Machine Learning

Lecture 2
Math Recap - Vectors, Matrices, Functions

Felix Bießmann

Beuth University & Einstein Center for Digital Future

April 10, 2019

Machine Learning – The Big Picture

Given data $x \in \mathcal{X}$ (neural signals, internet user data, ...)
$\rightarrow$ predict variable $y \in \mathcal{Y}$ (thoughts, user intentions, ...)

ML algorithms learn a function $f(.)$ from examples (x,y)

$$f(x) = y \quad (1)$$

Most of ML is about:

- Find the right function class for $f(.)$
- Fitting $f(.)$ correctly

Machine Learning In Practice

After importing and cleaning a data set we have:

- D-dimensional data points $\mathbf{x}_i \in \mathbb{R}^D,\ i \in \{1, 2, \ldots, N\}$
  Often the data is stored in matrix form

$$X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N} \tag{2}$$

- Corresponding labels $\mathbf{y}_i \in \mathbb{R}^L$

The data points and labels are **vectors**.

$\rightarrow$ Today we recap some basics about vector spaces and matrices

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

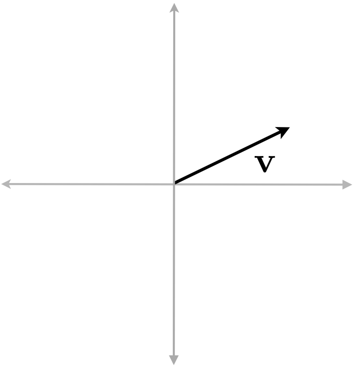
Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

Overview Math Recap

- Vectors
  - What is a vector, a scalar product, a norm?
  - What is orthogonality?
  - What are the rules for vector operations?
- Matrices
  - What is a linear mapping?
  - What are the rules for matrix multiplication?
- Derivatives
  - How to compute the gradient of a univariate function?
  - How to compute the gradient of a vector-valued function?

What is a Vector?



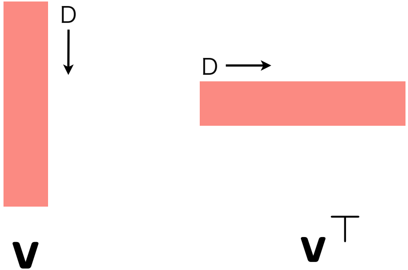
A vector $\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{bmatrix}$

can be understood as a point in a D-dimensional space

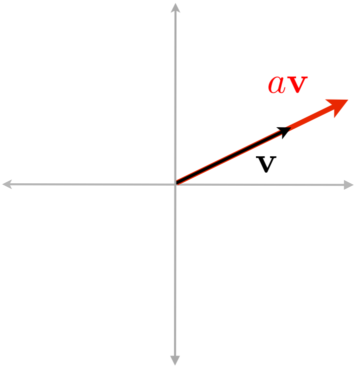
For comprehensibility we will restrict examples to $\mathbf{v} \in \mathbb{R}^2$.

What is a transposed Vector?

$\mathbf{v}^\top$ denotes $\mathbf{v}$−**transposed**



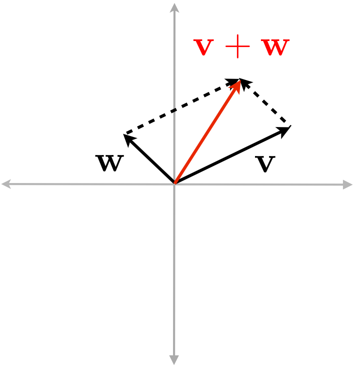
Vector Operations



Vectors can be scaled

$$a\mathbf{v} = \begin{bmatrix} av_1 \\ av_2 \\ \vdots \\ av_D \end{bmatrix} \quad (3)$$

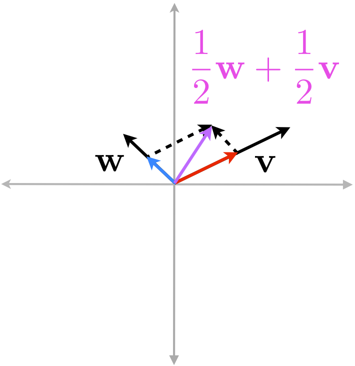
Vector Operations



Vectors from the same space can be added to each other

$$\mathbf{v} + \mathbf{w} = \begin{bmatrix} v_1 + w_1 \\ v_2 + w_2 \\ \vdots \\ v_D + w_D \end{bmatrix} \tag{4}$$

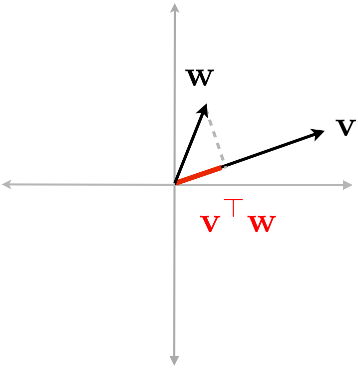
Vector Operations



Scaling and addition can be combined

$$a\mathbf{v} + b\mathbf{w} = \begin{bmatrix} av_1 + bw_1 \\ av_2 + bw_2 \\ \vdots \\ av_D + bw_D \end{bmatrix} \tag{5}$$

Vector Operations

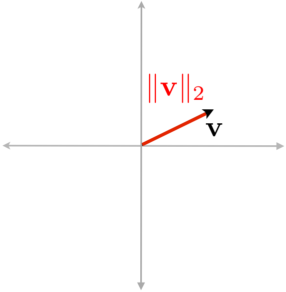


The **scalar product** computes the *similarity* between two vectors $\mathbf{v}$ and $\mathbf{w}$

$$\langle \mathbf{v}, \mathbf{w} \rangle = \mathbf{v}^\top \mathbf{w} = [v_1 \ v_2] \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \sum_i^D v_i w_i \tag{6}$$

$\mathbf{v}^\top \mathbf{w}$ is often called *projecting* $\mathbf{w}$ onto $\mathbf{v}$

Length of a Vector



Sometimes we are not interested in the direction of a vector, but only in its length.

How can we measure the **length of a vector**?

Norm of a Vector

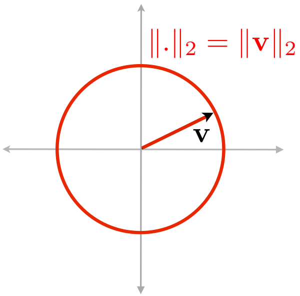
The p-**norm** $\|\mathbf{v}\|_p$ of a D-dimensional vector $\mathbf{v}$ is defined as

$$\|\mathbf{v}\|_p = \left(\sum_i^D |v_i|^p) \right)^{1/p} \tag{7}$$

Usually (and in the following) $\|.\|$ denotes the **euklidean norm** ($p = 2$)

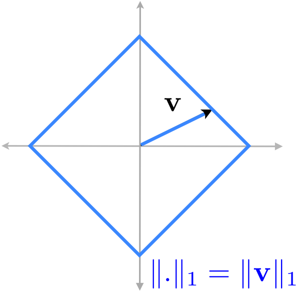
$$\|\mathbf{v}\|_2 = \left(\sum_i^D |v_i|^2) \right)^{1/2} = \sqrt{\sum_i^D v_i^2} = \sqrt{\mathbf{v}^\top \mathbf{v}} \tag{8}$$

The Euklidean (2-Norm)



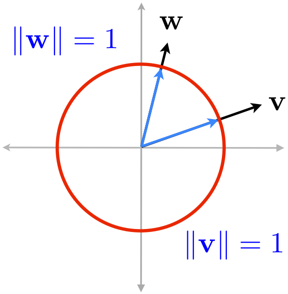
$$\|\mathbf{v}\|_2 = \left(\sum_i^D |v_i|^2)\right)^{1/2} = \sqrt{\mathbf{v}^\top \mathbf{v}} \tag{9}$$

The 1-Norm



$$\|\mathbf{v}\|_1 = \left(\sum_i^D |v_i|\right) = \sum_i^D |v_i| \qquad (10)$$

Unit vectors



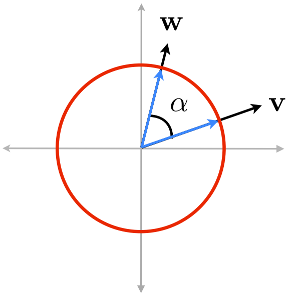
Every vector $\mathbf{v}$ can be scaled to length 1 by

$$\mathbf{v} \leftarrow \mathbf{v}/\|\mathbf{v}\| \quad (11)$$

This is useful. Why?

Because after normalizing each vector to unit length, our vector similarity measure $\mathbf{v}^\top \mathbf{w}$ has a standardized meaning.

Angle between two vectors



The scalar product of unit vectors is the cosine of the angle between the vectors

$$\frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|} = cos(\alpha) \tag{12}$$

So the angle α between **v** and **w** is

$$\alpha = acos\left(\frac{\mathbf{v}^\top \mathbf{w}}{\|\mathbf{v}\|\|\mathbf{w}\|}\right). \tag{13}$$

Note that $-1 \leq cos(\alpha) \leq 1$.

Does that remind you of something?

Angles between vectors and correlation coefficients

Consider T zero-mean[1] samples from two univariate time series $x_i \in \mathbb{R}^1, y_i \in \mathbb{R}^1, i \in \{1, \ldots, T\}$ stored in vectors $\mathbf{x}$, $\mathbf{y}$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_T \end{bmatrix} \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \tag{14}$$

The **correlation coefficient** between $\mathbf{x}$ and $\mathbf{y}$ is defined as

$$\frac{\sum_i^T x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|\|\mathbf{y}\|} \tag{15}$$

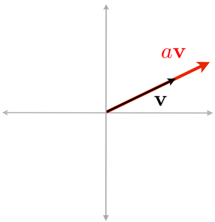
$\Rightarrow$ The cosine of the angle between $\mathbf{x}$ and $\mathbf{y}$ is their correlation!

[1] $\sum_i^T x_i = \sum_i^T y_i = 0$

Summary Vector Operations

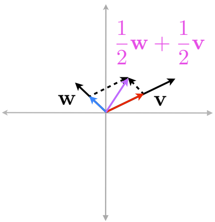
Vector Scaling

$$a\mathbf{v} = \begin{bmatrix} av_1 \\ av_2 \\ \vdots \\ av_D \end{bmatrix}$$



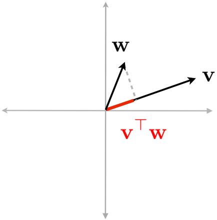
Vector Addition

$$a\mathbf{v} + b\mathbf{w} = \begin{bmatrix} av_1 + bw_1 \\ av_2 + bw_2 \\ \vdots \\ av_D + bw_D \end{bmatrix}$$



Scalar Product

$$\mathbf{v}^\top \mathbf{w} = \sum_i^D v_i w_i$$



Axioms of Vector Operations

1. Associativity of addition

$$\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w} \tag{16}$$

2. Commutativity of addition

$$\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v} \tag{17}$$

3. Distributivity of scaling

$$a(\mathbf{v} + \mathbf{w}) = a\mathbf{w} + a\mathbf{v} \tag{18}$$

Matrices

This is an example of a matrix $A \in \mathbb{R}^{U \times V}$

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1V} \\ A_{21} & A_{22} & \dots & A_{2V} \\ \vdots & \vdots & \ddots & \vdots \\ A_{U1} & A_{U2} & \dots & A_{UV} \end{bmatrix}$$

Matrix Multiplication

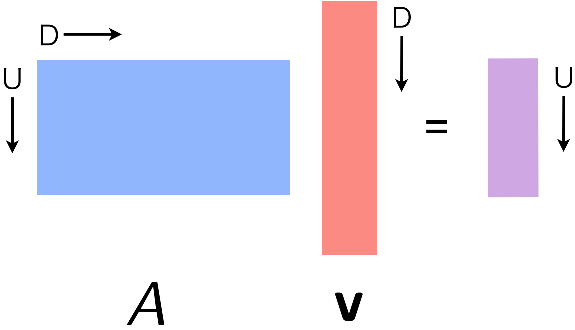
Matrix $A \in \mathbb{R}^{U \times D}$ can be multiplied with a vector $\mathbf{v} \in \mathbb{R}^D$ by

$$A\mathbf{v} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1D} \\ A_{21} & A_{22} & \dots & A_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ A_{U1} & A_{U2} & \dots & A_{UD} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_D \end{bmatrix} \tag{19}$$

$$= \begin{bmatrix} A_{11}v_1 & + & A_{12}v_2 & \dots & + & A_{1D}v_D \\ A_{21}v_1 & + & A_{22}v_2 & \dots & + & A_{2D}v_D \\ \vdots & & \vdots & & \ddots & \vdots \\ A_{U1}v_1 & + & A_{U2}v_2 & \dots & + & A_{UD}v_D \end{bmatrix} \tag{20}$$

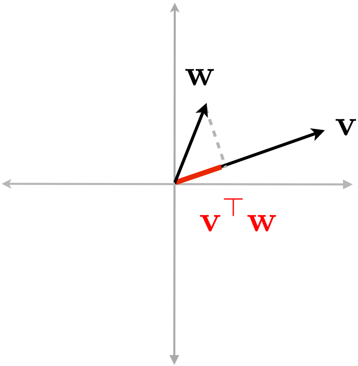
Matrix Multiplication

Matrix multiplication as pictorial illustration



Matrices as Linear Mappings

A matrix $A \in \mathbb{R}^{U \times V}$ can be understood as a **linear map** from a U-dimensional space to a V-dimensional space



Consider the scalar product $\mathbf{v}^\top \mathbf{w}$
$\mathbf{w}$ is mapped from $\mathbb{R}^2 \rightarrow \mathbb{R}^1$

But the mapping $\mathbf{v}^\top \in \mathbb{R}^{1 \times 2}$ is a very simple case

Matrices as Linear Mappings

- Permutation of dimensions:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

- Mirroring:

$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

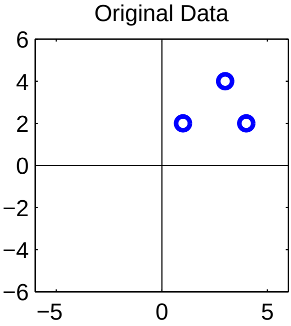
- Scaling:

$$A = \begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix}$$

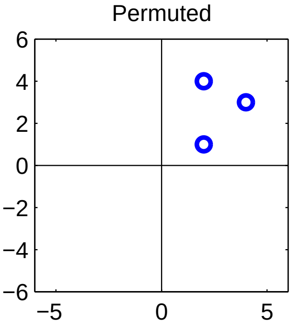
- Rotation by an angle of α in radians:

$$A = \begin{bmatrix} cos(\alpha) & -sin(\alpha) \\ sin(\alpha) & cos(\alpha) \end{bmatrix}$$

Matrices as Linear Mappings: Permutation

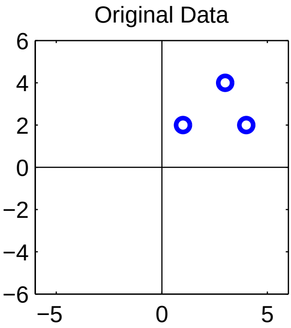


$AX \quad \rightarrow$

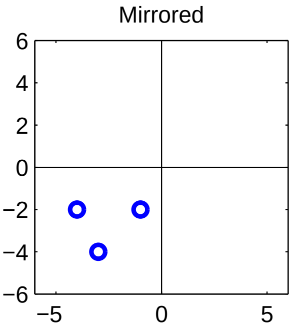


$$X = \begin{bmatrix} 1 & 3 & 4 \\ 2 & 4 & 2 \end{bmatrix} \qquad A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \qquad AX = \begin{bmatrix} 2 & 4 & 2 \\ 1 & 3 & 4 \end{bmatrix}$$

Matrices as Linear Mappings: Mirroring

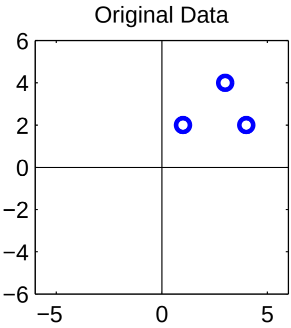


$AX \quad \rightarrow$

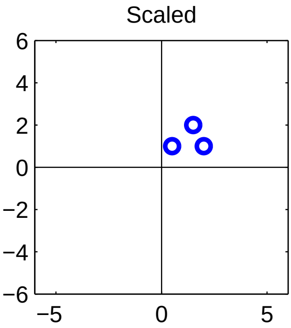


$$A = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Matrices as Linear Mappings: Scaling

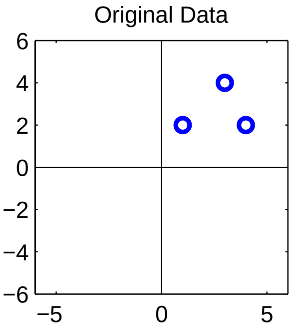


$AX \quad \rightarrow$

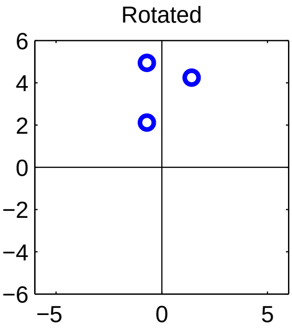


$$A = \begin{bmatrix} 1/2 & 0 \\ 0 & 1/2 \end{bmatrix}$$

Matrices as Linear Mappings: Rotation



$AX \quad \rightarrow$



$$A = \begin{bmatrix} cos(\pi \cdot 45/180) & -sin(\pi \cdot 45/180) \\ sin(\pi \cdot 45/180) & cos(\pi \cdot 45/180) \end{bmatrix}$$

Axioms of Matrix Operations

1. Associativity

$$A(BC) = (AB)C \tag{21}$$

2. Distributivity

$$A(B + C) = AB + AC, \qquad (A + B)C = AC + BC \tag{22}$$

3. **No Commutativity!**

$$AB \neq BA \tag{23}$$

Matrix Multiplication – Some Examples

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 15 & -17 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -2 & 7 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ -1 & 5 \end{bmatrix} \begin{bmatrix} 3 & 2 & 8 \\ 7 & 5 & 1 \end{bmatrix}$$

Matrix Multiplication – Some Examples

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 15 & -17 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ 15 & -17 \end{bmatrix}$$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ -2 & 7 \end{bmatrix} = \begin{bmatrix} -2 & 7 \\ 3 & 2 \end{bmatrix}$$

$$\begin{bmatrix} 2 & 3 \\ -1 & 5 \end{bmatrix} \begin{bmatrix} 3 & 2 & 8 \\ 7 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 27 & 19 & 19 \\ 32 & 23 & -3 \end{bmatrix}$$

Minimizing Functions

Remember:

- ML algorithms are functions f that need to be fitted to data
- Training ML algorithms is finding minima/maxima of functions
- Function minimization/maximization is often done by **Gradient Descent**

Taking Derivatives of Univariate Functions

Function	Derivative
$f(x) = x^n$	$f'(x) = nx^{n-1}$
$cf(x)$	$cf'(x)$
$f(x) + g(x)$	$f'(x) + g'(x)$
$f(x)g(x)$	$f'(x)g(x) + f(x)g'(x)$
$\frac{f(x)}{g(x)}$	$\frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2}$
$f(g(x))$	$f'(g(x))\, g'(x)$

Taking Derivatives of Vector-Valued Functions

Consider a vector-valued function $f(\mathbf{x}) : \mathbb{R}^D \rightarrow \mathbb{R}^U, \mathbf{x} \in \mathbb{R}^D$

The gradient $\nabla f(\mathbf{x})$ at position $\mathbf{x}$ is

$$\nabla f(x) = \left[\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_D}\right] \tag{24}$$

Taking Derivatives of Vector-Valued Functions

Function	Derivative
$f(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$	
$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$	
$f(\mathbf{x}) = A\mathbf{x}$	
$f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$	

More complicated examples: The Matrix Cookbook

Taking Derivatives of Vector-Valued Functions

Function	Derivative
$f(\mathbf{x}) = \mathbf{u}^\top \mathbf{x}$	$f'(\mathbf{x}) = \mathbf{u}$
$f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$	$f'(\mathbf{x}) = 2\mathbf{x}$
$f(\mathbf{x}) = A\mathbf{x}$	$f'(\mathbf{x}) = A$
$f(\mathbf{x}) = \mathbf{x}^\top A\mathbf{x}$	$f'(\mathbf{x}) = \mathbf{x}^\top (A + A^\top)$

More complicated examples: The Matrix Cookbook