_{0,N}) \end{bmatrix} = \text{diag}[\sigma'(\boldsymbol{a}_0)]$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}\boldsymbol{x}_0+\boldsymbol{b}} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \text{diag}[\sigma'(\mathbf{W}\boldsymbol{x}_0 + \boldsymbol{b})]\mathbf{W}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} =$$

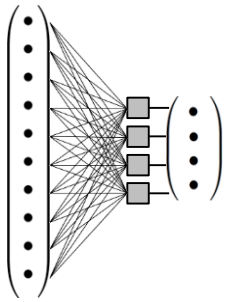
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = x_d \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} = \begin{bmatrix} \sigma'(a_{0,1}) & 0 & \dots & 0 \\ 0 & \sigma'(a_{0,2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma'(a_{0,N}) \end{bmatrix} = \text{diag}[\sigma'(\boldsymbol{a}_0)]$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}\boldsymbol{x}_0 + \boldsymbol{b}} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \text{diag}[\sigma'(\mathbf{W}\boldsymbol{x}_0 + \boldsymbol{b})]\mathbf{W}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}\boldsymbol{x} + \boldsymbol{b}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} = \text{diag}[\sigma'(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}_0)]$$

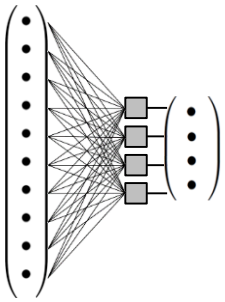
$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} =$$

Exercises: *Jacobians*

$$\left.\frac{d\boldsymbol{y}}{d\boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \left.\frac{\partial y_1}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_1}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_1}{\partial x_D}\right|_{x_{D,0}} \\ \left.\frac{\partial y_2}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_2}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_2}{\partial x_D}\right|_{x_{D,0}} \\ \vdots & \vdots & \ddots & \vdots \\ \left.\frac{\partial y_N}{\partial x_1}\right|_{x_{1,0}} & \left.\frac{\partial y_N}{\partial x_2}\right|_{x_{2,0}} & \dots & \left.\frac{\partial y_N}{\partial x_D}\right|_{x_{D,0}} \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{x}} f_1(\boldsymbol{x}_0)^\top \\ \nabla_{\boldsymbol{x}} f_2(\boldsymbol{x}_0)^\top \\ \vdots \\ \nabla_{\boldsymbol{x}} f_N(\boldsymbol{x}_0)^\top \end{bmatrix}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0}$$

Remember: *the* ***simple perceptron***



$$\boldsymbol{x} \in \mathbb{R}^D \quad \boldsymbol{y} = \sigma(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}) \in \mathbb{R}^N$$

$$= \sigma(\boldsymbol{a}) \text{ where } \boldsymbol{a} = \mathbf{W}\boldsymbol{x} + \boldsymbol{b} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \boldsymbol{x} \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \boldsymbol{x} \end{bmatrix} + \boldsymbol{b} = \sum_{d=1}^{D} \boldsymbol{w}_d x_d + \boldsymbol{b}$$

- Calculate the following Jacobians:

$$\left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \begin{bmatrix} \hat{\boldsymbol{w}}_1^\top \\ \vdots \\ \hat{\boldsymbol{w}}_N^\top \end{bmatrix} = \mathbf{W} \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{x}_0} = \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{x}_0} = x_d \mathbf{I}_N \qquad \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\boldsymbol{a}_0} = \begin{bmatrix} \sigma'(a_{0,1}) & 0 & \dots & 0 \\ 0 & \sigma'(a_{0,2}) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma'(a_{0,N}) \end{bmatrix} = \text{diag}[\sigma'(\boldsymbol{a}_0)]$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}\boldsymbol{x}_0 + \boldsymbol{b}} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{x}}\right|_{\boldsymbol{x}_0} = \text{diag}[\sigma'(\mathbf{W}\boldsymbol{x}_0 + \boldsymbol{b})]\mathbf{W}$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}\boldsymbol{x} + \boldsymbol{b}_0} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{b}}\right|_{\boldsymbol{b}_0} = \text{diag}[\sigma'(\mathbf{W}\boldsymbol{x} + \boldsymbol{b}_0)]$$

$$\left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} = \left.\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{a}}\right|_{\mathbf{W}_d\boldsymbol{x} + \boldsymbol{b}} \times \left.\frac{\partial \boldsymbol{a}}{\partial \boldsymbol{w}_d}\right|_{\boldsymbol{w}_d} = \text{diag}[\sigma'(\mathbf{W}_d\boldsymbol{x} + \boldsymbol{b})]x_d$$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = ?$$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = \frac{\partial r}{\partial \boldsymbol{a}}\Big|_{\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y}} \times \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}$$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = \frac{\partial r}{\partial \boldsymbol{a}}\Big|_{\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y}} \times \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}$$
$$= \nabla_{\boldsymbol{a}} g(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \times \mathbf{W}$$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = \frac{\partial r}{\partial \boldsymbol{a}}\Big|_{\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y}} \times \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}$$

$$= \nabla_{\boldsymbol{a}} g(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \times \mathbf{W}$$

We have: $\nabla_{\boldsymbol{a}} g(\boldsymbol{a}_0) = 2\boldsymbol{a}_0$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = \frac{\partial r}{\partial \boldsymbol{a}}\Big|_{\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y}} \times \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}$$

$$= \nabla_{\boldsymbol{a}} g(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \times \mathbf{W}$$

We have: $\nabla_{\boldsymbol{a}} g(\boldsymbol{a}_0) = 2\boldsymbol{a}_0$

$$= 2(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \mathbf{W}$$

Exercise: *Least Squares*

- Calculate the gradient of $f(\boldsymbol{\theta}) = \|\mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$ where $\mathbf{W} \in \mathbb{R}^{N \times D}$

Hint: Let $\boldsymbol{a} = \mathbf{W}\boldsymbol{\theta} - \boldsymbol{y}$, $g(\boldsymbol{a}) = \|\boldsymbol{a}\|_2^2$ and $r = g(\boldsymbol{a}) = f(\boldsymbol{\theta})$.

$$\nabla_\theta f(\boldsymbol{\theta}_0)^\top = \frac{\partial r}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0} = \frac{\partial r}{\partial \boldsymbol{a}}\Big|_{\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y}} \times \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}_0}$$

$$= \nabla_{\boldsymbol{a}} g(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \times \mathbf{W}$$

We have: $\nabla_{\boldsymbol{a}} g(\boldsymbol{a}_0) = 2\boldsymbol{a}_0$

$$= 2(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})^\top \mathbf{W}$$

$$\nabla_\theta f(\boldsymbol{\theta}_0) = 2\mathbf{W}^\top(\mathbf{W}\boldsymbol{\theta}_0 - \boldsymbol{y})$$