Mathematical Basics

Machine Learning 1 — Lecture 2

12th March 2024

Robert Peharz
Institute of Theoretical Computer Science
Graz University of Technology

Mathematics

"Mathematics" comes from Greek:

μάθημα: knowledge, study, learning
μαθηματικός: on the matter of that which is learned

Mathematics

"Mathematics" comes from Greek:

> μάθημα: knowledge, study, learning
> μαθηματικός: on the matter of that which is learned

Wait, is machine learning just... mathematics?

Mathematics

"Mathematics" comes from Greek:

μάθημα: knowledge, study, learning
μαθηματικός: on the matter of that which is learned

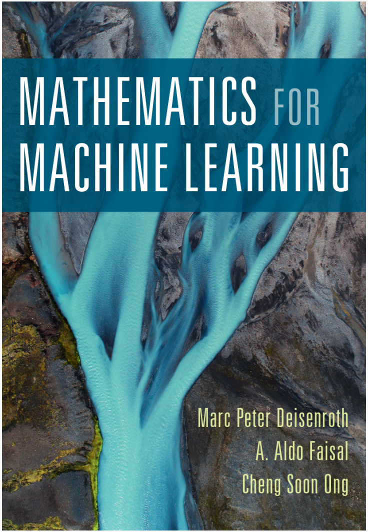
Wait, is machine learning just... mathematics?

Frankly, **yes!** Machine learning is mostly applied math.

I do hope you will have a boring lecture...

- This lecture is intended as a **refresher** at a fairly quick pace, not as a complete introduction
- Most of today's lecture should be familiar to you
- If you severely struggle following today's lecture, the remainder of the course will be very challenging!

Reading Material



https://mml-book.github.io/

I can recommend this excellent book, either for further reading or if you need to catch up.

Core Topics

The core mathematical disciplines in Machine Learning are:

- **Linear algebra**
- **Calculus**
- **Probability**

Euclidean Vector Space

- The Euclidean space is the set of all **D-tuples** $\boldsymbol{x}$ of real numbers:

$$\mathbb{R}^D = \mathbb{R} \times \mathbb{R} \times \cdots \times \mathbb{R} = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_D \end{pmatrix} : \; x_1, \ldots, x_D \in \mathbb{R} \right\}$$

- By default, we will interpret all vectors $\boldsymbol{x} \in \mathbb{R}^D$ as **column-vectors**
- **Row-vectors** are denoted as $\boldsymbol{x}^T$ ($\boldsymbol{x}$ transposed)
- Thus, we also write $\boldsymbol{x} = (x_1, \ldots, x_D)^T$
- D is the **dimensionality** of $\mathbb{R}^D$

- We can **add** and **subtract** vectors $\boldsymbol{x} = (x_1, x_2, \dots, x_D)^T$, $\boldsymbol{y} = (y_1, y_2, \dots, y_D)^T$:

$$\boldsymbol{x} + \boldsymbol{y} = (x_1 + y_1, x_2 + y_2, \dots, x_D + y_D)^T$$

$$\boldsymbol{x} - \boldsymbol{y} = (x_1 - y_1, x_2 - y_2, \dots, x_D - y_D)^T$$

- We can **scale** vectors, i.e. multiply with a scalar a:

$$a\boldsymbol{x} = (ax_1, ax_2, \dots, ax_D)^T$$

A **norm** $\|\cdot\|$ defines a notion of length to a vector $\boldsymbol{x} \in \mathbb{R}^D$.

- **Manhattan norm**: $\|\boldsymbol{x}\|_1 := \sum_i |x_i|$
- **Euclidean norm**: $\|\boldsymbol{x}\|_2 := \sqrt{\sum_i x_i^2}$
- **Max norm**: $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$

These are special cases of the **p-norm**, $p \geq 1$

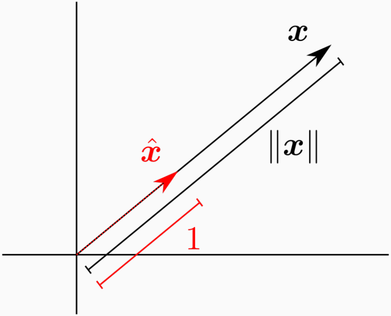
$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{D} |x_i|^p \right)^{1/p}$$

denoted ℓ_p-norm, L_p-norm, etc.

A **unit vector** has length (norm) 1. We can **normalize** a vector $\boldsymbol{x}$ by dividing by its norm:

$$\hat{\boldsymbol{x}} = \frac{1}{\|\boldsymbol{x}\|}\boldsymbol{x} = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}$$

$\hat{\boldsymbol{x}}$ is the unit vector "showing in the same direction" as $\boldsymbol{x}$.



Unit Spheres

Unitspheres (sets containing all unit vectors) for various p-norms:

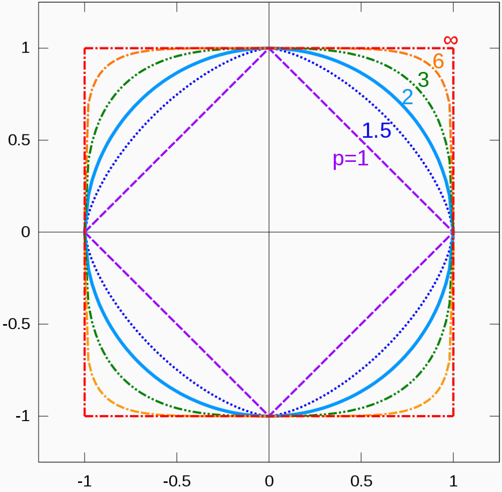


Image: wikipedia

As soon as we have defined a norm $\|\cdot\|$, we can define a **distance** $d(\cdot,\cdot)$ between two vectors:

$$d(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|$$

Of course the distance depends on the used norm. The Euclidean norm $\|\cdot\|_2$ measure distance differently than the max norm $\|\cdot\|_\infty$.

- For any two vectors $\boldsymbol{x} = (x_1, \ldots, x_D)^T$, $\boldsymbol{y} = (y_1, \ldots, y_D)^T$, the **dot product** is defined as

$$\boldsymbol{x}^T \boldsymbol{y} := \sum_{i=1}^{D} x_i y_i$$

- E.g., for $\boldsymbol{x} = (3, 1, 2.2)^T$ and $\boldsymbol{y} = (4, 0, -1)^T$

$$\boldsymbol{x}^T \boldsymbol{y} = 3 \times 4 + 1 \times 0 + 2.2 \times (-1) = 9.8$$

- Note that the dot product and the Euclidean norm (ℓ_2-norm) are related via

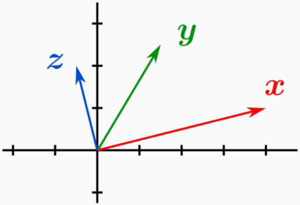
$$\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^T \boldsymbol{x}}$$

The dot products measures the **angle** between two vectors. In particular, the cosine of the angle ω between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ is given via:

$$-1 \leq \cos\omega =: \frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|_2\,\|\boldsymbol{y}\|_2} \leq 1$$

If $\boldsymbol{x}^T\boldsymbol{y} = 0 = cos(\pi/2)$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthogonal**, written $\boldsymbol{x}\perp\boldsymbol{y}$.

If additionally, $\|\boldsymbol{x}\|_2 = \|\boldsymbol{y}\|_2 = 1$, then $\boldsymbol{x}$ and $\boldsymbol{y}$ are **orthonormal**.



$\boldsymbol{x} = (4, 1)^T \quad \boldsymbol{y} = (1.5, 2.5)^T \quad \boldsymbol{z} = (-0.5, 2)^T$

$$\frac{\boldsymbol{x}^T\boldsymbol{y}}{\|\boldsymbol{x}\|\|\boldsymbol{y}\|} = \frac{4\times1.5+1\times2.5}{4.123\times2.915} = 0.7071 = cos(\tfrac{\pi}{4}) \triangleq 45^\circ$$

$$\frac{\boldsymbol{x}^T\boldsymbol{z}}{\|\boldsymbol{x}\|\|\boldsymbol{z}\|} = \frac{4\times(-0.5)+1\times2}{4.123\times2.062} = 0 = cos(\tfrac{\pi}{2}) \triangleq 90^\circ$$

$$\frac{\boldsymbol{y}^T\boldsymbol{z}}{\|\boldsymbol{y}\|\|\boldsymbol{z}\|} = \frac{1.5\times(-0.5)+2.5\times2}{2.915\times2.062} = 0.7071 = cos(\tfrac{\pi}{4}) \triangleq 45^\circ$$

Vector Projection

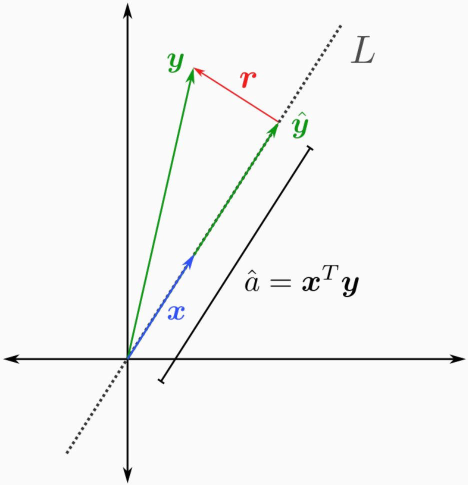
Assume a normalized vector $\boldsymbol{x}$. Then $f(a) = a\boldsymbol{x}$ parametrizes a **line** L going through the origin $(-\infty < a < \infty)$.

Assume another vector $\boldsymbol{y}$, not necessarily normalized. The **projection** $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ onto L is given as

$$\hat{\boldsymbol{y}} = (\boldsymbol{y}^T \boldsymbol{x})\boldsymbol{x} = \hat{a}\boldsymbol{x}$$

where $\hat{a} = \boldsymbol{y}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{y}$ is given by the dot product. $\hat{\boldsymbol{y}}$ is the **closest** vector (in Euclidean distance) to $\boldsymbol{y}$ which lies on L.

$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}$ is the **residual vector**, since $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{r}$. Vectors $\hat{\boldsymbol{y}}$ and $\boldsymbol{r}$ are orthogonal.

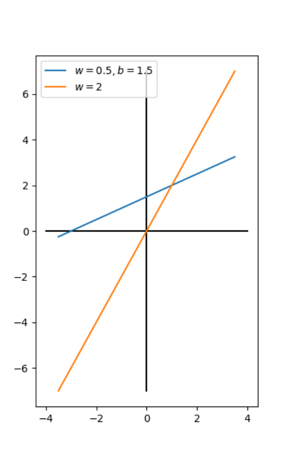

L
$\boldsymbol{y}$
$\boldsymbol{r}$
$\hat{\boldsymbol{y}}$
$\hat{a} = \boldsymbol{x}^T\boldsymbol{y}$
$\boldsymbol{x}$

A vector $\boldsymbol{w} \in \mathbb{R}^D$ defines a **linear function** $f : \mathbb{R}^D \mapsto \mathbb{R}$ via the dot product:

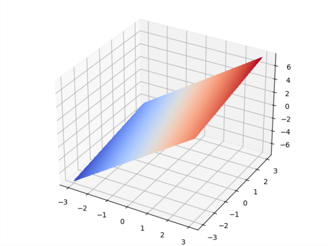
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$$

By adding a **bias** b, we get an **affine function**:

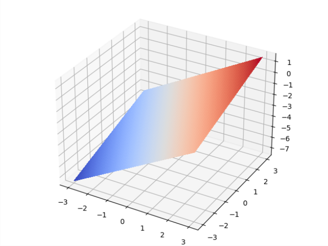
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$$



$\boldsymbol{w} = (2, 0.5)^T,\ b = 0$



$\boldsymbol{w} = (1, 0.5)^T,\ b = -3$



A vector $\boldsymbol{w} \in \mathbb{R}^D$ and a bias $b \in \mathbb{R}$ define an affine function $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + b$. The equation

$$f(\boldsymbol{x}) = 0$$

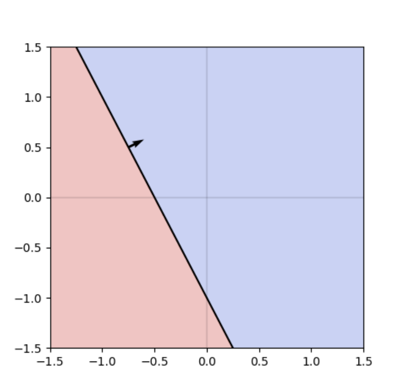
defines a **hyperplane** (a $D-1$ dimensional sub-space).

Moreover, two **half-spaces** are defined via

$$f(\boldsymbol{x}) < 0 \qquad \textbf{negative half-space}$$
$$f(\boldsymbol{x}) \geq 0 \qquad \textbf{positive half-space}$$

The vector $\boldsymbol{w}$ is **orthogonal** to the hyperplane and points into the positive half-space.

In 2-D, the hyperplane is a line, cutting $\mathbb{R}^D$ in two parts. For $\boldsymbol{w} = (2, 1)^T$ and $b = 1$:



blue: positive half-space
red: negative half-space

In 3-D, a hyperplane is an actual 2-D plane. In higher dimensions it is a corresponding generalization, hence the name *hyper*plane.

Matrices

A **matrix** is a rectangular array of real (or complex) numbers.
With $M, N \in \mathbb{N}$, a matrix $A \in \mathbb{R}^{M \times N}$ can be written as

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{pmatrix}$$

where M is the number of **rows** and N is the number of **columns**.
The **entry** a_{ij} is a scalar located at the i^{th} row and j^{th} column.

Like vectors, matrices are added *element-wise*:

$$A + B = C = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1N} \\ c_{21} & c_{22} & \cdots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \cdots & c_{MN} \end{pmatrix}$$

where $c_{ij} = a_{ij} + b_{ij}$.

Row and Column Vectors

A matrix can be understood as a collection of M **row-vectors**, or as a collection of N **column-vectors**:

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1N} \\ a_{21} & a_{22} & \dots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{M1} & a_{M2} & \dots & a_{MN} \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_{1:}^T \\ \boldsymbol{a}_{2:}^T \\ \vdots \\ \boldsymbol{a}_{M:}^T \end{pmatrix} = \begin{pmatrix} \boldsymbol{a}_{:1} & \boldsymbol{a}_{:2} & \dots & \boldsymbol{a}_{:N} \end{pmatrix}$$

where

$\boldsymbol{a}_{i:}^T = (a_{i1}, a_{i2}, \dots, a_{iN})$ **row-vectors**

$\boldsymbol{a}_{:j} = (a_{1j}, a_{2j}, \dots, a_{Mj})^T$ **column-vectors**

Note: we use the "colon-notation" $\boldsymbol{a}_{i:}$, $\boldsymbol{a}_{:j}$ to address rows and columns, similar as in `python`. Read "i-everything" or "everything-j".

The **transpose** A^T of a matrix A is a matrix A^T whose rows are the columns of A, (or, whose columns are the rows of A). For example:

$$A = \begin{pmatrix} 2 & 3 & 5 & 7 \\ 11 & 13 & 17 & 19 \\ 23 & 29 & 31 & 37 \end{pmatrix}$$

$$A^T = \begin{pmatrix} 2 & 11 & 23 \\ 3 & 13 & 29 \\ 5 & 17 & 31 \\ 7 & 29 & 37 \end{pmatrix}$$

A matrix A which has the same number of rows and columns is called a **square matrix**. For example:

$$A = \begin{pmatrix} 0 & 1 & 1 \\ 2 & 3 & 5 \\ 8 & 13 & 21 \end{pmatrix}$$

A square matrix A for which it holds that $A = A^T$ is called **symmetric**. For example:

$$A = \begin{pmatrix} 3 & 1 & 7 \\ 1 & 2 & 9 \\ 7 & 9 & 1 \end{pmatrix}$$

A square matrix which contains only 0's, except on the main diagonal, is called a **diagonal** matrix. For example:

$$A = \begin{pmatrix} 3.1415 & 0 & 0 \\ 0 & -7 & 0 \\ 0 & 0 & 42 \end{pmatrix}$$

Of course, a diagonal matrix is symmetric.

Matrix multiplication is defined in terms of the **dot products** between row-vectors and column vectors. In particular let $A \in \mathbb{R}^{M \times R}$ and $B \in \mathbb{R}^{R \times N}$ (note the same R):

$$A = \begin{pmatrix} \boldsymbol{a}_{1:}^T \\ \vdots \\ \boldsymbol{a}_{M:}^T \end{pmatrix}, \quad B = \begin{pmatrix} \boldsymbol{b}_{:1} & \dots & \boldsymbol{b}_{:N} \end{pmatrix}$$

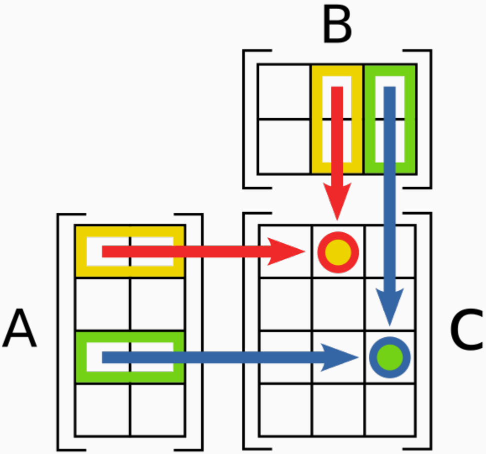
Then the **matrix product** $C = AB$ is given as a matrix $C \in \mathbb{R}^{M \times N}$

$$C = AB = \begin{pmatrix} c_{11} & c_{12} & \dots & c_{1N} \\ c_{21} & c_{22} & \dots & c_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \dots & c_{MN} \end{pmatrix}$$

where $c_{ij} = \boldsymbol{a}_{i:}^T \boldsymbol{b}_{:j} = \sum_{k=1}^{R} a_{ik} b_{kj}$

$$B = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{pmatrix}$$



$$C = \begin{pmatrix} 9 & 12 & 15 \\ 19 & 26 & 33 \\ 29 & 40 & 51 \\ 39 & 54 & 69 \end{pmatrix}$$

Image: wikipedia

Properties of Matrix Product

For matrices A, B, C, D with compatible dimensions, the following properties hold:

- $(AB)C = A(BC)$ **associativity**
- $A(B + C) = AB + AC$ **distributivity**
- $(AB)^T = B^T A^T$
- AA^T is always symmetric

Attention: matrix multiplication is **not commutative**, i.e. in general:

$$AB \neq BA$$

By interpreting vectors as “thin” matrices we can also multiply

- $A \in \mathbb{R}^{M \times N}$ and vector $\boldsymbol{x} \in \mathbb{R}^N$, yielding $\boldsymbol{y} \in \mathbb{R}^M$:

$$\boldsymbol{y} = A\boldsymbol{x}$$

- $\boldsymbol{x} \in \mathbb{R}^M$ and $A \in \mathbb{R}^{M \times N}$, yielding $\boldsymbol{y} \in \mathbb{R}^N$:

$$\boldsymbol{y}^T = \boldsymbol{x}^T A$$

Matrices as Linear Functions

Recall that the **dot product** with vector $\boldsymbol{w} \in \mathbb{R}^D$ defines a linear function $f(\boldsymbol{x})\colon \mathbb{R}^D \mapsto \mathbb{R}$, $f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$.

Similarly, the **matrix-vector product** with matrix $A \in \mathbb{R}^{M \times N}$ defines a **linear function** $f\colon \mathbb{R}^N \mapsto \mathbb{R}^M$

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

We get an **affine function** by adding an M-dimensional bias vector $\boldsymbol{b}$:

$$f(\boldsymbol{x}) = A\boldsymbol{x} + \boldsymbol{b}$$

The diagonal matrix I which contains only 1's in the diagonal is called the **identity matrix**. There is an identity matrix for each $M \in \mathbb{N}$. For $M = 3$:

$$I = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The identity matrix represents the **identity function**, since

$$I\boldsymbol{x} = \boldsymbol{x}$$

Let A be a square matrix. If there exists a matrix A^{-1} such that

$$A^{-1}A = I$$

then A^{-1} is called the **inverse (matrix)** of A.

If the inverse matrix exists, then it also holds that

$$AA^{-1} = I$$

The computational complexity of computing the inverse of a matrix $A \in \mathbb{R}^{M \times M}$ is $\mathcal{O}(M^3)$.

Linear Basis, Linear Subspace

Linear Combination, Linear Independence **Definition**

Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$ be K vectors in $\mathbb{R}^D$. A **linear combination** of these vectors yields another vector written as

$$\sum_{k=1}^{K} z_k \boldsymbol{x}_k$$

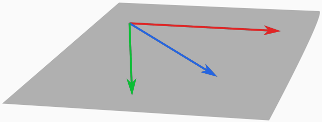
with coefficients $z_k \in \mathbb{R}$.

The vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_K$ are called **linearly independent**, when **none of them can be expressed as a linear combination of the other vectors.**

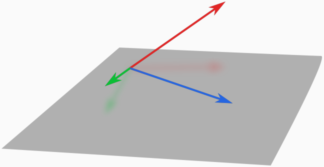
That is, for each $1 \leq i \leq K$ and any coefficients z_k

$$\boldsymbol{x}_i \neq \sum_{k=1, k \neq i}^{K} z_k \boldsymbol{x}_k$$

Linearly dependent vectors



Linearly independent vectors



Let $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_D$ be D **linearly independent vectors** in $\mathbb{R}^D$.
Any such collection $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_D$ is called a **basis** of $\mathbb{R}^D$.

For any basis, it is possible to express any vector $\boldsymbol{x} \in \mathbb{R}^D$ as a linear combination

$$\boldsymbol{x} = \sum_{d=1}^{D} \boldsymbol{b}_d z_d,$$

with **unique** coefficients z_d.

The coefficient vector $\boldsymbol{z} = (z_1, \ldots, z_D)^T$ represents $\boldsymbol{x}$ in the basis $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_D$, i.e. a **change of basis**.

How to find $\boldsymbol{z}$?

Let $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_D$ be a basis of $\mathbb{R}^D$. We can collect the basis vectors as columns of a **basis matrix** B:

$$B = \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \ldots & \boldsymbol{b}_D \end{pmatrix}$$

Then we can express any $\boldsymbol{x} \in \mathbb{R}^D$ as

$$\boldsymbol{x} = \sum_{d=1}^{D} \boldsymbol{b}_d z_d = B\boldsymbol{z}.$$

Thus $\boldsymbol{z}$ is given as:

$$B^{-1}\boldsymbol{x} = \boldsymbol{z}$$

Note: B^{-1} exists if and only if $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_D$ are a basis.

Things become particularly nice if the basis vectors are **orthonormal**, i.e. if they are

- normalized (i.e. $\|\boldsymbol{b}_i\|_2 = 1$)
- orthogonal

This can be compactly described as

$$\boldsymbol{b}_i^T \boldsymbol{b}_j = \underbrace{\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}}_{\textbf{Kronecker delta}}$$

In this case $B = \begin{pmatrix} \boldsymbol{b}_1 & \boldsymbol{b}_2 & \dots & \boldsymbol{b}_D \end{pmatrix}$ is called an **orthonormal matrix**.

Properties of Orthonormal Matrices

Let B be an orthonormal matrix.

- Since $\boldsymbol{b}_i^T \boldsymbol{b}_j = \delta_{ij}$ it holds that

$$B^T B = I$$

- Thus $B^{-1} = B^T$
- Thus also $BB^T = I$

Properties of Orthonormal Matrices cont'd

Multiplication with B corresponds to a **rotation** of a vector.

Hence, multiplication with B (or B^T) leaves the dot product (angle) between two vectors unchanged:

$$(B\boldsymbol{x})^T(B\boldsymbol{y}) = \boldsymbol{x}^T(B^TB)\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y}$$

$$(B^T\boldsymbol{x})^T(B^T\boldsymbol{y}) = \boldsymbol{x}^T(BB^T)\boldsymbol{y} = \boldsymbol{x}^T\boldsymbol{y}$$

Properties of Orthonormal Matrices cont'd

Also the (quadratic) Euclidean norm of a vector is unchanged:

$$\|B\boldsymbol{x}\|_2^2 = (B\boldsymbol{x})^T(B\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$$

$$\|B^T\boldsymbol{x}\|_2^2 = (B^T\boldsymbol{x})^T(B^T\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{x} = \|\boldsymbol{x}\|_2^2$$

Change of Basis for Orthonormal Matrices

Recall from before that for any $\boldsymbol{x} \in \mathbb{R}^D$ as

$$\begin{aligned} \boldsymbol{x} &= B\boldsymbol{z} \\ B^{-1}\boldsymbol{x} &= \boldsymbol{z} \end{aligned}$$

For orthonormal matrices $B^T = B^{-1}$, thus

$$\begin{aligned} \boldsymbol{x} &= B\boldsymbol{z} \\ B^T\boldsymbol{x} &= \boldsymbol{z} \end{aligned}$$

Linear Subspace

So far we considered a **complete** basis:

$$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_D$$

describing the same vector space $\mathbb{R}^D$.

If we use only $K < D$ basis vectors

$$\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K, \cancel{\boldsymbol{b}_{K+1}}, \ldots, \cancel{\boldsymbol{b}_D}$$

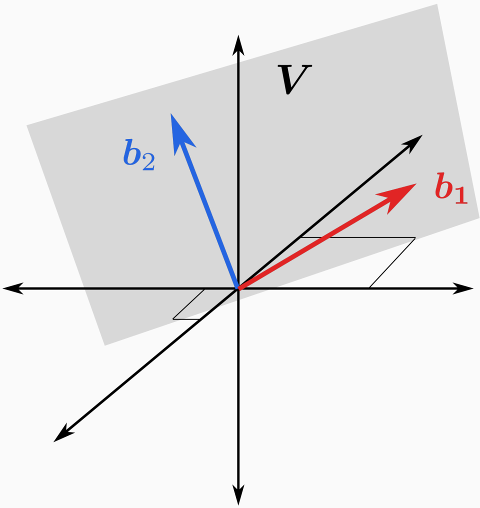
we describe a K-dimensional **linear subspace**.

Let $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots, \boldsymbol{b}_K)$ be a $D \times K$-matrix with orthonormal vectors as columns, where $K < D$.

The **linear subspace** spanned by B is defined as all possible linear combinations of B's columns:

$$\boldsymbol{V} = \left\{ \boldsymbol{v} = B\boldsymbol{z} \mid \boldsymbol{z} = (z_1, \ldots, z_K)^T \in \mathbb{R}^K \right\}$$

Note that $\boldsymbol{V}$ still contains tuples of length D, but it is a K-dimensional space, i.e. in one-to-one correspondence with $\mathbb{R}^K$.



2-dimensional subspace $\boldsymbol{V}$ of $\mathbb{R}^3$ spanned by 2 vectors $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$.

Let $B = (\boldsymbol{b}_1, \boldsymbol{b}_2, \dots, \boldsymbol{b}_K)$ be a matrix with K orthonormal vectors as columns, spanning a K-dimensional subspace $\boldsymbol{V}$ of $\mathbb{R}^D$.

We can **project** an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^D$ onto $\boldsymbol{V}$ by

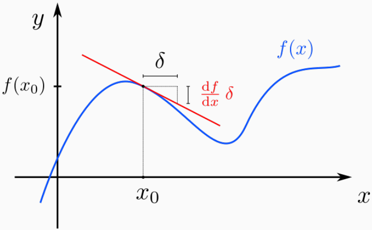
- computing the **projection coefficients** $B^T\boldsymbol{x} =: \boldsymbol{z} \in \mathbb{R}^K$
- computing the **projection**/**reconstruction** $B\boldsymbol{z} =: \hat{\boldsymbol{x}} \in \boldsymbol{V}$
- thus, $\hat{\boldsymbol{x}} = \underbrace{BB^T}_{\textbf{projection matrix}} \boldsymbol{x}$
- $\hat{\boldsymbol{x}}$ is the **closest point** in $\boldsymbol{V}$ (in Euclidean distance) to $\boldsymbol{x}$
- the **residual** $\boldsymbol{r} = \boldsymbol{x} - \hat{\boldsymbol{x}}$ is always orthogonal to $\hat{\boldsymbol{x}}$

Differential Calculus

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a univariate function. If it exists, the **derivative** at a point x is defined as

$$\frac{\mathrm{d}f}{\mathrm{d}x}(x) = f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

- rate of change
- slope of the tangent line at some point x_0, i.e. best local linear approximation: $f(x) \approx f_{linear}(x) = \underbrace{f(x_0)}_{b} + \underbrace{f'(x_0)}_{a} \underbrace{(x - x_0)}_{\delta}$



Standard Rules for Derivatives

Let f and g be differentiable univariate functions and a, b arbitrary constants.

- **Derivative is linear:**

$$(af + bg)'(x) = af'(x) + bg'(x)$$
$$\left(\quad \frac{\mathrm{d}(af + bg)}{\mathrm{d}x} = a\frac{\mathrm{d}f}{\mathrm{d}x} + b\frac{\mathrm{d}g}{\mathrm{d}x} \quad \right)$$

- **Product rule:**

$$(fg)'(x) = f'g(x) + fg'(x)$$
$$\left(\quad \frac{\mathrm{d}fg}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}x}g + \frac{\mathrm{d}g}{\mathrm{d}x}f \quad \right)$$

- **Chain rule:**

$$f(g(x))' = f'(g(x))g'(x)$$
$$\left(\quad \frac{\mathrm{d}f \circ g}{\mathrm{d}x} = \frac{\mathrm{d}f}{\mathrm{d}g}\frac{\mathrm{d}g}{\mathrm{d}x} \quad \right)$$

Note: $f \circ g$ denotes function composition: $(f \circ g)(x) = f(g(x))$

Derivatives of some well-known functions

Name	$f(x)$	$f'(x)$
Constant	a	0
Affine	$ax + b$	a
Polynomial	x^k	$k\,x^{k-1}$
Exponential	$\exp(x),\;\; e^x$	$\exp(x),\;\; e^x$
	b^x	$b^x \log b$
Logarithm	$\log x$	$\frac{1}{x}$
Sine	$\sin(x)$	$\cos(x)$
Cosine	$\cos(x)$	$-sin(x)$

Note: log denotes the **natural logarithm** in this course.