Recap 18.06

Tristan Ozuch

Contents

1 Vectors and linear combinations

Linear Algebra is the study of vectors together with operations on vectors: linear combinations.

1.1 Vector

Mathematically, a vector is an element of a vector space... More on that later.

Geometrically, a vector is characterized by a length or magnitude, and a direction. It is typically represented by an **arrow**. The translation of a vector represents de **same** vector

In coordinates, a vector in $\mathbb{R}^n$ (that is, in dimension n) is given as n numbers called components or coordinates. It may be represented horizontally or vertically as a matrix:

$$\vec{x} = (x_1, x_2, \dots, x_n) = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}.$$

1.2 Addition and multiplication by a factor

Geometrically, adding vectors means concatenating the arrows.

In coordinates, this means adding the coordinates. For instance, if $\vec{u} = (u_1, u_2, \dots, u_n)$ and $\vec{v} = (v_1, v_2, \dots, v_n)$, then one has:

$$\vec{u} + \vec{v} = \begin{bmatrix} u_1 + v_1 \\ u_2 + v_2 \\ \dots \\ u_n + v_n \end{bmatrix}.$$

Geometrically, multiplying by a factor (or scalar, or number) $c \in \mathbb{R}$ means multiplying the length by $|c|$, and keeping the same direction if $c \geqslant 0$ or considering its opposite if $c < 0$.

In coordinates, this means multiplying all of the coordinates by the factor c. For instance, one has:

$$c\vec{u} = \begin{bmatrix} cu_1 \\ cu_2 \\ \dots \\ cu_n \end{bmatrix}.$$

1.3 Linear combination and span

A **linear combination** of $\vec{x}_1, \dots, \vec{x}_k$ (note that k has nothing to do with the dimension!) is a vector of the form

$$c_1\vec{x}_1 + \dots + c_k\vec{x}_k, \text{ where } c_1, \dots, c_k \in \mathbb{R}.$$

The **linear span** of vectors $\vec{x}_1, \dots, \vec{x}_k$ is the *set* of linear combinations of $\vec{x}_1, \dots, \vec{x}_k$. Our notation is going to be

$$\text{Span}(\vec{x}_1, \dots, \vec{x}_k) = \left\{ c_1\vec{x}_1 + \dots + c_k\vec{x}_k, \;\; c_1, \dots, c_k \in \mathbb{R} \right\}.$$

If a family $\vec{x}_1, \dots, \vec{x}_k$ spans the whole space $\mathbb{R}^n$ (i.e. $\text{Span}(\vec{x}_1, \dots, \vec{x}_k) = \mathbb{R}^n$), then one has $k \geqslant n$.

2 Linearly independent vectors and bases

2.1 Linear independent families of vectors

We say that a family of vectors $\vec{x}_1, \ldots, \vec{x}_k$ is **linearly independent** if the only vanishing linear combination of $\vec{x}_1, \ldots, \vec{x}_k$ is the trivial one with zero coefficients. In more mathematical terms: if

$$c_1\vec{x}_1 + \cdots + c_k\vec{x}_k = 0 \qquad \Longrightarrow \qquad c_1 = \cdots = c_k = 0.$$

If a family $\vec{x}_1, \ldots, \vec{x}_k$ in $\mathbb{R}^n$ is linearly independent, then one has $k \leqslant n$.

2.2 Bases and dimension

A **basis** of $\mathbb{R}^n$ is a family of vectors $\vec{x}_1, \ldots, \vec{x}_k$ which is linearly independent and spans $\mathbb{R}^n$. It implies that $k = n$ – *the number of vectors in a basis is always equal to the dimension*.

Equivalent definitions are: $\vec{x}_1, \ldots, \vec{x}_k$ is a basis if one of the following equivalent properties is satisfied

1. $k = n$ and the family is linearly independent (typically the most important/useful),
2. $k = n$ and the family spans $\mathbb{R}^n$.

3 Dot product

The **dot product** of $\vec{u} = (u_1, \ldots, u_n)$ and $\vec{v} = (v_1, \ldots, v_n)$ is:

$$\vec{u} \cdot \vec{v} = u_1v_1 + u_2v_2 + \cdots + u_nv_n.$$

The **length** of a vector $\vec{u}$ is:

$$\|\vec{u}\| = \sqrt{\vec{u} \cdot \vec{u}}.$$

Denoting θ the **angle** between $\vec{u}$ and $\vec{v}$, we have the relation:

$$\vec{u} \cdot \vec{v} = \|\vec{u}\|\|\vec{v}\| \cos\theta,$$

which leads to the inequality

$$\left|\vec{u} \cdot \vec{v}\right| \leqslant \|\vec{u}\|\|\vec{v}\|.$$

Two vectors $\vec{u}$ and $\vec{v}$ are said to be **orthogonal** or **perpendicular** if:

$$\vec{u} \cdot \vec{v} = 0,$$

or equivalently $\theta = \frac{\pi}{2}$.

Another important inequality is the **triangle inequality**:

$$\|\vec{u} + \vec{v}\| \ \leqslant \|\vec{u}\| + \|\vec{v}\|.$$

4 Operations on matrices

A **matrix** is a rectangular array of number. A $p \times q$ matrix has p **rows** and q **columns**.

We may write $p \times q$ matrices "in components" as $M = (m_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}}$ corresponding to the array of numbers:

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1q} \\ m_{21} & m_{22} & \dots & m_{2q} \\ \dots & \dots & \dots & \dots \\ m_{p1} & m_{p2} & \dots & m_{pq} \end{bmatrix}.$$

A **square matrix** has $p = q$.

A **vector** only has one column, it is a $p \times 1$ matrix.

4.1 Matrix addition and multiplication by a factor

The **sum of two matrices** is the sum of its components:

$$(a_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}} + (b_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}} = (a_{ij} + b_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}},$$

Similarly, one has for any $c \in \mathbb{R}$: $c\ (a_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}} = (c\ a_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}}$.

4.2 Matrix-vector and matrix-matrix multiplication

In order to multiply a matrix $p \times q$ and a $r \times s$ matrix, one **must have** $q = r$.

4.2.1 Matrix-vector multiplication

In coordinates, the (left) multiplication of $\vec{x} = (x_1, \dots, x_q)$ by the matrix $M = (m_{ij})_{i\in\{1,\dots,p\},j\in\{1,\dots,q\}}$ is the following vector of $\mathbb{R}^p$ (or $p \times 1$ matrix)

$$M\vec{x} = \Big(\sum_{j=1}^{q} m_{1j}x_j\ ,\ \sum_{j=1}^{q} m_{2j}x_j\ ,\ \dots\ ,\ \sum_{j=1}^{q} m_{pj}x_j \Big) \in \mathbb{R}^p.$$

If one considers the **rows** of M as vectors in $\mathbb{R}^q$: $\vec{m}_i = (m_{i1}, \dots, m_{iq})$, one ca rewrite:

$$M\vec{x} = \Big(\vec{m}_1 \cdot \vec{x}\ ,\ \vec{m}_2 \cdot \vec{x}\ ,\ \dots\ ,\ \vec{m}_p \cdot \vec{x} \Big) \in \mathbb{R}^p.$$

If on the other hand, one considers the **columns** of M as vectors of $\mathbb{R}^p$, $\vec{\mu}_j = (m_{1j}, m_{2j}, \dots, m_{pj})$, one may also rewrite:

$$M\vec{x} = x_1\vec{\mu}_1 + x_2\vec{\mu}_2 + \dots + x_q\vec{\mu}_q \in \mathbb{R}^p.$$

In conclusion, it is useful to notice that a matrix can be written as the set of its columns or as the set of its rows.

$$M = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1q} \\ m_{21} & m_{22} & \dots & m_{2q} \\ \dots & \dots & \dots & \dots \\ m_{p1} & m_{p2} & \dots & m_{pq} \end{bmatrix} = \begin{bmatrix} \vec{m}_1^T \\ \vec{m}_2^T \\ \dots \\ \vec{m}_p^T \end{bmatrix} = \begin{bmatrix} \vec{\mu}_1 & \vec{\mu}_2 & \dots & \vec{\mu}_q \end{bmatrix}.$$

where $\vec{m}_i^T$ will be defined very soon as the "*transpose*" of the vector $\vec{m}_i$, that is the matrix composed of one row whose coefficients are the coefficients of $\vec{m}_i$. The $\vec{m}_i^T$ are the rows of M, and the $\vec{\mu}_j$ are the columns of M.

4.2.2 Matrix-matrix multiplication

The product of a $p \times q$ matrix and a $q \times r$ matrix is a $p \times r$ matrix. Denote c_{ij} the $(i, j) \in \{1, \dots, p\} \times \{1, \dots, r\}$ component of the product of $(a_{ij})_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant q}$ and $(b_{ij})_{1 \leqslant i \leqslant q, 1 \leqslant j \leqslant r}$, i.e. $C = AB$ with obvious notations. Then, the coefficients of the product are given by:

$$c_{ij} = \sum_k a_{ik} b_{kj}.$$

Each column of C is the matrix-vector product of A with the corresponding column of B.

Note: Even if A and B are square, we generally do not have $AB = BA$.

4.3 Matrices, linear combinations and linear systems

A **linear system of equations** of size $p \times q$ is generally of the form

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q = b_2 \\ \quad \cdots \qquad\quad \cdots \qquad\quad \cdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pq}x_q = b_p. \end{cases}$$

This corresponds exactly to the equation

$$A\vec{x} = \vec{b},$$

for $A = (a_{ij})_{1 \leqslant i \leqslant p, 1 \leqslant j \leqslant q}$, $\vec{x} = (x_1, \dots, x_q)$, and $\vec{b} = (b_1, \dots, b_p)$.

Consider the matrix A whose columns are vectors $\vec{\alpha}_j$:

$$A = \begin{bmatrix} \vec{\alpha}_1 & \vec{\alpha}_2 & \dots & \vec{\alpha}_q \end{bmatrix},$$

and the vector

$$\vec{x} = (x_1, \dots, x_q).$$

As seen above, the linear combination with coefficients $x_1, \dots, x_q$ of the vectors $\vec{\alpha}_1, \dots, \vec{\alpha}_q$ can be written in matrix form as

$$x_1\vec{\alpha}_1 + x_2\vec{\alpha}_2 + \cdots + x_q\vec{\alpha}_q = A\vec{x}.$$

The question of whether or not there **exists** a solution to the linear system is exactly the question of whether or not $\vec{b} \in \text{Span}(\vec{\alpha}_1, \dots, \vec{\alpha}_q)$. The **uniqueness** is exactly the question of the linear independence of $(\vec{\alpha}_1, \dots, \vec{\alpha}_q)$. More precisely, the linear independence of $(\vec{\alpha}_1, \dots, \vec{\alpha}_q)$ is equivalent to the statement that the only solution $\vec{x}$ of $A\vec{x} = \vec{0}$ is $\vec{x} = \vec{0}$.

5 Solving linear systems

A streamlined method for solving linear systems is the so-called (Gaussian) *elimination* or *elimination*.

5.1 Linear system of equations

A linear systems of equations is of the form:

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q = b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q = b_2 \\ \quad \cdots \qquad\quad \cdots \qquad\quad \cdots \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pq}x_q = b_p. \end{cases}$$

The a_{ij} and b_i are given, while the x_j are the unknowns that we search for.

The situation is best when you have as many equations as unknowns, i.e. $p = q$. We will see why later, but typically, to expect a solution, one needs $p \geqslant q$. "Forgetting" some equations, you can recover the $p = q$ situation.

5.1.1 In dimension 2

When $p = q = 2$, linear systems are of the form

$$\begin{cases} a_{11}x + a_{12}y = b_1 \\ a_{21}x + a_{22}y = b_2. \end{cases}$$

Row picture: Both equations are equations of lines. We are therefore looking for the the intersection of two lines in the plane. There is typically a **unique solution**.

However, if the two lines are **parallel**, the intersection is either empty or the whole line: there are either no solutions or there are infinitely many of them.

Column picture: Denoting $\vec{\alpha}_1 = (a_{11}, a_{21})$, $\vec{\alpha}_2 = (a_{12}, a_{22})$ and $\vec{b} = (b_1, b_2)$, the system of equations can be written as:

$$x\vec{\alpha}_1 + y\vec{\alpha}_2 = \vec{b}.$$

We are looking for a **linear combination** of the columns that reaches $\vec{b}$. The coefficients of the linear combination are our unknowns x and y.

5.1.2 In higher dimensions

In higher dimensions, the row and column pictures still hold. One either looks for the intersection of p hyperplanes in $\mathbb{R}^q$, or for a linear combination with q terms of vectors in $\mathbb{R}^p$.

The geometric interpretation however becomes less and less useful as the dimensions grow. We need a more systematic and even algorithmic method.

5.2 Solving linear systems of equations

5.2.1 Gaussian elimination

The elimination method is algorithmic and iterative. We want to reduce a $p \times p$ system to a $(p-1) \times (p-1)$ system and iterate to find a $(p-2) \times (p-2)$ system... until we reach a 1×1 system that you all know how to solve.

Let us illustrate the general process on

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1q}x_q = b_1 & (R_1) \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2q}x_q = b_2 & (R_2) \\ \quad \cdots \qquad\quad \cdots \qquad\quad \cdots & \\ a_{p1}x_1 + a_{p2}x_2 + \cdots + a_{pq}x_q = b_p. & (R_p) \end{cases}$$

1. if $a_{11} \neq 0$, we call it the **first pivot** and will use it to get rid of the variable x_1 in every other row.
2. for $i \geqslant 2$, replace (R_i) by $(R_i') = (R_i) - l_{i1}R_1$, where $l_{i1} = \frac{a_{i1}}{a_{11}}$ is the ith multiplier.
3. This gets rid of every x_1 except in the first equation. The (equivalent) system now looks like:

$$\begin{cases} a_{11}x_1 \quad + a_{12}x_2 + \cdots + a_{1q}x_q = b_1 & (R_1) \\ \qquad\quad + a'_{22}x_2 + \cdots + a'_{2q}x_q = b'_2 & (R'_2) \\ \qquad \cdots \qquad\quad \cdots \qquad\quad \cdots & \\ \qquad\quad + a'_{p2}x_2 + \cdots + a'_{pq}x_q = b'_p. & (R'_p) \end{cases}$$

Retaining only the last $p-1$ lines, we are now looking at a $(p-1) \times (p-1)$ system whose pivot is a'_{22} if it does not vanish. By iterating this process, we obtain an upper triangular system that can be solved by back substitution.

5.2.2 Matrix picture of Gaussian elimination

The key to this method is that once the system is **upper triangular**, everything becomes easy to solve.

All of the above operations can be written as matrix multiplications on the left. These matrix multiplications transform the initial matrix A into an **upper triangular matrix** U (i.e. with $u_{ij} = 0$ if $i > j$), and the vector $\vec{b}$ into another vector $\vec{c}$.

5.2.3 When the method fails

The method fails when one **pivot vanishes**. There may be a fix to this: for instance if $a_{11} = 0$ but $a_{i1} \neq 0$, we may exchange rows (R_1) and (R_j) and apply the same method to the new system.

If all of the a_{i1} vanish however, there is no fix and the method is typically doomed. This is exactly where there are either **no solutions** or **infinitely many**. Think of the intersection of two lines!

6 Elimination as matrix multiplications

Recall that the matrix multiplication of $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{q\times r}$ is the matrix whose columns are the products of A and the columns of B.

An important property is the **associativity** $A(BC) = (AB)C$. In general, we do not have **commutation**: generally $AB \neq BA$!!

6.1 The identity matrix

The identity matrix $\mathrm{I}_p \in \mathbb{R}^{p\times p}$ is the **neutral** element for the matrix multiplication: for any matrix $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{p\times q}$, we have:

$$A\,\mathrm{I}_p = A \text{ and } \mathrm{I}_p\,B = B,$$

in particular, if $p = n$, $A\,\mathrm{I}_p = \mathrm{I}_p\,A = A$.

The identity matrix is **diagonal**, i.e. its off-diagonal coefficients for $i \neq j$ vanish. Its diagonal coefficients are all 1:

$$\mathrm{I}_p = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 1 \end{bmatrix}.$$

6.2 Elimination and matrix operations

In the elimination process, we saw two types of **elementary operations**: linear combinations of two rows, and exchanging two rows.

6.2.1 Canonical basis for matrices

$\mathbb{R}^n$ has a canonical basis $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, 0 \dots, 0)$, ..., $e_n = (0, 0, \dots, 0, 1)$, i.e. e_i is the vector with only zero coordinates, except 1 at the i-th component.

Similarly, the space of matrices $\mathbb{R}^{p\times q}$ also has a canonical basis. Its elements are E_{ij}, the matrix with only zeros except at the (i, j) component. For instance, E_{32} in $\mathbb{R}^{p\times q}$ has the form

$$E_{32} = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & 0 \end{bmatrix}.$$

Multiplying A on the left by E_{ij} results in the matrix with only zeros, except on row i where the j-th row of A is copied-pasted.

Note: This E_{ij} is a widespread notation. In the book we follow it means something different: $\mathrm{I}_p - l_{ij}E_{ij}$

6.2.2 Linear combination of rows as a matrix operation

One of the elementary operations consisted in substracting l_{ij} times the row j to the row i. This can be written as a matrix multiplication: the left multiplication by

$$\mathrm{I}_p - l_{ij}E_{ij}.$$

Indeed, I_p just repeats the same matrix while $-l_{ij}E_{ij}$ substracts l_{ij} times the row j to the row i.

6.2.3 Permutation as a matrix operation

Another important operation is the permutation of two rows. This can be obtained by the left multiplication by the **permutation matrix** P_{ij} equal to the identity on which one exchanges rows i and j.

7 Inverse matrices

In this section, we only consider **square** matrices $p \times p$.

7.1 Some properties of matrix operations

$$A + B = B + A \qquad c(A + B) = cA + cB$$
$$A + (B + C) = (A + B) + c \qquad C(A + B) = CA + CB$$
$$(A + B)C = AC + BC \qquad A(BC) = (AB)C.$$

7.2 Powers of a square matrix

We define $A^0 = \mathrm{I}_p$ and for $k, l \geqslant 1$, one has:

$$A^k = \underbrace{AA \dots A}_{k \text{ times}},$$

and the formulas:

$$(A^k)^l = A^{kl} \qquad A^k A^l = A^{k+l} = A^l A^k.$$

7.3 Negative powers and inverse

A matrix $A \in \mathbb{R}^{p \times p}$ is **invertible** if it has an **inverse**, i.e., there exists $B \in \mathbb{R}^{p \times p}$ such that:

$$AB = BA = \mathrm{I}_p\,.$$

Such a B is **unique** when it exists and we will denote it A^{-1}.

A^{-1} can be computed by elimination and exists if and only if the pivots do not vanish (up to exchanging rows).
The typical picture is to start by the $p \times 2p$ matrix consisting of A and I_p and operate by left multiplication in order to "eliminate" terms on the left matrix. At the end of the process, one is left with the $p \times 2p$ matrix consisting of I_p and A^{-1}.

If A is invertible and one tries to solve the linear system $A\vec{x} = \vec{b}$, then the **unique** solution is $\vec{x} = A^{-1}\vec{b}$.

A $p \times p$ matrix is invertible if and only if the **only** solution of $A\vec{x} = \vec{0}$ is $\vec{x} = \vec{0}$.

A $p \times p$ matrix is invertible if and only if its columns are **linearly independent**.

7.4 Properties and formulas for the inverse

Some properties:

$$(AB)^{-1} = B^{-1}A^{-1} \qquad (ABC)^{-1} = C^{-1}B^{-1}A^{-1}.$$

For a 2×2 matrix $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, and when $ad - bc \neq 0$, one has:

$$A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

8 LU factorization of matrices

A computationally useful decomposition of matrices is the LU or sometimes LDU decomposition.

8.1 Triangular matrices

Upper (resp. lower) triangular matrices only have zero coefficients below (resp. above) the diagonal.

A **product** of upper (resp. lower) triangular matrices is an upper (resp. lower) triangular matrices.

An upper (resp. lower) triangular matrices is **invertible** if and only if all of its diagonal terms are non zero. And if it is the case, their **inverses** are also upper (resp. lower) triangular matrices.

8.2 Elimination and LU decomposition

The matrices appearing in the process of **elimination** $\mathrm{I}_p - l_{ij}E_{ij}$ with $i > j$ (resp. with $i < j$) are lower (resp. upper) triangular matrices. Their **inverses** are simply $\mathrm{I}_p + l_{ij}E_{ij}$.

Let us forget about **permutations** (which are almost never needed). From a matrix point of view, the elimination process for a linear system can be written as:

$$(\mathrm{I}_p - l_{p,p-1}E_{p,p-1}) \cdots (\mathrm{I}_p - l_{31}E_{31})\ (\mathrm{I}_p - l_{21}E_{21})A\vec{x} = U\vec{x}$$

for an upper triangular matrix U. Now the observation is that $(\mathrm{I}_p - l_{p,p-1}E_{p,p-1}) \cdots (\mathrm{I}_p - l_{31}E_{31})\ (\mathrm{I}_p - l_{21}E_{21})$ is *lower* triangular so its inverse

$$\begin{aligned} L &= \left[(\mathrm{I}_p - l_{p,p-1}E_{p,p-1}) \cdots (\mathrm{I}_p - l_{31}E_{31})\ (\mathrm{I}_p - l_{21}E_{21})\right]^{-1} \\ &= (\mathrm{I}_p - l_{21}E_{21})^{-1}\ (\mathrm{I}_p - l_{31}E_{31})^{-1} \cdots (\mathrm{I}_p - l_{p,p-1}E_{p,p-1})^{-1} \\ &= (\mathrm{I}_p + l_{21}E_{21})\ (\mathrm{I}_p + l_{31}E_{31}) \cdots (\mathrm{I}_p - l_{p,p-1}E_{p,p-1}), \end{aligned}$$

is also lower triangular and easy to compute.

The **LU decomposition** of A is $A = LU$.

8.3 Use of the LU decomposition

Once you know that $A = LU$ and you want to solve the system $A\vec{x} = \vec{b}$, there is a trick: introduce an undetermined $\vec{c}$ and solve the two successive triangular (hence easy to solve by back substitution!) systems:

1. $L\vec{c} = \vec{b}$,
2. $U\vec{x} = \vec{c}$.

The vector $\vec{x}$ is then solution to the equation $A\vec{x} = \vec{b}$.

9 Linear systems and uniqueness of the solution

9.1 A simple observation

If a linear system $A\vec{x} = \vec{b}$ has two solutions $\vec{x}_1$ and $\vec{x}_2$, then $\vec{y} = \vec{x}_1 - \vec{x}_2$ is a solution of $A\vec{y} = \vec{0}$.
We deduce:

1. the **uniqueness** of the solution is independent of $\vec{b}$,
2. it is equivalent to: $A\vec{y} = \vec{0}$ if and only if $\vec{y} = \vec{0}$, and
3. it is also equivalent to: the columns of A are linearly independent.

9.2 The nullspace or kernel

The **nullspace** (or **kernel**) of a matrix A is the set of vectors $\vec{y}$ satisfying $A\vec{y} = \vec{0}$. If A is a $p \times q$ matrix, we will denote it:

$$N(A) = \left\{ y,\ A\vec{y} = \vec{0} \right\} \in \mathbb{R}^q.$$

Some common other notations are Null(A) and in mathematics, the most common would be ker(A) for "kernel".

If there exists x_0 such that $A\vec{x}_0 = \vec{b}$ he set of solutions of $A\vec{x} = \vec{b}$ is $\vec{x}_0 + N(A)$, that is the set of vectors $\vec{x} = \vec{x} + \vec{y}$ for $\vec{y} \in N(A)$. $\vec{x}_0$ is a **particular solution**, and $N(A)$ is the **direction** of the space of solutions.

In particular, the solution $\vec{x}_0$ is **unique** if and only if $N(A) = \{\vec{0}\}$.

If $A \in \mathbb{R}^{p \times q}$ with $q > p$ ("short and wide"), then $N(A) = \{\vec{0}\}$ is **never** satisfied, we will see that $N(A)$ is at least $(q-p)$-dimensional. On the other hand, if $q \leqslant p$ ("tall and thin"), then *generically* (almost surely), one has $N(A) = \{\vec{0}\}$.

9.3 Computing the nullspace

Elimination is once again the first technique to solve equations $A\vec{y} = \vec{0}$. This time, in particular if $q > p$, we need an extension of triangular matrices, and they are **echelon matrices**.

An **echelon matrix** is a rectangular matrix transformed such that all zero rows are at the bottom and each leading entry of a nonzero row is to the right of the leading entry of the previous row. The **pivots** are these nonzero leading entries on each row.

On an echelon matrix, if $\vec{y} = (y_1, ..., y_q)$ we distinguish **pivot variables** as the y_j such that the *column j* has a pivot, and the other components of $\vec{y}$ are called **free variables**.

The free variables can be chosen freely and the pivots will be determined by these choices. The free variables parametrize the set of solutions (when it is not empty): the **number of free variables** is the **dimension** of $N(A)$.

The solutions are vectors whose coordinates are linear functions of the free variables.

10 Criterion for solving a linear system

An even more important question than that of uniqueness, is that of the **existence** of solutions to your equations.

10.1 The column space of a matrix (or its image, or range)

The **column space** of a matrix $A \in \mathbb{R}^{p \times q}$ (or "image of A" or "range of A") is the span of its column vectors: the set of linear combinations of its column vectors. It is denoted $C(A)$ in our textbook, common other notations are Im(A) or $A(\mathbb{R}^q)$. It is a subset of $\mathbb{R}^p$.

An equation $A\vec{x} = \vec{b}$ **admits a solution** if and only if $\vec{b} \in C(A)$.

10.2 Computing the column space

The column space cannot be computed as directly as the nullspace of a matrix, but its dimension is actually the number of pivots. We call it the **rank** of a matrix A. We will see that this number is at the core of our structure theorem about the set of solutions of a linear system.

Elimination on a matrix $A \in \mathbb{R}^{p\times q}$ consists in multiplying the matrix A from the **left** by elementary $p \times p$ matrices. The result (with notations compatible with the LU decomposition) is

$$L^{-1}A = U, \text{ an echelon matrix.}$$

Note that U is not quite triangular, but just an **echelon** matrix.

As we saw, the nullspace of A is equal to the nullspace of the echelon matrix $U = L^{-1}A$. The column space of A is $C(A) = L(C(U))$. More precisely, if $U = \begin{bmatrix}\vec{u}_1 & \vec{u}_2 & \dots & \vec{u}_q\end{bmatrix}$ is decomposed in column vectors, then one has

$$C(U) = \text{Span}(\vec{u}_1, \vec{u}_2, \dots, \vec{u}_q), \text{ and}$$

$$C(A) = \text{Span}(L(\vec{u}_1), L(\vec{u}_2), \dots, L(\vec{u}_q)).$$

One can then reduce this further: $C(U)$ is also the span of the **pivot** columns only, and $C(A)$ the span of the $L(\vec{u}_i)$ for $\vec{u}_i$ a pivot vector. The $L(\vec{u}_i)$ form a basis of $C(A)$.

10.3 First part of the "fundamental theorem of linear algebra"

We have seen that the sum of number of pivot columns and the number of free columns was always q, the number of columns. This leads to the following central formula:

Theorem 10.1.

$$q = \dim N(A) + \dim C(A).$$

This is the first part of what is sometimes called the **fundamental theorem of linear algebra**.

11 Vector spaces and linear maps

It is useful to think about our vector operations and matrices more abstractly and maybe more geometrically.

11.1 Vector (Sub)spaces

The full definition of a **vector space** is tedious to verify. For us, it will typically reduce to recalling the properties of the addition and multiplication by a scalar. I copy it below if you are curious: A vector space V over the real numbers $\mathbb{R}$ must satisfy the following axioms:

Closure Axioms

1. *Closure under Addition*: $\forall \vec{u}, \vec{v} \in V, \vec{u} + \vec{v} \in V$
2. *Closure under Scalar Multiplication*: $\forall \vec{v} \in V, \forall c \in \mathbb{R}, c\vec{v} \in V$

Arithmetic Axioms

3. *Associativity of Addition*: $(\vec{u} + \vec{v}) + \vec{w} = \vec{u} + (\vec{v} + \vec{w})$
4. *Commutativity of Addition*: $\vec{u} + \vec{v} = \vec{v} + \vec{u}$
5. *Identity Element of Addition*: $\exists \vec{0} \in V \,:\, \forall \vec{v} \in V, \vec{v} + \vec{0} = \vec{v}$
6. *Inverse Elements of Addition*: $\forall \vec{v} \in V, \exists -\vec{v} \in V \,:\, \vec{v} + (-\vec{v}) = \vec{0}$
7. *Compatibility of Scalar Multiplication with Scalar Addition*: $\forall c, d \in \mathbb{R}, \forall \vec{v} \in V, (c + d)\vec{v} = c\vec{v} + d\vec{v}$

8. *Compatibility of Scalar Multiplication with Scalar Multiplication*: $\forall c, d \in \mathbb{R}, \forall \vec{v} \in V, (cd)\vec{v} = c(d\vec{v})$
9. *Identity Element of Scalar Multiplication*: $\forall \vec{v} \in V, 1\vec{v} = \vec{v}$
10. *Distributivity of Scalar Multiplication with respect to Vector Addition*: $\forall c \in \mathbb{R}, \forall \vec{u}, \vec{v} \in V, c(\vec{u} + \vec{v}) = c\vec{u} + c\vec{v}$

We will instead mostly use the definition of **vector subspace**. Indeed, it is much easier to verify that a subset is a vector space when it is already included in a vector space: the main properties of the addition and multiplication by a scalar are already verified!

Let W be a vector space. A **vector subspace** $V \subset W$ is a set that satisfies: for all $\vec{v}_1, \vec{v}_2 \in V$ and all $c \in \mathbb{R}$,

1. $\vec{0} \in V$,
2. $\vec{v}_1 + \vec{v}_2 \in V$, and
3. $c\vec{v}_1 \in V$

(where $\vec{0}$, the addition and multiplication are those of W). In other words, the subspace is a vector subspace is stable by linear combinations. The first axiom is a condition of the last one if the set is not empty, but it is very easy to forget it! For instance, the line of equation $3x + 4y = 1$ is **not** a vector subspace because $(0, 0)$ is not inside it. This should always be your first test to verify that a subset is a vector space.

The key point is: a **vector subspace** is **a vector space**. It just happens to be inside another.

Let $A \in \mathbb{R}^{p \times q}$, then the **nullspace** $N(A)$ is a vector subspace of $\mathbb{R}^q$, and the column space $C(A)$ is a vector subspace of $\mathbb{R}^p$.

The **dimension** of a vector space is the number of a vector in a basis. It is the largest possible number of vectors in a linearly independent family, and the smallest possible number of vectors required to span the whole space.

11.2 Linear maps

We call **linear maps** the operations from one vector space to another that respect linear combinations.

A map $f : V \to W$ between vector spaces is linear if for all $\vec{v}_1, \vec{v}_2 \in V$ and $c \in \mathbb{R}$, one has:

1. $f(\vec{v}_1 + \vec{v}_2) = f(\vec{v}_1) + f(\vec{v}_2)$, and
2. $f(c\vec{v}_1) = cf(\vec{v}_1)$.

This is satisfied by many operations you have learned about in calculus for instance. Another key example is the following one: let $A \in \mathbb{R}^{p \times q}$. Then, the following map is linear:

$$\begin{array}{rccc} f : & \mathbb{R}^q & \to & \mathbb{R}^p \\ & \vec{x} & \mapsto & A\vec{x}. \end{array}$$

11.3 Linear maps and matrices

It turns out that any linear map between *finite dimensional* vector spaces can be represented by a matrix.

11.3.1 From n-dimensional vector spaces to $\mathbb{R}^n$

Any basis of a vector space provides an **identification** with $\mathbb{R}^n$ as follows. Let V be a vector space, and $(\vec{v}_1, \dots, \vec{v}_n)$ be a basis of V. Then, there is a direct one-to-one correspondence between vectors in V and elements of $\mathbb{R}^n$:

$$\begin{array}{ccc} V & \leftrightarrow & \mathbb{R}^n \\ x_1\vec{v}_1 + \cdots + x_n\vec{v}_n & \tilde{\mapsto} & (x_1, \dots, x_n). \end{array}$$

This crucially tells us that *intrinsically*, all finite-dimensional vector spaces are the same and that understanding $\mathbb{R}^n$ is enough to understand them all!

11.3.2 Matrices associated to a linear map

This correspondence extends to matrices.

To define the matrix associated to a linear map, we need to choose a **basis** in the starting space, and a **basis** in the target space. Consider $f : V \to W$, and bases $(\vec{v}_1, \dots, \vec{v}_q)$ of V, and $(\vec{w}_1, \dots, \vec{w}_p)$ of W. Then, the **matrix** A **associated to** f **in the bases** $(\vec{v}_1, \dots, \vec{v}_q)$, and $(\vec{w}_1, \dots, \vec{w}_p)$ is obtained as follows:

1. compute $f(\vec{v}_1), \dots, f(\vec{v}_q)$,
2. express them as linear combinations of the $\vec{w}_i$: $f(\vec{v}_j) = a_{1j}\vec{w}_1 + a_{2j}\vec{w}_2 + \cdots + a_{pj}\vec{w}_p$, and
3. define $A \in \mathbb{R}^{p\times q}$ by $A = (a_{ij})_{1\leqslant i\leqslant p, 1\leqslant j\leqslant q}$.

The point is that now, if $\vec{x} \in V$ is equal to $x_1\vec{v}_1 + \cdots + x_q\vec{v}_q$, then $A(x_1, \dots, x_q) = (y_1, \dots, y_p) \in \mathbb{R}^p$ lets us write

$$f(\vec{x}) = y_1\vec{w}_1 + y_2\vec{w}_2 + \cdots + y_p\vec{w}_p.$$

12 Linear maps and special operations

Having a more abstract approach to matrices lets one understand more conceptually what elimination and other operations are really doing and why they work.

12.1 Image and kernel/nullspace of maps

Let $f : V \to W$ be a linear map. We define the **image** of f as

$$\mathrm{Im}(f) = f(V) = \{f(v), v \in V\} \subset W.$$

Note that for $f : \vec{x} \mapsto A\vec{x}$, $\mathrm{Im}(f) = C(A)$, the column space.

We similarly define the **nullspace** or **kernel** of f as

$$\ker(f) = f^{-1}(\{0\}) = \{f(v), v \in V\} \subset W.$$

I am not a fan of the notation $f^{-1}(\{0\})$, f might not be invertible, hence f^{-1} may not make sense – this is just a notation. For $f : \vec{x} \mapsto A\vec{x}$, $\mathrm{Ker}(f) = N(A)$, the nullspace.

12.2 Changes of bases and changes of matrices

The process of elimination consisted of **left** multiplications by matrices. We may rephrase this as looking at the same linear map but **in different bases**.

Denote $(\vec{v}_1, ..., \vec{v}_q)$ and $(\vec{w}_1, ..., \vec{w}_p)$. The map $f : \vec{x} \to A\vec{x}$ has the matrix $A \in \mathbb{R}^{p\times q}$ between $(\vec{v}_1, ..., \vec{v}_q)$ and $(\vec{w}_1, ..., \vec{w}_p)$. But between the bases $(\vec{v}_1, ..., \vec{v}_q)$ and $(L\vec{w}_1, ..., L\vec{w}_p)$ for some $L \in \mathbb{R}^{p\times p}$, its matrix is $L^{-1}A$.

Note in particular that if $A \in \mathbb{R}^{p\times p}$ is invertible, then the matrix of A in the bases $(\vec{v}_1, ..., \vec{v}_q)$ and $(A\vec{v}_1, ..., A\vec{v}_p)$ is the identity.

Elimination is a **change of basis** in the target space. Most matrix reductions we will encounter will be exactly about choosing convenient bases in which to look at our linear maps.

A key point is that the spaces $\mathrm{Im}(f)$ and $\ker(f)$ do not depend on choices of bases. One can therefore choose the most convenient bases for them.

13 Inner product, dot-product, adjoint transpose

An **inner product** is a generalization of the dot-product to more general vector spaces. Formally, an *inner product* on V is a function $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ that satisfies the following properties:

1. *Linearity in the first argument:* $\langle a\vec{u} + b\vec{v}, \vec{w}\rangle = a\langle\vec{u}, \vec{w}\rangle + b\langle\vec{v}, \vec{w}\rangle$ for all $\vec{u}, \vec{v}, \vec{w} \in V$ and $a, b \in \mathbb{R}$.
2. *Symmetry:* $\langle\vec{u}, \vec{v}\rangle = \langle\vec{v}, \vec{u}\rangle$ for all $\vec{u}, \vec{v} \in V$.
3. *Positivity:* $\langle\vec{u}, \vec{u}\rangle \geq 0$ for all $\vec{u} \in V$.
4. *Definiteness:* $\langle\vec{u}, \vec{u}\rangle = 0 \iff \vec{u} = \vec{0}$.

When a vector space is equipped with an inner product, it is commonly called **Euclidean**. This gives additional structure to linear maps and matrices.

Our main examples of such spaces will be: $\mathbb{R}^n$ with the usual dot-product, and $\mathbb{R}^{p\times q}$ with the operation

$$\langle A, B\rangle_{\mathbb{R}^{p\times q}} = \mathrm{tr}(A^T B) = \mathrm{tr}(A^T B) = \sum_{ij} A_{ij} B_{ij},$$

where the tr operation is the sum of the diagonal terms.

We may then measure the **length** of a vector $\vec{x}$ as $\sqrt{\langle\vec{x}, \vec{x}\rangle}$.

13.1 Orthonormal bases

An **orthogonal basis** $(\vec{v}_1, \ldots, \vec{v}_n)$ satisfies for all $i \neq j \in \{1, \ldots, n\}$,

$$\langle\vec{v}_i, \vec{v}_j\rangle = 0,$$

it is **orthonormal** if it additionally satisfies $\langle\vec{v}_i, \vec{v}_i\rangle = 1$.

Orthonormal bases are the most convenient ones to compute dot-products: if $\vec{x}$ and $\vec{y}$ respectively have coordinates $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$ in an *orthonormal* basis $(\vec{v}_1, \ldots, \vec{v}_n)$, then

$$\langle\vec{x}, \vec{y}\rangle = x_1 y_1 + \cdots + x_n y_n.$$

13.2 Adjoint of a linear map and transpose of a matrix

Consider $f : V \to W$ where $(V, \langle\cdot,\cdot\rangle_V)$ and $(W, \langle\cdot,\cdot\rangle_W)$ are Euclidean. There is a unique linear map f^* defined by the property: $\forall \vec{x} \in \mathbb{R}^q, \vec{y} \in \mathbb{R}^p$

$$\langle f(\vec{x}), \vec{y}\rangle_W = \langle \vec{x}, f^*(\vec{y})\rangle_V. \tag{1}$$

The linear map $f^* : W \to V$ is called the **adjoint** of f. If $A \in \mathbb{R}^{p\times q}$ is the matrix associated to f **between orthonormal bases**, then the matrix associated to f^* is $A^T \in \mathbb{R}^{q\times p}$ defined as follows: if $A = (a_{ij})_{1\leqslant i\leqslant p, 1\leqslant j\leqslant q} \in \mathbb{R}^{p\times q}$, then

$$A^T = (a_{ji})_{1\leqslant i\leqslant p, 1\leqslant j\leqslant q} \in \mathbb{R}^{q\times p}$$

it is called the **transpose** of A (sometimes **adjoint** too). It corresponds to replacing columns and rows – or a reflection along the line $x = -y$.

Some **properties**: $(A^T)^T = A$, $(A + B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, $(A^{-1})^T = (A^T)^{-1}$ (and the associated operations for f).

The transpose lets one rewrite the dot-product: for $\vec{x}, \vec{y} \in \mathbb{R}^n$:

$$\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \vec{y}^T \vec{x},$$

where by vector, we mean column vectors in $\mathbb{R}^n$. Note for instance that for any matrix A, one has the useful identity (compare with (1)):

$$(A\vec{x}) \cdot \vec{y} = (A\vec{x})^T \vec{y} = \vec{x}^T (A^T \vec{y}) = \vec{x} \cdot (A^T \vec{y}).$$

13.3 Self-adjoint operators and symmetric matrices

A linear map $f : \mathbb{R}^p \to \mathbb{R}^p$ is called **self-adjoint** if $f = f^*$. This corresponds to its associated matrix being **symmetric**: $A = A^T$.

For any matrix $B \in \mathbb{R}^{p\times q}$, the matrices $B^T B \in \mathbb{R}^{q\times q}$ and $BB^T \in \mathbb{R}^{p\times p}$ are symmetric. Moreover, one always has:

$$(B\vec{x}) \cdot (B\vec{y}) = (B\vec{x})^T (B\vec{y}) = \vec{x}^T (B^T B)\vec{y}.$$

Note that $(\vec{x}, \vec{y}) \mapsto \vec{x}^T (B^T B)\vec{y}$ is also an inner product that may differ from the dot product. This is the effect a **change of basis** can have on the dot product.

14 Orthogonal subspaces and "fundamental theorem" part 2

14.1 The four subspaces

The **four fundamental subspaces** of a matrix $A \in \mathbb{R}^{p\times q}$ are the following:

1. the nullspace $N(A) \in \mathbb{R}^q$,
2. the column space $C(A) \in \mathbb{R}^p$,
3. the nullspace $N(A^T) \in \mathbb{R}^p$, and
4. the column space $C(A^T) \in \mathbb{R}^q$.

14.2 Orthogonal subspaces

In a vector subspace equipped with an inner product $\langle \cdot, \cdot \rangle$, we say that two linear subspaces W_1 and W_1 if for any $w_1 \in W_1$ and $w_2 \in W_2$, one has $\langle w_1, w_2 \rangle = 0$. We will denote this relationship by $W_1 \perp W_2$, just like we write $w_1 \perp w_2$ for orthogonal vectors.

The **intersection** of orthogonal subspaces is always $\{0\}$.

Given a linear subspace $W \subset V$, we define its **orthogonal complement** $W^\perp$ as the set of vectors orthogonal to all the vectors in W. One obtains an **orthogonal direct sum**: $V = W \overset{\perp}{\oplus} W^\perp$ meaning that for any vector $v \in V$, there exist unique vectors $v_W \in W$, and $v_\perp \in W^\perp$ such that: $v = v_W + v_\perp$. This decomposition is **unique**, and we call v_W and $v_\perp$ the orthogonal projections of v on W and $W^\perp$. They define linear maps: also called **orthogonal projections**

$$\pi_W : V \to W, \quad v \mapsto v_W, \qquad \text{and} \qquad \pi_{W^\perp} : V \to W^\perp, \quad v \mapsto v_\perp.$$

Note that we always have $(W^\perp)^\perp = W$ and that $\dim V = \dim W + \dim W^\perp$. A nice example of orthogonality for instance is that **symmetric** matrices satisfying $A = A^T$ are orthogonal to **anti-symmetric matrices** $B = -B^T$.

14.3 Orthogonality of the four subspaces

The "**second part of the fundamental theorem of linear algebra**" is as follows, it adds one property: for a matrix $A \in \mathbb{R}^{p \times q}$ of rank r.

Theorem 14.1. *One has* $\dim C(A) = \dim C(A^T) = r$*: "the dimension of the image is the rank" Similarly,* $\dim N(A) = q - r$ *and* $\dim N(A^T) = p - r$*: "the* ***codimension*** *of the kernel is the rank" because:*

$$N(A) = C(A^T)^\perp \in \mathbb{R}^q, \qquad \textit{and} \qquad C(A) = N(A^T)^\perp \in \mathbb{R}^p.$$

In the context of linear maps, it states the very general fact that the image of f is the orthogonal complement of the kernel of its adjoint f^*.

14.4 Existence of solutions

Remember that we have a solution to $A\vec{x} = \vec{b}$ if and only if $\vec{b} \in C(A)$. However, it can be tedious to compute $C(A)$ which is the span of vectors, while computing a nullspace is often easier. Another criterion is the following:

Theorem 14.2. *There is a solution to* $A\vec{x} = \vec{b}$ *if and only if* $\vec{b} \perp N(A^T)$.

15 Least squares approximations

When a linear equation $A\vec{x} = \vec{b}$ cannot be solved, i.e. when $\vec{b} \notin C(A)$, one still searches the "best" $\vec{x}^*$ or the closest to a solution one can find.

These situations occur all the time, and often there is solution **by design**. One wants to summarize a complicated data set by something much simpler, the goal is not to match exactly every data point–what would be an actual solution– but rather to find the closest simple summary of the situation. The "best" $\vec{x}^*$ depends on a **loss function** that we set to measure how far we are to an actual solution.

15.1 Typical setup

We will use the **mean squared error** as our loss function: it leads us to search $\vec{x}^*$ minimizing the (Euclidean) distance between $A\vec{x}$ and $\vec{b}$.

In data science and statistics, the problem $A\vec{x} = \vec{b}$ commonly corresponds to:

- the rows of $A \in \mathbb{R}^{p\times q}$ correspond to the measures from different **samples**,
- the columns correspond to the measures of specific **features**,
- the components of $\vec{b}$ are the **outcomes** of each sample, and
- the components of $\vec{x}$ are the **weights** allocated to each feature.

The goal is to **learn** the vector of weights $\vec{x}$ that best summarizes the law: features observed on sample $(i) \longrightarrow$ outcome in sample (i). We measure it by searching for $\vec{x}$ minimizing $\text{dist}(A\vec{x}, \vec{b})$ or equivalently $\|A\vec{x}-\vec{b}\|^2 = \langle A\vec{x}-\vec{b}\ ,\ A\vec{x}-\vec{b}\rangle$.

A nice way to think of the i^{th} component of $\vec{x}$ is the following. If a feature is increased by 1, then, the outcome is changed by x_i.

15.2 How to solve in practice

One typically has many more samples than features measured, so the matrix A is in practice very tall and thin, i.e. $q \ll p$. In this context, $C(A)$–the set of $\vec{b}$ for which the equation can be solved–is at most q-dimensional in the much larger p-dimensional ambient space: almost no equation $A\vec{x} = \vec{b}$ can be solved. Instead, we turn to the minimization problem:

$$\min_{x\in\mathbb{R}^q} \|A\vec{x} - \vec{b}\|^2.$$

There, one can take different, but equivalent approaches:

1. multivariable calculus tells us that the minimizer should satisfy $A^TA\vec{x}^* = A^T\vec{y}$, so if A^TA is invertible, we take $\vec{x}^* = (A^TA)^{-1}A^T\vec{y}$, here $(A^TA)^{-1} \in \mathbb{R}^{q\times q}$ with q the number of features so inverting it by elimination for instance does not cost that much,
2. if there are too many features (and for more difficult loss functions), one would run a **gradient descent** on the function $\vec{x} \mapsto \|A\vec{x} - \vec{b}\|^2$, where for $L(\vec{x}) := \|A\vec{x} - \vec{b}\|^2$ one computes that $\nabla L(\vec{x}) = 2A^TA\vec{x} - 2A^T\vec{y}$, and
3. geometrically, by the Pythagorean theorem, $A\vec{x}^*$ should be the **orthogonal projection** of $\vec{b}$ on $C(A)$, and we will see that this leads again to the formula $\vec{x}^* = (A^TA)^{-1}A^T\vec{y}$.

15.3 The linear algebra behind the method

Let us focus on the last point above. By the second part of the "fundamental theorem" of linear algebra, we have the following orthogonal direct sum:

$$\mathbb{R}^p = C(A) \overset{\perp}{\oplus} N(A^T),$$

so $\vec{b}$ decomposes uniquely as

$$\vec{b} = \vec{p} + \vec{e}, \qquad \text{for } \vec{p} \in C(A), \text{ and } \vec{e} \in N(A^T).$$

We call $\vec{p}$ the **projection** of $\vec{b}$ onto $C(A)$, and $\vec{e}$, the error, i.e. what measures how far we are to $C(A)$. Indeed, the Pythagorean theorem tells us that for any $\vec{x} \in C(A)$, one has:

$$\|\vec{b} - A\vec{x}\|^2 = \|\vec{p} - A\vec{x}\|^2 + \|\vec{e}\|^2,$$

where $\|\vec{e}\|^2$ is fixed, but where one may choose $\vec{x} = \vec{x}^*$ such that $\vec{p} = A\vec{x}^*$ because $\vec{p} \in C(A)$. In order to make this $\vec{x}^*$ **unique**, one may additionally impose that $\vec{x}^* \in C(A^T) = N(A)^\perp$.

16 Linear regression, and vector projections

16.1 Linear regression

A very common situation is that of a cloud of points in $\mathbb{R}^2$ that we want to fit with a straight line. This is **linear regression**, which is a particular case of least squares approximation.

We are given couples $(t_1, y_1), \dots, (t_n, y_n)$ corresponding to the coordinates of the points. Our linear system would ideally correspond to finding a line $y = Cx + D$ through its **slope** coefficient C and the constant D that passes through all of the (t_i, y_i), that is solves the system of n equations $y_i = Ct_i + D$ for all $i \in \{1, \dots, n\}$. This rarely has a solutions, so one instead wants to make the line as close to the set of points $(t_i, y_i)_i$ as possible through:

$$min_{C,D} \sum_{i=1}^{n} \left(y_i - (Ct_i + D)\right)^2.$$

If one denotes $\vec{x} = (C, D)$, $\vec{b} = (y_1, \dots, y_n)$ and $A = \begin{bmatrix} t_1 & 1 \\ \dots & \dots \\ t_n & 1 \end{bmatrix}$, one recovers the previous least squares approximation setting.

One then computes $A^T A = \begin{bmatrix} \sum_i t_i^2 & \sum_i t_i \\ \sum_i t_i & n \end{bmatrix}$, and $A^T \vec{b} = \begin{bmatrix} \sum_i t_i y_i \\ \sum_i y_i \end{bmatrix}$ so that the solution (C, D) must satisfy

$$\begin{bmatrix} \sum_i t_i^2 & \sum_i t_i \\ \sum_i t_i & n \end{bmatrix} \begin{bmatrix} C \\ D \end{bmatrix} = \begin{bmatrix} \sum_i t_i y_i \\ \sum_i y_i \end{bmatrix}.$$

We can always solve this except if all of the t_i are equal, since we know the formula for the inverse of a 2×2 matrix.

16.2 A trick

A trick however simplifies the solution: if the t_i are centered, i.e. if $\sum_i t_i = 0$, then $A^T A = \begin{bmatrix} \sum_i t_i^2 & 0 \\ 0 & n \end{bmatrix}$, and the solution is obviously

$$C = \frac{\sum_{i=1}^n y_i t_i}{\sum_{i=1}^n t_i^2}, \text{ and } D = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

where one recognizes the $\frac{1}{n} \sum_{i=1}^n y_i$ is the **mean** or **expected value** of $(y_1, \dots, y_n)$, as well as the **covariance** $\frac{1}{n} \sum_{i=1}^n y_i t_i$ between $(t_i)_i$ and $(y_i)_i$ (which is an inner product!) and the **variance** $\frac{1}{n} \sum_{i=1}^n t_i^2$ (which is a norm squared!).

The simplest method often becomes the following one (that extends to higher dimensions). Given $(t_1, y_1), \dots, (t_n, y_n)$,

1. Compute the average $\bar{t} = \frac{1}{n} \sum_{j=1}^n t_j$
2. you first center the data by defining $T_i = t_i - \bar{x}$,
3. you can then solve the problem for the set of points (T_i, y_i):

$$C = \frac{\sum_{i=1}^n y_i T_i}{\sum_{i=1}^n T_i^2}, \text{ and } D = \frac{1}{n} \sum_{i=1}^{n} y_i,$$

4. the best line fitting $(t_1, y_1), \dots, (t_n, y_n)$ is then

$$C(t - \bar{t}) + D.$$

Remark 16.1. *We will see that this "centering" of the data can be seen as an* ***orthogonalization*** *of the matrix A.*

Remark 16.2. *In the context of finance, the "β" measures the sensitivity of a stock's returns to the overall market returns. It is essentially the slope coefficient* **C** *in a linear regression model between a stock's returns and the average market's returns. Using least squares, you find the best-fitting line that describes this relationship, and the β quantifies how much the stock is expected to move relative to the market. If* $\beta > 1$*, then the stock is very influenced by the market, and if* $\beta < 1$*, it is relatively not impacted by overall market movements.*

In a similar way, one can fit more **complicated functions** to a set of points. One may fit some polynomial, exponential, logarithm etc depending on the context.

16.3 Vector projection

16.3.1 Projection on a line

When $\vec{b} \in \mathbb{R}^p$ is a vector that you want to project on the line generated by $\vec{a} \neq 0$, the **orthogonal projection** $\vec{p}$ of $\vec{b}$ on $\vec{a}$ is:

$$\vec{p} = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\|^2}\vec{a} = \frac{\vec{a}^T\vec{b}}{\vec{a}^T\vec{a}}\vec{a} = \left((\vec{a}^T\vec{a})^{-1}\vec{a}^T\vec{b}\right)\vec{a}.$$

This recovers our formula $\vec{x}^* = (\vec{a}^T\vec{a})^{-1}\vec{a}^T\vec{b}$ for the solution of the least squares.

The projection on the orthogonal of the line is then $\vec{e} = \vec{b} - \vec{p}$ and one recovers the direct sum decomposition $\vec{b} = \vec{e} + \vec{p}$ with $\vec{e} \perp \vec{p}$.

16.3.2 Projection on a subspace

Consider a subspace spanned by $(\vec{a}_1, \dots, \vec{a}_q)$ inside $\mathbb{R}^p$ and the matrix $A \in \mathbb{R}^{p\times q}$ whose columns are $(\vec{a}_1, \dots, \vec{a}_q)$. The subspace spanned by $(\vec{a}_1, \dots, \vec{a}_q)$ is exactly the column space of A, $C(A)$. We want to project $\vec{b}$ on this subspace.

It is often easier to check that the vector $\vec{e} = \vec{b} - \vec{p}$ in the orthogonal of $C(A)$ or is equivalently in $N(A^T)$. This corresponds to the equations $\vec{a}_i^T\vec{e} = \vec{a}_i^T(\vec{b} - \vec{p}) = 0$ for $i \in \{1, \dots, q\}$ or equivalently $A^T\vec{e} = A^T(\vec{b} - \vec{p}) = 0$. If we force $\vec{p} = A\vec{x}$ (i.e. $\vec{p} \in C(A)$), we recover the previous formula when A^TA is invertible,

$$\vec{p} = A(A^TA)^{-1}A^T\vec{b}.$$

The matrix $P = A(A^TA)^{-1}A^T \in \mathbb{R}^{p\times p}$ is an **orthogonal projection** matrix onto $C(A)$. **Projection matrices** are characterized by the property

$$P^2 = P,$$

and they are orthogonal if they additionally satisfy $P^T = P$ or equivalently if $N(P) \perp C(P)$.

17 Gram-Schmidt orthonormalization

In all of the above methods, the bottleneck is often computing the inverse $(A^TA)^{-1}$. However, if A is expressed in a "nice" enough basis, we will see that we always have $A^TA = \mathrm{I}_q$.

17.1 Orthonormal bases and matrices

Recall that an **orthogonal basis** $(\vec{v}_1, \dots, \vec{v}_n)$ satisfies for all $i \neq j \in \{1, \dots, n\}$,

$$\langle \vec{v}_i, \vec{v}_j \rangle = 0, \text{ and } \langle \vec{v}_i, \vec{v}_i \rangle = 1.$$

We will be interested in **matrices with orthonormal columns**. Such a matrix $Q \in \mathbb{R}^{p\times q}$ satisfies exactly

$$Q^T Q = \mathrm{I}_q\,.$$

One extremely important feature of matrices with orthogonal columns is that **they preserve the dot product**, that is, for any $\vec{x}, \vec{y} \in \mathbb{R}^q$, one has

$$(Q\vec{x}) \cdot (Q\vec{y}) = \vec{x} \cdot \vec{y}$$

where we will note that the first dot-product is in $\mathbb{R}^p$ and the second in $\mathbb{R}^q$.

If (**and only if**) $p = q$, i.e. the matrix is square, we call this matrix **orthogonal** and we find the relationship

$$Q^{-1} = Q^T.$$

Permutation matrices and rotation matrices are orthogonal.

17.2 Gram-Schmidt process

The Gram-Schmidt process is an algorithm which produces an **orthonormal basis** $(\vec{c}_1, \ldots, \vec{c}_n)$, from **any basis** $(\vec{a}_1, \ldots, \vec{a}_n)$. The steps are as follows:

- choose a first element $\vec{a}_1$ of your basis, define
$$\vec{b}_1 = \vec{a}_1,$$
- choose a second one $\vec{a}_2$ and only consider its part that is orthogonal to $\vec{b}_1$, i.e. define
$$\vec{b}_2 := \vec{a}_2 - \frac{\vec{b}_1^T \vec{a}_2}{\vec{b}_1^T \vec{b}_1}\vec{b}_1,$$
- choose a third vector $\vec{a}_3$ of your basis and only consider its part that is orthogonal to $\vec{b}_1$ and $\vec{b}_2$, i.e. define
$$\vec{b}_3 := \vec{a}_3 - \frac{\vec{b}_1^T \vec{a}_3}{\vec{b}_1^T \vec{b}_1}\vec{b}_1 - \frac{\vec{b}_2^T \vec{a}_3}{\vec{b}_2^T \vec{b}_2}\vec{b}_2,$$
- iterate until done for all vectors of the basis, you obtain an **orthogonal basis** $(\vec{b}_1, \ldots, \vec{b}_n)$,
- to further make the basis **orthonormal**, define $\vec{c}_i = \frac{\vec{b}_i}{\|\vec{b}_i\|}$.

Again, a change of basis lets us understand a matrix better.

17.3 QR decomposition

And again, this change of basis can be understood as a *matrix decomposition*. This is the $A = QR$ decomposition. The operations we made on the vectors $\vec{a}_i$ to reach the vectors $\vec{c}_i$ always corresponded to linear combinations of the **columns** of a matrix A whose columns are the $\vec{a}_i$. This means that we were multiplying **from the right** by some matrix $R \in \mathbb{R}^{q\times q}$.

If we look more carefully, we see that the matrix R is additionally **upper triangular**. This can also be seen as follows: since $Q^T Q = \mathrm{I}_q$, from $A = QR$, we find

$$R = Q^T A.$$

The matrix R is then upper triangular and if Q has columns $\vec{q}_1, \ldots, \vec{q}_q$ its coefficient i, j is equal to

$$\vec{q}_i \cdot \vec{a}_j$$

They vanish if $i > j$ by construction.

18 Simplifications with orthonormal bases

18.1 Least squares approximation

Let us come back to the least squares problem: for $A \in \mathbb{R}^{p\times q}$,

$$\min_{\vec{x}} \|A\vec{x} - \vec{b}\|^2.$$

18.2 If Q has orthogonal columns

If the columns of Q form an **orthonormal** family of vectors, then one still has $(Q^TQ)^{-1} = \mathrm{I}_q$, and we find

$$\vec{x}^* = (Q^TQ)^{-1}Q^T\vec{b} = Q^T\vec{b}.$$

18.2.1 In general

In general Gram-Schmidt procedure obtains a matrix Q as elementary operations on the columns of A, and gives

$$Q = AR^{-1}$$

for $R \in \mathbb{R}^{q\times q}$ upper triangular.

This simplifies drastically the least squares approximation. Indeed, one finds $A^TA\vec{x}^* = A^T\vec{b}$ becomes $R^TQ^TQR\vec{x} = R^TQ^T\vec{b}$. Since Q is orthogonal and R invertible, this is equivalent to:

$$R\vec{x} = Q^T\vec{b}.$$

It is a simple system to solve since R is triangular!

18.3 Projections

Recall that the projection of $\vec{b}$ on the subspace $C(A)$ was as follows (when A had linearly independent columns):

$$\vec{p} = A(A^TA)^{-1}A^T\vec{b}.$$

18.3.1 When A has orthonormal columns

If Q has orthonormal columns, then, $(Q^TQ)^{-1} = \mathrm{I}_q$, and we find

$$\vec{p} = QQ^T\vec{b}.$$

We also verify that $P = QQ^T$ is a **projection** matrix since it satisfies $P^2 = P$.

18.3.2 In general

In general, after Gram-Schmidt orthonormalization, we obtain $A = QR$ which simplifies the formula $\vec{p} = A(A^TA)^{-1}A^T\vec{b}$ as:

$$\begin{aligned}\vec{p} &= QR(R^T(Q^TQ)R)^{-1}R^TQ^T\vec{b}\\ &= QRR^{-1}(R^T)^{-1}Q^T\vec{b}\\ &= QQ^T\vec{b}.\end{aligned}$$

Notice that it does not depend on R. Indeed, this is a projection on a subspace: it is **geometric**. It means that it should not depend on the basis we take: it is the same formula with the columns of A or the columns of Q.

Remark 18.1. *The above denominations are **confusing**. A square matrix with ortho**normal** columns is called ortho**gonal**.*

*A matrix with ortho**normal** columnns satisfyies $Q^T Q = \mathrm{I}_q$ while for a matrix with ortho**gonal** columns, $Q^T Q$ is only diagonal.*

19 Determinant

The determinant is a scalar value that can be computed from a square matrix and encapsulates various properties of the matrix. It is denoted as $\det(A)$ or $|A|$ for a square matrix A.

19.1 Formulas for the determinant in dimensions 2 and 3

For a 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

the determinant is given by

$$\det(A) = ad - bc.$$

For a 3×3 matrix

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix},$$

the determinant is

$$\det(A) = a(ei - fh) - b(di - fg) + c(dh - eg).$$

19.2 Defining properties of determinant

The determinant function has several defining properties:

1. **Identity matrix:** The determinant of the identity matrix is 1.

$$\det(I) = 1$$

2. **Row exchange:** Exchanging two rows of A negates $\det(A)$.

$$\det(A') = -\det(A)$$

3. **Determinant and linearity:** The determinant is a linear function of each row when others are held constant.

$$\det(a_1 \vec{R}_1 + a_2 \vec{R}_2, \vec{R}_3, \dots, \vec{R}_n) = a_1 \det(\vec{R}_1, \vec{R}_3, \dots, \vec{R}_n) + a_2 \det(\vec{R}_2, \vec{R}_3, \dots, \vec{R}_n)$$

19.3 Other important properties

4. **If two rows are equal, then** $\det A = 0$**.**
5. **Subtracting a row to another** leaves det unchanged.
6. **A row of zeros gives zero:** $\det A = 0$ if A has a row of zeros.
7. **If A is triangular**, then $\det(A)$ is the product of diagonal entries: $\det(A) = a_{11} \times a_{22} \times \cdots \times a_{nn}$

8. $\det(A) \neq 0$ **iff** A **is invertible.**

9. **Multiplicative property:** $\det(AB) = \det(A)\det(B)$

10. **Transpose property:** $\det(A^T) = \det(A)$.

That last property tells us that all of the previous properties that we prove about row operations can also be performed on columns with analogue conclusions!

20 Permutations and cofactors

From the above properties 7. and 9., we see that the $A = LU$ decomposition is a way to compute the determinant of A. If the diagonal terms in U are $d_1, \dots, d_p$, then determinant of A is the product $d_1 \dots d_p$.

Permutations of rows multiply the determinant by -1, so permutations in the elimination process also works.

This is however typically overkill, and we present another method today.

20.1 Determinants and pivots

One simple but deep observation is that when applying the elimination process to a matrix A, *the first k pivots only depend on the matrix entries* a_{ij} *for* $i \leqslant k$, $j \leqslant k$, i.e. the terms in the $k \times k$ matrix A_k in the top-left corner of A.

We therefore obtain subdecompositions: $A_k = L_k U_k$ where L_k and U_k are respectively lower and upper triangular $k \times k$ matrices. In particular, $\det A_k = d_1 \dots d_k = d_k \det A_{k-1}$, which yields (if the matrix A_{k-1} is invertible):

$$d_k = \frac{\det A_k}{\det A_{k-1}}.$$

20.2 General formula for the determinant in terms of the entries of A

One may write a general but mostly useless in practice (but sometimes useful abstractly) formula for the determinant:

$$\det A = \sum_{\substack{\text{permutation matrices } P \\ P(1,\dots,p)=(b_1,\dots,b_p)}} (\det P)\ a_{1b_1} a_{2b_2} \dots a_{pb_p}.$$

20.3 Determinants and cofactors

Take a matrix $A \in \mathbb{R}^{p\times p}$ and denote $A_{ij} \in \mathbb{R}^{(p-1)\times(p-1)}$ the matrix obtained by removing the i^{th} row and j^{th} column of A. We may write the determinant of A in terms of the determinants of the smaller matrices A_{ij}. More precisely, define the **cofactors**:

$$c_{ij} := (-1)^{i+j} \det A_{ij},$$

then, if the first row is the simplest (more zeros, only small integers...) you may use:

$$\det A = \sum_j a_{1j} c_{1j}.$$

If the i^{th} row is simpler, then use

$$\det A = \sum_j a_{ij} c_{ij}.$$

Remember the pattern

$$\text{sign}(-1)^{i+j} = \begin{bmatrix} + & - & + & - \\ - & + & - & + \\ + & - & + & - \\ - & + & - & + \end{bmatrix}.$$

21 Determinants, cofactors and inverses

Determinants of well-chosen submatrices may be used to solve linear systems and compute inverses of matrices.

21.1 Solving linear systems by computing determinants

Consider $A \in \mathbb{R}^{p\times p}$, $\vec{x}, \vec{b} \in \mathbb{R}^p$, and consider the linear system $A\vec{x} = \vec{b}$. We can directly write the solution in terms of determinants obtained from specific matrices.

Consider the matrices B_j such that all columns are equal to those of A except the j^{th} column that is equal to $\vec{b}$.

Define the matrices F_j such that all columns are equal to those of I_p except the j^{th} column that is equal to $\vec{x}$. One computes:

$$AF_j = B_j,$$

and since a direct computation (from the j^{th} row) shows that $\det F_j = x_j$, we find:

$$x_j \det A = \det B_j,$$

and in particular, if A is invertible, we obtain the formula

$$x_j = \frac{\det B_j}{\det A},$$

determining all the coefficients of $\vec{x}$ at the cost of computing determinants. This technique is known as Cramer's rule.

Note: Cramer's rule is inefficient algorithmically on large general matrices as it requires many determinants. It is however appealing theoretically. For instance, A and $\vec{b}$ only have integer coefficients and if $\det A = \pm 1$, then, the solution $\vec{x}$ only has integer coefficients.

21.2 Computing inverse matrices thanks to determinants

We also have a **direct formula** for the inverse of a matrix $A \in \mathbb{R}^{p\times p}$. Denote C the **matrix** of cofactors c_{ij}. One has the general formula:

$$AC^T = \det A\, \mathrm{I}_p,$$

when A is invertible, this yields the beautiful formula:

$$\frac{1}{\det A} C^T = A^{-1}.$$

22 Eigenvalues and Eigenvectors

A $p \times p$ matrix A often has **preferred directions** in which its action is simple. These are the **eigenvectors** of A.

22.1 Definition

An **eigenvector** of a $p \times p$ matrix A is a non-zero vector $\vec{x}$ such that $A\vec{x} = \lambda\vec{x}$ for some scalar λ. The scalar λ is called the **eigenvalue** corresponding to the **eigenvector** $\vec{x}$.

22.2 Computing eigenvalues and eigenvectors

In some special cases, one can find **eigenvalues** and **eigenvectors** directly by inspection. For example, the identity matrix has eigenvalue 1 for all vectors.

In other cases, a robust method is:

1. To find the *eigenvalues*, one can solve the **characteristic equation** $\det(A - \lambda I) = 0$. The roots of this equation are the *eigenvalues*.
2. Once an *eigenvalue* λ is found, the corresponding *eigenvectors* are in the *nullspace* of $A - \lambda I$. To find the *eigenvectors*, solve $(A - \lambda I)\vec{x} = \vec{0}$.

22.3 Important properties

- The **eigenvalues** of a triangular matrix are simply its **diagonal entries**.
- The **determinant** of a matrix is equal to the product of its **eigenvalues**.
- The sum of the **eigenvalues** is equal to the **trace** of the matrix.

Some complications:

1. **Gaussian elimination** does not preserve **eigenvalues** or **eigenvectors**.
2. **Eigenvalues** may be **complex numbers**, even for real matrices.
3. Not all matrices have p distinct **eigenvalues** or **eigenvectors**. **Eigenvalues** can have algebraic multiplicities greater than 1, and this may complicate finding a basis of **eigenvectors**.

23 Diagonalizing a matrix

23.1 Diagonalization

Let $A \in \mathbb{R}^{p \times p}$. To transform a matrix A into a **diagonal** matrix D, one constructs the matrix S whose columns are **eigenvectors**. Then one finds $S^{-1}AS = D$, which is a diagonal matrix composed of the **eigenvalues** of A.

Important remarks: exchanging the order of eigenvectors in S will correspondingly change the order of eigenvalues in the diagonal matrix D. If a matrix has p **distinct** eigenvalues (a generic condition), it has a basis of eigenvectors and can therefore be diagonalized.

23.2 Mistakes to avoid

It is critical to remember that some matrices are nondiagonalizable, which means they do not have a sufficient number of distinct eigenvectors to form a basis. As such, no matrix S of eigenvectors exists that can diagonalize them.

Furthermore, it is a common mistake to assume that the eigenvalues of $A + B$ or AB can be easily inferred from the eigenvalues of A and B; this is not generally the case.

23.3 Matrix powers

Diagonalization can be particularly useful for the fast computation of matrix powers. If a matrix A is diagonalizable, one can compute high powers of A efficiently:

$$A^k = S D^k S^{-1},$$

where D^k is just the matrix whose diagonal entries are k^{th} powers of the entries of D.

Diagonalization simplifies the computation of sequences of vectors defined by a recurrence relation such as $\vec{u}_{k+1} = A\vec{u}_k$, since it allows for leveraging the powers of a diagonal matrix.

24 Eigenvalues: Properties and Applications

An **eigenspace** is the set of eigenvectors associated with an eigenvalue plus zero. It is always a vector subset since it is the nullspace of a matrix.

24.1 Determinant and Trace

The **determinant** of matrix $A \in \mathbb{R}^{p\times p}$ with p eigenvalues (counted with multiplicity as roots of the characteristic polynomial) is the product of its eigenvalues,

$$\det(A) = \prod_{i=1}^{p} \lambda_i = \lambda_1 \times \cdots \times \lambda_p,$$

and the **trace** of A, which is the sum of its diagonal entries, is the sum of its eigenvalues,

$$\mathrm{tr}(A) = \sum_{i=1}^{p} \lambda_i = \lambda_1 + \cdots + \lambda_p.$$

In particular, if one writes that the characteristic polynomial is

$$\det(A - s\,\mathrm{I}_p) = (-1)^p s^p + a_{p-1} s^{p-1} + \cdots + a_1 s + a_0,$$

we find: $a_{p-1} = (-1)^{p-1}\,\mathrm{tr}(A)$ and $a_0 = \det(A)$.

One also deduces an important formula: the derivative of a determinant is a trace. Geometrically, the derivative of the volume can be seen as a trace. More concretely, define $f : \mathbb{R} \to \mathbb{R}$, $f : t \mapsto \det(I_p + tA)$, then f is differentiable (it is a polynomial) and one has:

$$f'(0) = \mathrm{tr}(A).$$

More generally, for a differentiable family of invertible matrices $t \mapsto A(t)$, one finds:

$$\frac{d}{dt}\det(A) = \det(A)\,\mathrm{tr}\left(A^{-1}\frac{dA}{dt}\right).$$

24.2 Symmetric Matrices

For any symmetric matrix $A \in \mathbb{R}^{n\times n}$, all eigenvalues are real, and there exists an orthonormal basis of $\mathbb{R}^n$ consisting of eigenvectors of A. This leads to a special form of diagonalization known as **spectral decomposition**,

$$A = Q D Q^T,$$

where Q is an orthogonal matrix whose columns are eigenvectors of A, and D is a diagonal matrix of the eigenvalues.

24.3 Eigenvalues and Matrix Functions

If $f : \mathbb{R} \to \mathbb{R}$ is a function, we can make sense of $f(A)$ for A diagonalizable real eigenvalues (if not diagonalizable, one typically asks that f is analytic) through

$$f(A) = Qf(D)Q^T,$$

where $f(D)$ means simply applying f to each of the eigenvalues on the diagonal of D.

24.4 Applications of Eigenvalues

Eigenvalues play a crucial role in many areas, including:

- **Differential Equations:** Eigenvalues are used to solve systems of linear differential equations.
- **Stability Analysis:** In dynamical systems, the stability of an equilibrium point is often determined by the eigenvalues of the Jacobian matrix at that point.
- **Principal Component Analysis (PCA):** In statistics and machine learning, eigenvalues are used to determine the principal components of a data set. The larger the eigenvalues of the covariance matrix, the more variance there is in the associated eigenspaces.
- **Quantum Mechanics:** In physics, eigenvalues of self-adjoint linear maps (in infinite dimension) correspond to the observable quantities in quantum mechanics.
- **Ranking nodes of a graph:** See PageRank algorithm (Page-Brin, Google'98) from Perron-Frobenius.

24.5 Non-Diagonalizable Matrices

Not all matrices are diagonalizable. Matrices that cannot be brought to a diagonal form are called **non-diagonalizable**.

Non-diagonalizable matrices occur when the **geometric multiplicity** (i.e. the dimension of the eigenspace) of an eigenvalue is less than its **algebraic multiplicity** (i.e. its multiplicity as a root of the characteristic polynomial). In other words, there are not enough linearly independent eigenvectors to form a basis of $\mathbb{R}^n$.

For such matrices, we can still achieve a form of diagonalization using **Jordan's decomposition** the Jordan canonical form, which is a block diagonal matrix where each block is called a Jordan block. A Jordan block associated with an eigenvalue λ which is a matrix with λ in all of the diagonal entries, and a number of 1 and 0 on the "superdiagonal", i.e. the entries just above the diagonal.

24.6 Cayley-Hamilton Theorem

Jordan's decomposition relies on a beautiful and deep theorem.

The **Cayley-Hamilton theorem** is a fundamental result in linear algebra which states that every square matrix $A \in \mathbb{R}^{n \times n}$ satisfies its own characteristic equation. If $p(\lambda) = \det(A - \lambda I)$ is the characteristic polynomial of A, then

$$p(A) = 0.$$

This means that if we substitute the matrix A into its characteristic polynomial, the result is the zero matrix. This theorem has practical applications in computing functions of matrices (like the matrix exponential) and in simplifying expressions involving matrices.

For example, in dimension 2, for every possible matrix $A \in \mathbb{R}^{2\times 2}$, one has $p(\lambda) = \lambda^2 - \mathrm{tr}(A)\lambda + \det(A)$, so by the Cayley-Hamilton theorem, A satisfies

$$A^2 - \mathrm{tr}(A)A + \det(A)I = 0, \text{ or equivalently } A^2 = \mathrm{tr}(A)A - \det(A)I.$$

This property can be used to compute higher powers of A and inverse of A (if it exists) by expressing them in terms of A and I, reducing the computational complexity significantly for certain matrices.

25 Exponential of a matrix and systems of ODEs

We have two equivalent definitions for the *exponential of a matrix* M:

1. $e^{tM} = Y(t)$ is the unique solution to the ODE:

$$\begin{cases} Y'(t) = MY(t), \\ Y(0) = \mathrm{I}_p, \end{cases}$$

2. it is given by the sum:

$$e^{tM} = \sum_{k=0}^{+\infty} \frac{t^k}{k!} M^k.$$

It is an important definition since any solution of $Y'(t) = AY(t)$ is equal to

$$Y(t) = e^{(t-t_0)A} Y(t_0).$$

Computing the exponential of a diagonalizable matrix is simple from the second definition of a diagonalizable matrix A such that $A = SDS^{-1}$ for a diagonal matrix D:

$$e^{tA} = Se^{tD}S^{-1},$$

where if $D = diag(\lambda_1, ..., \lambda_p)$, then $e^{tD} = diag(e^{\lambda_1 t}, ..., e^{\lambda_p t})$.

From the first definition, we can also say that if $AV_i = \lambda_i V_i$, then $e^{tA}V_i = e^{t\lambda_i}V_i$. This is especially useful when solving ODEs: if $Y(t_0) = \sum y_i V_i$, then the solution of the ODE is given by

$$Y(t) = e^{(t-t_0)A} Y(t_0) = \sum_i e^{\lambda_i t} y_i V_i.$$

We have two simple and usual criteria implying that a matrix is diagonalizable.

Proposition 25.1. *If $A \in \mathbb{R}^{p\times p}$ has d distinct eigenvalues, then it is diagonalizable.*

(this implies that with probability 1 a random matrix is diagonalizable).

Proposition 25.2. *If $A \in \mathbb{R}^{p\times p}$ is symmetric, then it is diagonalizable in an orthonormal basis and has real eigenvalues.*

26 The Perron-Frobenius Theorem: Foundations and Applications

26.1 Perron-Frobenius Theorem for Positive Matrices

The **Perron-Frobenius theorem** asserts that for any square matrix with **positive entries**, there exists a unique largest real eigenvalue, known as the Perron root, which is associated with a strictly positive eigenvector. Formally, given a matrix $A \in ((0, +\infty)])^{p\times p}$, there exists $\lambda > 0$ and $x > 0$ such that $Ax = \lambda x$.

Properties of the Perron Root and Eigenvector:

- The Perron root is simple: its algebraic multiplicity is one.
- The associated eigenvector can be chosen to have all components positive.
- The Perron root dominates in magnitude: $\lambda > |\mu|$ for any other eigenvalue μ of A.

26.2 Applications to Stochastic Matrices and PageRank

The **Google PageRank algorithm** is an application of the Perron-Frobenius theorem to stochastic matrices representing the web graph.

In this context, the internet may be represented as a graph: **nodes** are webpages and links are **edges**. If you think of a person randomly clicking on links on each page, the question is: after a lot of clicks, what **proportion of time** will they spend on each page? Each page's rank (or importance) is determined by this steady-state probability distribution, whose existence and uniqueness is a consequence of Perron-Frobenius theorem.

Stochastic Matrices and PageRank:

- A web graph's adjacency matrix can be transformed into a stochastic matrix where each column sums to one, making it suitable for PageRank.
- The largest eigenvalue of such a matrix is always 1 due to its stochastic nature.
- The corresponding eigenvector gives the steady-state probability distribution of someone randomly browsing– guided by the , which is used to rank web pages.

The Perron-Frobenius theorem ensures the existence and uniqueness of the steady-state distribution.

Eigenvector Computation and Power Method: The computation of the PageRank vector (eigenvector associated with eigenvalue 1) can be efficiently done: powers of the matrix converge. Starting from a generic initial vector $\vec{x}_0$, a renormalization (so that the sum of the elements is 1) of the elements of the sequence $\vec{x}_{n+1} = A\vec{x}_n$ converge to the normalized eigenvector associated to the largest eigenvalue 1. Its coefficients represent the proportion of time spent at each node of the graph if one follows the probability given by the edges.

27 Symmetric Matrices and Their Diagonalization

Symmetric matrices are fundamental in linear algebra and its applications. A matrix A is **symmetric** if $A = A^T$, where A^T denotes the transpose of A.

27.1 Diagonalization of Symmetric Matrices

A key property of symmetric matrices is that they are **diagonalizable**. This means that for a symmetric matrix A, there exists an orthogonal matrix Q and a diagonal matrix D such that $A = QDQ^T$. Here, Q is the matrix whose columns are the elements of the orthonormal basis of eigenvectors.

Another important property is that all **eigenvalues** of a symmetric matrix are **real**. This is crucial for certain applications, like in optimization and quadratic forms.

Another property of symmetric matrices is that the signs of the **pivots** are the same as the signs of the **eigenvalues**. Their values can however be very different!

27.2 Positive Definite Matrices

A symmetric matrix $A \in \mathbb{R}^{p\times p}$ is **positive definite** if for any non-zero vector $\vec{x}$, the quadratic form $\vec{x}^T A\vec{x}$ is positive. Other characterizations are:

- the eigenvalues of A are all positive,
- the pivots of A are all positive,
- the determinants of the submatrices A_k composed of the top left $k \times k$ components of A are positive.

For any invertible matrix R, the matrix $A = R^T R$ is positive since

$$\vec{x}^T A\vec{x} = \vec{x}^T R^T R\vec{x} = (R\vec{x})^T (R\vec{x}) = \|R\vec{x}\|^2.$$

Conversely, any symmetric definite positive matrix A can be written as $A = R^T R$ for R invertible. Indeed, if $A = QDQ^T$, then $A = R^T R = R^2$ for $R = Q\sqrt{D}Q^T$.

These matrices are crucial as they exhaust the possible inner products on $\mathbb{R}^p$: any Euclidean distance that fits your problem will come from such a matrix.

27.3 Applications

Positive definite matrices have numerous applications in optimization, statistical estimation, and numerical analysis. They guarantee the uniqueness and stability of solutions in various problems.

In the upcoming lectures, we will extend these concepts to more general matrices using the **Singular Value Decomposition (SVD)**. SVD is a powerful tool in matrix analysis and has applications in data science, signal processing, and more.

The starting point of this singular value decomposition of a (rectangular!) matrix $A \in \mathbb{R}^{p\times q}$ is the above spectral decomposition applied to the **symmetric** matrices $AA^T \in \mathbb{R}^{p\times p}$ and $A^T A \in \mathbb{R}^{q\times q}$.

28 Singular Value Decomposition: Theory and Applications

Singular Value Decomposition (SVD) is a factorization of a matrix into three matrices, capturing essential geometric and algebraic properties.

28.1 The SVD Theorem

For any matrix $X \in \mathbb{R}^{p\times q}$, there exist orthogonal matrices $U \in \mathbb{R}^{p\times p}$, $V \in \mathbb{R}^{q\times q}$, and a diagonal matrix $\Sigma \in \mathbb{R}^{p\times q}$ such that

$$X = U\Sigma V^T.$$

The diagonal entries of Σ are the **singular values** of X, and the columns of U and V are the left and right **singular vectors**, respectively. We will require the terms in Σ to be nonnegative.

28.2 Connection to Eigendecomposition

SVD generalizes the concept of eigendecomposition from square matrices to rectangular ones. For symmetric matrices, eigendecomposition is

$$A = Q\Lambda Q^T,$$

whereas SVD of a general matrix X is

$$X = U\Sigma V^T.$$

This highlights the elegance of SVD in capturing the structure of a matrix.

The proof of the existence of this decomposition is based on the diagonalization in orthonormal bases (or spectral decomposition) of the symmetric semidefinite matrices $X^T X$ and XX^T. In fact, the columns of U form an orthonormal basis of eigenvectors of $X^T X$ while the columns of V form an orthonormal basis of eigenvectors of XX^T.

28.3 Truncated SVD

In many practical scenarios, a full SVD is not necessary. The **truncated SVD** uses only the largest k singular values (and corresponding vectors), providing a good approximation of the original matrix. This approach is particularly useful in applications like data compression and reduced-order modeling.

The intuition is that the large singular values are the most meaningful directions of the matrix action, and the rest can be considered as negligible in some applications. Typically, if a matrix is sparse, it will have very few low singular values which still capture the core behavior of the linear operation.

28.4 Numerical Computation

The computation of SVD can be efficiently done using iterative algorithms. Methods like the power method are employed to approximate the largest singular values and the associated singular vectors, especially in large-scale matrices common in data science applications.

28.5 Applications of SVD

SVD has profound applications in various fields, including:

- **Data Compression:** SVD is used in image and signal processing for data compression and noise reduction.
- **Machine Learning:** In machine learning, SVD is used in algorithms like PCA for dimensionality reduction.
- **Numerical Analysis:** SVD aids in solving linear systems and in matrix inversion approximations.
- **Natural Language Processing:** In NLP, SVD is used for semantic analysis and building latent semantic indexing models.
- **Computer Graphics:** SVD is pivotal in animations and graphics for transformations and decompositions.

29 SVD (continued)

29.1 Sum of matrices of rank 1

The SVD says that any matrix A of rank r is a weighted sum of normalized matrices of rank 1. The weights are the nonzero singular values σ_1,..., σ_r. If the left and right singular vectors associated to the σ_i are respectively the $\vec{u}_i$ and the $\vec{v}_i$, then the normalized rank 1 matrices are $\vec{u}_i\vec{v}_i^T$. More concretely:

$$A = \sigma_1\vec{u}_1\vec{v}_1^T + \cdots + \sigma_r\vec{u}_r\vec{v}_r^T.$$

29.2 Dimensional reduction

We have simplified a $p \times q$ matrix by r matrices of rank 1, which already reduces quite drastically the complexity.

In practice, there are often singular values which are much larger than the others. They capture the main contributions of the matrix A.

One may consider that the small singular values are negligible and do not truly contribute. For instance, if the singular values are ranked as $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r$ one could say that the σ_k that are less than $0.001 \times \sigma_1$ (or something more intelligent depending on r) are negligible. For instance, if $\sigma_4 < 0.001 \times \sigma_1$ one may start from $A = \sigma_1\vec{u}_1\vec{v}_1^T + \cdots + \sigma_r\vec{u}_r\vec{v}_r^T$ and just keep the leading terms:

$$A \approx \sigma_1\vec{u}_1\vec{v}_1^T + \sigma_2\vec{u}_2\vec{v}_2^T + \sigma_3\vec{u}_3\vec{v}_3^T.$$

29.3 Proving SVD

In order to prove SVD, one starts by diagonalizing the symmetric definite semi-positive matrix $X^T X$ as:

$$X = Q\Lambda Q^T,$$

with $\Lambda = \text{diag}(\sigma_1^2, \dots, \sigma_r^2, 0, \dots, 0)$, and find $V = Q$ and Σ from the $\sigma_i \geqslant 0$.

From $XV = U\Sigma$, one then finds the columns $\vec{u}_i$ from the $\vec{v}_i$ and σ_i for $i \leqslant r$–they are automatically orthonormal. The $\vec{u}_i$ for $i \geqslant r+1$ can be freely chosen as long as they complete the orthonormal basis.

As an example of application, a matrix can be represented by scales of gray. One nice application of SVD is in compressing images. In an image with thousands of components, often only the first few ones are the most meaningful as you can see below.

30 Moore-Penrose Inverse: Concepts and Applications

30.1 Definition and Conditions

For a matrix $A \in \mathbb{K}^{m\times n}$, the Moore-Penrose pseudoinverse, denoted $A^+ \in \mathbb{K}^{n\times m}$, satisfies four conditions:

1. $AA^+A = A$.
2. $A^+AA^+ = A^+$.
3. $(AA^+)^T = AA^+$.
4. $(A^+A)^T = A^+A$.

These conditions ensure that A^+A is the **orthogonal projection** onto $C(A^T) = N(A)^\perp$, and AA^+ is the **orthogonal projection** onto $C(A)$. Indeed $(A^+A)^2 = A^+A$ and $(AA^+)^2 = AA^+$ and the matrices are orthogonal.

30.2 Linear map/Geometric point of view

The linear map associated to $A \in \mathbb{R}^{p\times q}$. Then, from the four fundamental subspaces, we find the orthogonal direct sums:

$$\mathbb{R}^q = N(A) \overset{\perp}{\oplus} C(A^T),$$

$$\mathbb{R}^p = N(A^T) \overset{\perp}{\oplus} C(A).$$

where we recall that $\dim C(A) = \dim C(A^T) = r$ is the **rank** of A.

If we look at the map $f : \mathbb{R}^q \to \mathbb{R}^p$ defined by $f(\vec{x}) = A\vec{x}$ in this splitting, we obtain:

- the restriction of f to $C(A^T)$, $f_{|C(A^T)}$ is **invertible** between $C(A^T)$ and $C(A)$. Let us denote $(f_{|C(A^T)})^{-1} : C(A) \to C(A^T)$ its inverse, and
- the restriction of f to the orthogonal complement $N(A)$ is the zero map by definition.
- Consequently, for any $\vec{x} \in \mathbb{R}^q$, there is a unique orthogonal decomposition $\vec{x} = \vec{x}_{N(A)} + \vec{x}_{C(A^T)}$ with $(\vec{x}_{N(A)}, \vec{x}_{C(A^T)}) \in N(A) \times C(A^T)$. This lets us understand f as:

$$f(\vec{x}) = f_{|C(A^T)}(\vec{x}_{C(A^T)}).$$

Now, the Moore-Penrose pseudo-inverse A^+ is defined as the matrix associated to the linear map $f^+ : \mathbb{R}^p \to \mathbb{R}^q$ defined below between canonical bases:

- the restriction of f^+ to $C(A)$, is $(f_{|C(A^T)})^{-1} : C(A) \to C(A^T)$, and
- the restriction of f to the orthogonal complement $N(A^T)$ is the zero map.
- Consequently, for any $\vec{y} \in \mathbb{R}^p$, there is a unique orthogonal decomposition $\vec{y} = \vec{y}_{N(A^T)} + \vec{y}_{C(A)}$ with $(\vec{y}_{N(A^T)}, \vec{y}_{C(A)}) \in N(A^T) \times C(A)$. This lets us understand f^+ as:

$$f^+(\vec{y}) = f_{|C(A)}(\vec{y}_{C(A)}) = (f_{|C(A^T)})^{-1}(\vec{y}_{C(A)}).$$

One verifies that this satisfies the four defining properties and that

$$f \circ f^+(\vec{y}) = \vec{y}_{C(A)}, \qquad \text{and} \qquad f^+ \circ f(\vec{x}) = \vec{x}_{C(A^T)}.$$

30.3 Singular Value Decomposition (SVD) Method

The pseudoinverse can be computed using SVD, where if $A = U\Sigma V^T$, then $A^+ = V\Sigma^+ U^T$, where if $\Sigma \in \mathbb{R}^{p \times q}$ is the matrix with $r = \text{rank}(A)$ nonzero diagonal values $\sigma_1, \ldots, \sigma_r$, then $\Sigma^+ \in \mathbb{R}^{q \times p}$ is the matrix with r nonzero diagonal values $\sigma_1^{-1}, \ldots, \sigma_r^{-1}$.

This ensures that $AA^+ = UU^T$ and $A^+A = VV^T$, which are projections onto the image and support of A, respectively.

30.4 Properties

- Existence and Uniqueness: Each matrix has a unique pseudoinverse.
- If A is invertible, then $A^+ = A^{-1}$.
- For a squared diagonal matrix D, D^+ is obtained by taking reciprocals of nonzero diagonal elements.
- The pseudoinverse of the pseudoinverse is the original matrix: $(A^+)^+ = A$.
- $(AB)^+ = B^+A^+$.
- $\ker(A^+) = \ker(A^T)$ and $\text{ran}(A^+) = \text{ran}(A^T)$.
- If A^TA is invertible, then $A^+ = (A^TA)^{-1}A^T$ is a left-inverse for A, i.e. $A^+A = I_q$.

30.5 Application in Linear Systems and Least Squares

The Moore-Penrose pseudoinverse is instrumental in solving linear systems or more generally least squares problems. When faced with a system of equations, represented by $A\vec{x} = \vec{b}$, a solution might not exist, and if it does it might not be unique.

The unique "best" solution using the pseudoinverse is given by $\vec{x} = A^{+}\vec{b}$. This solution minimizes the least squares error $\|A\vec{x} - \vec{b}\|^2$.

Among all possible least squares solutions,$\vec{x}$ is the smallest. More precisely for any other minimizer $\vec{y}$ such that $\|A\vec{y} - \vec{b}\|^2 = \|A\vec{x} - \vec{b}\|^2$, one has $\|\vec{y}\|^2 \geqslant \|\vec{x}\|^2$ with equality only if $\vec{y} = \vec{x}$.

This is the most efficient and stable solution in the presence of numerical errors or perturbations. This property is particularly useful in data fitting, regression analysis, and machine learning applications where least squares methods are prevalent.

31 Cayley-Hamilton Theorem and Jordan decomposition

31.1 Linear Maps and Matrices: Similar Matrices and Changes of Bases

31.1.1 Similar Matrices

Two matrices A and B are **similar** if there exists an invertible matrix P such that $A = PBP^{-1}$. Similar matrices represent the same linear transformation in different bases: if one goes from the canonical basis to a basis $\vec{x}_1, \dots, \vec{x}_p$, then P is the matrix with columns $\vec{x}_1, \dots, \vec{x}_p$.

31.2 Nilpotent Matrices

31.2.1 Definition

A matrix A is **nilpotent** if there exists some positive integer k such that $A^k = 0$. The only eigenvalue of nilpotent matrices is zero, and the non zero nilpotent matrices are not diagonalizable.

31.3 Cayley-Hamilton Theorem

The Cayley-Hamilton theorem states that every square matrix satisfies its own characteristic equation. That is, if $p(s) = \det(A - s\,\mathrm{I}_p)$ is the characteristic polynomial of a matrix A, then $p(A) = 0$.

This in particular tells us that all high powers of a matrix can be expressed from the p first powers of a matrix, this is in particular useful, when considering infinite sums such as the exponential of a matrix.

31.4 Jordan Decomposition

31.4.1 Definition

The Jordan decomposition of a matrix, also known as its canonical form, is a representation of a matrix as the sum of a diagonalizable matrix and a nilpotent matrix.

31.4.2 Jordan Blocks

Before proving the Jordan decomposition, we need to understand the concept of a Jordan block. A Jordan block is a square matrix of the form:

$$J = \begin{pmatrix} \lambda & * & 0 & \cdots & 0 & 0 \\ 0 & \lambda & * & \cdots & 0 & 0 \\ 0 & 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda & * \\ 0 & 0 & 0 & \cdots & 0 & \lambda \end{pmatrix}$$

where λ is an eigenvalue of the matrix, and where the $*$ can take the values 0 or 1. Jordan blocks are used to construct the Jordan form of a matrix, which is a block diagonal matrix where each block is a Jordan block.

31.4.3 Derivation from the Cayley-Hamilton Theorem

Step 1: Application of Cayley-Hamilton Theorem Let A be a square matrix, and let $p(\lambda)$ be its characteristic polynomial given by $p(\lambda) = \det(\lambda I - A)$. By the Cayley-Hamilton theorem, $p(A) = 0$. This implies that A satisfies its own characteristic equation.

Step 2: Characteristic Polynomial and Primary Decomposition The vector space into a direct sum of **invariant subspaces** (i.e. vector subspaces V such that $A(V) \subset V$) corresponding to the nullspace of $(A - \lambda)^{m_{\lambda_i}}$ for distinct eigenvalue λ_i, where m_{λ_i} is the multiplicity.

Step 3: Construction of Jordan Blocks For each eigenvalue λ_i, consider the subspace $N((A - \lambda)^{m_{\lambda_i}})$ associated with it. In this subspace, the matrix $(A - \lambda_i I)$ is nilpotent. We can find a basis such that $(A - \lambda_i I)$ is represented by a matrix in Jordan block form. This requires finding a chain of generalized eigenvectors that spans the subspace.

Step 4: Assembling the Jordan Form By repeating the process for each distinct eigenvalue, we obtain a collection of Jordan blocks. The Jordan form of A is then a block diagonal matrix where each block is one of these Jordan blocks.

Step 5: Jordan Decomposition The Jordan decomposition of A is given by $A = D + N$, where D is a diagonalizable matrix containing the eigenvalues of A, and N is a nilpotent matrix corresponding to the superdiagonal elements in the Jordan blocks. By construction, the matrices D and N **commute**.

31.5 Applications

The Jordan decomposition is used to solve systems of linear differential equations and is instrumental in the stability analysis of dynamical systems.