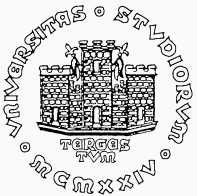
# Computer Vision and Pattern Recognition

Course ID: 554SM – Fall 2020

Felice Andrea Pellegrino

University of Trieste
Department of Engineering and Architecture



# Linear Algebra Review

The following material[1] is basically a collection of concepts and results of Linear Algebra that are frequently encountered in Computer Vision and Pattern Recognition. An in-depth treatment can be found in Strang (2016) and Meyer (2000).

[1] Part of which is taken from Kolter (2019).

# Basic concepts and notation

# Sets of linear equations

Linear algebra deals with sets of linear equations.

Linear algebra provides a way to represent compactly the sets of linear equations, analyzing their properties and operating on them.

For example:

$$\begin{aligned} x_1 + 2x_2 &= 5 \\ -x_1 + x_2 &= 2 \end{aligned}$$

is a set of two equations in two variables. It can be compactly represented as

$$Ax = b$$

where

$$A = \begin{bmatrix} 1 & 2 \\ -1 & 1 \end{bmatrix}, \qquad b = \begin{bmatrix} 5 \\ 2 \end{bmatrix}.$$

# Basic notation

- By $A \in \mathbb{R}^{m \times n}$ we denote a *matrix* having $m$ rows and $n$ columns, whose entries are real numbers;
- by $x \in \mathbb{R}^n$ we denote a *vector* of $n$ entries. We will treat an $n$-dimensional vector as a special case of a matrix, namely a matrix having $n$ rows and $1$ column (*column vector*). A *row vector*, i.e. a matrix having $1$ row and $n$ columns, will be typically denoted by $x^\top$;
- the $i$th element of vector $x$ is denoted by $x_i$:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix};$$

- we denote by $a_{ij}$ (or, sometimes, $A_{ij}$) the entry of $A$ in the $i$th row and $j$th column:

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix};$$

# Basic notation (cont.)

- we denote by $a_j$ the $j$th column of $A$:

$$A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix};$$

- we denote by $a_i^\top$ the $i$th row of $A$:

$$A = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_m^\top \end{bmatrix}.$$

Notice that above notation is ambiguous (for example, $a_1$ and $a_1^\top$ are not one the transpose[2] of the other), but the actual meaning of the symbols will be clear from the context.

[2] The transpose will be formally defined in the following.

# Matrix multiplication

# Product of two matrices

**Definition**

The product of two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{n \times p}$ is the matrix:

$$C = AB \in \mathbb{R}^{m \times p}$$

where

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}. \tag{1}$$

- Matrix multiplication is associative:
$$(AB)C = A(BC).$$
- Matrix multiplication is distributive:
$$A(B + C) = AB + AC.$$
- Matrix multiplication is, in general, not commutative. Indeed, in general, provided that both $AB$ and $BA$ exist, we have $AB \neq BA$. In the special case when $AB = BA$, we say that $A$ and $B$ *commute*.

# Product of two matrices (cont.)

The matrix product expression (1) holds also for *block matrices* (or *partitioned matrices*), i.e. matrices whose rows and columns are grouped in such a way that each"entry" is actually a submatrix:

$$\begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1p} \\ B_{21} & B_{22} & \cdots & B_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ B_{n1} & B_{n2} & \cdots & B_{np} \end{bmatrix} = \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1p} \\ C_{21} & C_{22} & \cdots & C_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ C_{m1} & C_{m2} & \cdots & C_{mp} \end{bmatrix}$$

where

$$C_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}.$$

Of course, he columns of $A$ and the rows of $B$ must be partitioned consistently, meaning that, for all $k$, $A_{ik}$ must have as many columns as the number of rows of $B_{kj}$.

# Vector-vector products

**Definition (Dot product)**

Given two vectors $x, y \in \mathbb{R}^n$, the *dot product* (or *inner product*) of $x$ and $y$ is the scalar

$$x^\top y = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=1}^{n} x_i y_i \in \mathbb{R}.$$

By definition, we have $x^\top y = y^\top x$.

# Vector-vector products (cont.)

**Definition (Outer product)**

Given two vectors $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$ the *outer product* of $x$ and $y$ is the matrix whose entries are given by $(xy^\top)_{ij} = x_i y_j$, i.e.

$$xy^\top = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} = \begin{bmatrix} x_1 y_1 & x_1 y_2 & \cdots & x_1 y_n \\ x_2 y_1 & x_2 y_2 & \cdots & x_2 y_n \\ \vdots & \vdots & \ddots & \vdots \\ x_m y_1 & x_m y_2 & \cdots & x_m y_n \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

By definition, we have $x^\top y = y^\top x$.

# Matrix-vector products

Matrix multiplication can be looked at in different ways. We will examine some of them in the following, starting from the special case of matrix-vector products.

Given $A \in \mathbb{R}^{m \times n}$ and $x \in \mathbb{R}^n$, their product is the vector

$$y = Ax \in \mathbb{R}^m.$$

A first way of interpreting $Ax$ is a stack of dot products: by writing $A$ by rows we get:

$$y = Ax = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_m^\top \end{bmatrix} x = \begin{bmatrix} a_1^\top x \\ a_2^\top x \\ \vdots \\ a_m^\top x \end{bmatrix}.$$

In words, the $i$th entry of $y$ is the dot product of the $i$th row of $A$ and $x$, i.e. $y_i = a_i^\top x$.

# Matrix-vector products (cont.)

Alternatively, by writing $A$ column-wise we get:

$$y = Ax = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \\ a_1 \\ \\ \end{bmatrix} x_1 + \begin{bmatrix} \\ a_2 \\ \\ \end{bmatrix} x_2 + \cdots + \begin{bmatrix} \\ a_n \\ \\ \end{bmatrix} x_n.$$

In other words, $y$ is a linear combinations of the columns of $A$, where the coefficients are the entries of $x$.

# Matrix-vector products (cont.)

If we multiply a matrix on the left by a row vector, we get a row vector $y^\top = x^\top A$, where $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^m$, $y \in \mathbb{R}^n$. The row vector $y^\top$ can be expressed in two ways, as before. If we write $A$ column-wise we have

$$y^\top = x^\top A = x^\top \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix} = \begin{bmatrix} x^\top a_1 & x^\top a_2 & \dots & x^\top a_n \end{bmatrix},$$

thus the $i$th entry of $y^\top$ is the inner product of $x$ and the $i$th column of $A$.

Conversely, if we write $A$ row-wise we have:

$$y^\top = x^\top A = \begin{bmatrix} x_1 & x_2 & \cdots & x_m \end{bmatrix} \begin{bmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_m^\top \end{bmatrix} = x_1 \begin{bmatrix} \quad a_1^\top \quad \end{bmatrix} + x_2 \begin{bmatrix} \quad a_2^\top \quad \end{bmatrix} + \cdots + x_m \begin{bmatrix} \quad a_m^\top \quad \end{bmatrix}$$

this $y^\top$ is a linear combination of the rows of $A$ where the coefficients are the entries of $x$.

# Matrix-matrix products

We now focus on the product $C = AB$ and show four different ways it can be thought of.

1. By partitioning $A$ row-wise and $B$ column-wise we get

$$C = AB = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_m^\top \end{bmatrix} \begin{bmatrix} b_1 & b_2 & \cdots & b_p \end{bmatrix} = \begin{bmatrix} a_1^\top b_1 & a_1^\top b_2 & \cdots & a_1^\top b_p \\ a_2^\top b_1 & a_2^\top b_2 & \cdots & a_2^\top b_p \\ \vdots & \vdots & \ddots & \vdots \\ a_m^\top b_1 & a_m^\top b_2 & \cdots & a_m^\top b_p \end{bmatrix}$$

   thus the product $C$ is a matrix whose $(i, j)$th entry is the dot product of the $i$th row of $A$ and the $j$th column of $B$.

2. By partitioning $A$ column-wise and $B$ row-wise we get

$$C = AB = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} b_1^\top \\ b_2^\top \\ \vdots \\ b_n^\top \end{bmatrix} = \sum_{i=1}^{n} a_i b_i^\top$$

   where the last equality follows from the generalization of the product matrix to partitioned matrices. Thus, the product $AB$ is written as a sum of outer products $a_i b_i^\top$, each of which is an $m \times p$ matrix.

3. By representing $B$ by columns, and interpreting the matrix-matrix multiplication as a set of matrix-vector products, we get

$$C = AB = A \begin{bmatrix} b_1 & b_2 & \cdots & b_p \end{bmatrix} = \begin{bmatrix} Ab_1 & Ab_2 & \cdots & Ab_p \end{bmatrix},$$

   thus the $j$th column of $C$ is a column vector obtained as the product of $A$ and the $j$th column of $b$: $c_j = Ab_j$.

4. Finally, by representing $A$ by rows we get:

$$C = AB = \begin{bmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_m^\top \end{bmatrix} B = \begin{bmatrix} a_1^\top B \\ a_2^\top B \\ \vdots \\ a_m^\top B \end{bmatrix},$$

thus the $i$th row of $C$ is the product of the $i$th row of $A$ and $B$: $c_i^\top = a_i^\top B$.

# Operations and properties

# Identity matrix

**Definition**

The *identity matrix* of size $n$ is the square matrix $I_n \in \mathbb{R}^{n\times n}$ with ones on the main diagonal and zeros elsewhere:

$$(I_n)_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}.$$

For example, the following are identity matrices:

$$I_1 = \begin{bmatrix} 1 \end{bmatrix},\ I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},\ \cdots,\ I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

.

For all $A \in \mathbb{R}^{m\times n}$ we have:

$$I_m A = A = A I_n.$$

# Transpose

**Definition**

Given a matrix $A \in \mathbb{R}^{m \times n}$, its *transpose* denoted by $A^\top \in \mathbb{R}^{n \times m}$ is the $n \times m$ matrix such that

$$\left(A^\top\right)_{ij} = A_{ji}.$$

For example:

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix}, \qquad A^\top = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix}.$$

The following properties hold:

- $\left(A^\top\right)^\top = A$
- $(AB)^\top = B^\top A^\top$
- $(A + B)^\top = A^\top + B^\top$

# Symmetric matrices

**Definition**

A square matrix $A \in \mathbb{R}^{n \times n}$ is said to be *symmetric* if

$$A = A^\top.$$

**Definition**

A square matrix $A \in \mathbb{R}^{n \times n}$ is said to be *skew-symmetric* or *anti-symmetric* if

$$A = -A^\top.$$

For example, the following matrices are respectively symmetric and skew-symmetric:

$$\begin{bmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{bmatrix} \qquad \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix}$$

# Symmetric matrices (cont.)

Any square matrix $A \in \mathbb{R}^{n \times n}$ can be written as the sum of a symmetric matrix and a skew-symmetric matrix, in view of the following identity:

$$A = \frac{1}{2}\left(A + A^\top\right) + \frac{1}{2}\left(A - A^\top\right).$$

For example:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 5 \\ 3 & 5 & 7 \\ 5 & 7 & 9 \end{bmatrix} + \begin{bmatrix} 0 & -1 & -2 \\ 1 & 0 & -1 \\ 2 & 1 & 0 \end{bmatrix}.$$

# Trace

**Definition**

The *trace* of a square matrix $A \in \mathbb{R}^{n \times n}$, denoted as $\mathrm{tr}(A)$ or $\mathrm{tr}\, A$, is the sum of the elements of the main diagonal:

$$\mathrm{tr}\, A = \sum_{i=1}^{n} a_{ii}.$$

The following properties hold:

- $\mathrm{tr}\, A = \mathrm{tr}\, A^\top$
- $\mathrm{tr}(A + B) = \mathrm{tr}\, A + \mathrm{tr}\, B$
- $\mathrm{tr}(\alpha A) = \alpha\, \mathrm{tr}\, A, \quad \forall \alpha \in \mathbb{R}$
- $\mathrm{tr}(AB) = \mathrm{tr}(BA)$

# Inverse

**Definition**

Let $A$ be an $n \times n$ square matrix: $A \in \mathbb{R}^{n \times n}$. If there exists a matrix $B$ such that

$$AB = BA = I_n,$$

then $B$ is called *inverse* of $A$, and denoted by $A^{-1}$.

The following properties hold true (in the identities below, we assume that all the inverses do exist):

- the inverse, if it does exist, is unique;
- $\left(A^{-1}\right)^{-1} = A$;
- $\left(A^{-1}\right)^{\top} = \left(A^{\top}\right)^{-1} \doteq A^{-\top}$;
- $(AB)^{-1} = B^{-1}A^{-1}$.

For example:

$$B = \begin{bmatrix} -1 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{bmatrix} \quad \text{is the inverse of} \quad A = \begin{bmatrix} 0 & 2 \\ 2 & 4 \end{bmatrix}.$$

# Inverse (cont.)

**Definition**

A square matrix $A$ is said to be *invertible* or *non-singular* if $A^{-1}$ exists.

It is said to be *non-invertible* or *singular* otherwise.

In the following we will give necessary and sufficient conditions for a matrix being invertible.

The inverse can be used to solve a linear systems of equations. Let the system be $Ax = b$, where $A \in \mathbb{R}^{n \times n}$ and $x, b \in \mathbb{R}^n$. Then, provided that $A$ is invertible, by multiplying both sides by $A^{-1}$ we get

$$x = A^{-1} b.$$

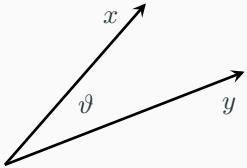
# Geometric interpretation of the dot product

The *dot product* (or *scalar product* or *inner product*) between the vectors $x$ and $y$ has the following geometrical meaning:

$$x \cdot y = \|x\| \, \|y\| \cos \vartheta$$

where $\vartheta$ is the acute angle between the two arrows that represent $x$ and $y$. When the vectors are treated as single-column matrices, the dot product is obtained by a matrix product as follows:

$$x \cdot y = x^\top y = y^\top x.$$




**Definition**

Two vectors $x$ and $y$ are said to be *orthogonal* if their dot product is zero:

$$x^\top y = 0.$$

If $x$ and $y$ are orthogonal, the arrows representing $x$ and $y$ are mutually perpendicular.

# Norms

**Definition (Vector norm)**

A *norm* is any function $\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}$ that satisfies the following properties:

1. $\|x\| > 0, \forall x \neq 0$ (positivity);
2. $\|ax\| = |a| \, \|x\| \, , \forall a \in \mathbb{R}$ (homogeneity);
3. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality).

Examples of norms are:

- Euclidean norm:

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

- $L_1$ norm:

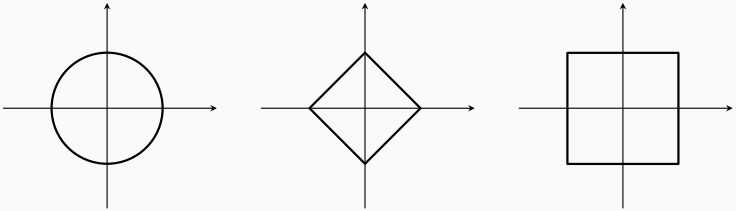
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

- $L_\infty$ norm (or *max norm*):

$$\|x\|_\infty = \max_i |x_i|$$

# Norms (cont.)

The level surfaces of the previous norms are, respectively, spheres, diamonds and cubes.



$\|\cdot\|_2$ $\|\cdot\|_1$ $\|\cdot\|_\infty$

The previous norms are a special case of the $L_p$ norm, defined as

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

for $p \in \mathbb{R}, p \geq 1$.

# Norms (cont.)

Any vector norm *induces* a corresponding matrix norm in the following way.

**Definition (Induced matrix norm)**

Let $\|\cdot\|$ be a vector norm. The corresponding induced matrix norm is defined as

$$\|A\| = \max_{\|x\|=1} \|Ax\|$$

Thus, we can define $\|A\|_1$, $\|A\|_2$, $\|A\|_\infty$.

Informally speaking, the induced $L_p$ norm of $A$ is a measure of the maximum "amplification" that a vector $x \in \mathbb{R}^n$ may incur when multiplied by $A$, when the length of the vectors are measured using the vector norm $L_p$.

**Definition (Frobenius norm)**

Given a matrix $A$, its *Frobenius norm* is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$$

Thus, the Frobenius norm is the Euclidean norm of a column vector obtained by stacking all the entries of $A$.

It can be easily checked that

$$\|A\|_F = \sqrt{\mathsf{tr}\left(A^\top A\right)}.$$

**Definition (Determinant)**

Let $A \in \mathbb{R}^{n \times n}$. The *determinant* of $A$, denoted as $\det(A)$, $\det A$ or $|A|$, is a scalar defined, by either of the following recursions:

$$\begin{cases} \det(A) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det(A_{\setminus i, \setminus j}) & (i \text{ fixed}) \\ \det(a) = a \end{cases} \quad \text{or} \quad \begin{cases} \det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j} \det(A_{\setminus i, \setminus j}) & (j \text{ fixed}) \\ \det(a) = a \end{cases}$$

where $A_{\setminus i, \setminus j}$ is a matrix obtained from $A$ by suppressing the $i$th row and the $j$th column. The determinant does not depend on fixed indices $1 \leq i \leq n$ and $1 \leq j \leq n$.

For example, the determinant of a $2 \times 2$ matrix takes the form:

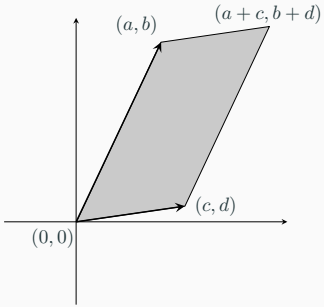
$$\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc.$$

# Determinant (cont.)

The determinant of a $2 \times 2$ matrix has the following geometric meaning: its absolute value represents the area of the parallelogram defined by the rows of the matrix, as shown in the figure.

It represents also the area of a parallelogram (in general, different) defined by the columns.

The above geometrical meaning extends to three dimensional volume and $n$-dimensional volume.



# Determinant (cont.)

The following properties hold, for $A \in \mathbb{R}^{n \times n}$:

- $\det(AB) = \det(A)\det(B)$
- $\det(\alpha A) = \alpha^n \det(A)$, where $\alpha \in \mathbb{R}$
- $\det(I) = 1$
- $\det(A^{-1}) = \dfrac{1}{\det(A)}$
- $A$ is invertible $\iff \det(A) \neq 0$
- if $A$ is invertible, then the inverse can be written as

$$A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$$

where the adjoint matrix $\operatorname{adj}(A) \in \mathbb{R}^{n \times n}$ is such that $(\operatorname{adj}(A))_{ij} = (-1)^{i+j} \det(A_{\setminus j, \setminus i})$

**Definition**

The set of vectors $x_1, x_2, \ldots, x_m \in \mathbb{R}^n$ is said to be *linearly independent* if

$$\sum_{i=1}^{m} \alpha_i x_i = 0 \quad \Longrightarrow \quad \alpha_i = 0, \ \ \forall i$$

Equivalently, a set of vectors is linearly independent if no vector can be written as a linear combination of the remaining vectors. For example, the vectors

$$x_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} \qquad x_2 = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} \qquad x_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$

are not linearly independent because $x_3 = -2x_1 + x_2$.

# Orthogonal matrices

Recall that two vectors $x, y \in \mathbb{R}^n$ are orthogonal if $x^\top y = 0$.

**Definition**

A vector $x \in \mathbb{R}^n$ is *normalized* if $\|x\|_2 = 1$.

**Definition**

A square matrix $U \in \mathbb{R}^{n \times n}$ is *orthogonal* if all its columns are *orthonormal*, i.e.:

1. normalized, and
2. orthogonal to each other (mutually orthogonal).

It follows that

$$U^\top U = I = U U^\top,$$

thus the inverse of an orthogonal matrix is its transpose.

It can be shown that, if $U$ is orthogonal, then

$$\|Ux\|_2 = \|x\|_2$$

thus operating on a vector with an orthogonal matrix will not change its Euclidean norm.

# Rank

**Definition (Column rank)**

The *column rank* of matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of columns of $A$ that constitute a linearly independent set.

**Definition (Row rank)**

The *row rank* of matrix $A \in \mathbb{R}^{m \times n}$ is the size of the largest subset of rows of $A$ that constitute a linearly independent set.

**Proposition**

For any $A \in \mathbb{R}^{m \times n}$ the column rank and the row rank are equal.

In view of the above Proposition, we can simply refer to the *rank* of a matrix $A$, denoted by $\mathbf{rank}(A)$ or $\mathbf{rank}\, A$.

# Rank (cont.)

The following properties hold:

- for $A \in \mathbb{R}^{m \times n}$, $\mathsf{rank}(A) \leq \mathsf{min}(m, n)$. If $\mathsf{rank}(A) = \mathsf{min}(m, n)$, then $A$ is said to be *full rank* (more precisely, if the rank is $n$ it is *full column rank*, if the rank is $m$ it is *full row rank*).
- $\mathsf{rank}(A) = \mathsf{rank}(A^\top)$
- $\mathsf{rank}(AB) \leq \mathsf{min}(\mathsf{rank}(A), \mathsf{rank}(B))$
- $\mathsf{rank}(A + B) \leq \mathsf{rank}(A) + \mathsf{rank}(B)$
- $A \in \mathbb{R}^{n \times n}$ is full rank $\iff \mathsf{det}(A) \neq 0$
- if $B$ is invertible, then $\mathsf{rank}(AB) = \mathsf{rank}(A)$
- if $B$ is invertible, then $\mathsf{rank}(BA) = \mathsf{rank}(A)$

**Definition (Span of a set of vectors)**

The *span* of a set of vectors $\{x_1, x_2, \ldots, x_m\}$ is the set of all vectors that can be written as a linear combination of $\{x_1, x_2, \ldots, x_m\}$:

$$\text{span}\{x_1, x_2, \ldots, x_m\} = \left\{ v : v = \sum_{i=1}^{m} \alpha_i x_i, \quad \alpha_i \in \mathbb{R} \right\}$$

For example, the span of $\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ is the whole $\mathbb{R}^2$ since any vector $v \in \mathbb{R}^2$ can be written as

$$v = \alpha_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

for some $\alpha_1, \alpha_2$.

# Span (cont.)

Let $\{x_1, x_2, \ldots, x_m\}$ be a set of vectors and suppose that $x_i$ can be expressed a linear combination of the remaining vectors:

$$x_i = \sum_{j \neq i} \alpha_j x_j.$$

It is easy to prove that:

$$\mathsf{span}(\{x_1, x_2, \ldots, x_m\}) = \mathsf{span}(\{x_1, x_2, \ldots, x_m\} \setminus x_i).$$

As a consequence, the span of a set of vectors is equal to the span of the largest subset of independent vectors.

# Subspaces

**Definition (Subspace)**

Let $\mathcal{V}$ be a subset of $\mathbb{R}^n$: $\mathcal{V} \subseteq \mathbb{R}^n$. It is said to be a *subspace* of $\mathbb{R}^n$ if it is closed with respect to linear combinations:

$$u, v \in \mathcal{V} \quad \Longrightarrow \quad \alpha u + \beta v \in \mathcal{V}, \ \ \forall \alpha, \beta \in \mathbb{R}.$$

Clearly, $\mathbb{R}^n$ is a subspace of itself. Moreover, the span of any set of vectors of $\mathbb{R}^n$ is a subspace of $\mathbb{R}^n$. Conversely, any subspace of $\mathbb{R}^n$ can be written as the span of a suitable set of vectors of $\mathbb{R}^n$.

**Definition (Basis)**

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^n$. A set of vectors $\{v_1, \ldots, v_m\}$ is said to be a *basis* for $\mathcal{V}$ if both the following properties hold:

- $V = \textbf{span}\{v_1, \ldots, v_m\}$ ($\mathcal{V}$ is *generated* by $\{v_1, \ldots, v_m\}$)
- the set $\{v_1, \ldots, v_m\}$ is linearly independent.

For example, the sets

$$\left\{\begin{bmatrix}1\\0\end{bmatrix},\begin{bmatrix}0\\1\end{bmatrix}\right\} \quad \text{and} \quad \left\{\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}-1\\0\end{bmatrix}\right\}$$

are both basis of $\mathbb{R}^2$. The set

$$\left\{\begin{bmatrix}1\\2\end{bmatrix},\begin{bmatrix}-1\\0\end{bmatrix},\begin{bmatrix}5\\1\end{bmatrix}\right\}$$

is not a basis, because the set of vectors is not linearly independent. A subspace has an infinite number of different bases, nut they all share the same number of elements, as state by the following proposition.

**Proposition**

If $\{v_1, \dots v_p\}$ and $\{w_1, \dots, w_q\}$ are both bases of the same (sub)space $\mathcal{V}$, then $p = q$.

Thus we can formulate the following definition.

**Definition (Dimension)**

The dimension of a (sub)space $\mathcal{V}$, denoted by $\dim \mathcal{V}$, is the number of vectors of any basis of $\mathcal{V}$. The null subspace $\mathcal{V} = \{0\}$ has dimension zero.

# Range and nullspace

**Definition (Range of a matrix)**

The *range* (or *columnspace*, or *image*) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\mathsf{im}\, A$, is the subspace of $\mathbb{R}^m$ given by the span of the columns of $A$:

$$\mathsf{im}\, A = \{v \in \mathbb{R}^m : v = Ax, \quad x \in \mathbb{R}^n\}\,.$$

Properties:

- $\dim(\mathsf{im}\, A) = \mathsf{rank}(A)$
- $\mathsf{im}\, A = \mathsf{im}(AA^\top)$

**Definition (Rowspace)**

The span of the rows of $A \in \mathbb{R}^{m \times n}$ (a subset of $\mathbb{R}^n$) is said to be the *rowspace* of $A$ and denoted by $\mathsf{im}(A^\top)$. Its dimension is equal to the rank of $A$: $\dim(\mathsf{im}(A^\top)) = \mathsf{rank}(A^\top) = \mathsf{rank}(A)$.

# Range and nullspace (cont.)

**Definition (Nullspace)**

The *nullspace* (or *kernel*) of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\ker A$, is the set of all vectors of $\mathbb{R}^n$ that equal $0$ when multiplied by $A$:

$$\ker A = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Since $Au = 0, Bv = 0 \implies A(\alpha u + \beta v) = 0$, it follows that $\ker A$ is a subspace.

The following fundamental theorem establishes a relationship between $\dim(\ker A)$ and $\dim(\operatorname{im} A)$.

**Theorem (Rank-nullity theorem)**

Let $A \in \mathbb{R}^{m \times n}$. Then

$$n = \underbrace{\dim(\operatorname{im} A)}_{\operatorname{rank} A} + \dim(\ker A).$$

# Range and nullspace (cont.)

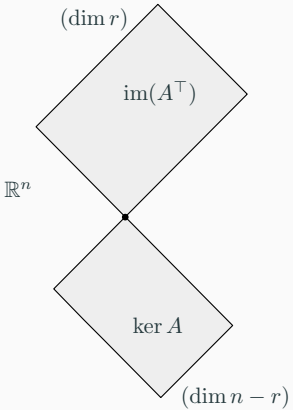
Another fundamental result is the following.

**Theorem**

For $A \in \mathbb{R}^{m \times n}$ we have

$$\left\{ w : w = u + v, u \in \mathsf{im}(A^\top), v \in \ker A \right\} = \mathbb{R}^n \quad \text{and} \quad \mathsf{im}(A^\top) \cap \ker A = \{0\}.$$

In other words, the rowspace and the nullspace of a matrix have trivial intersection and together span the whole $\mathbb{R}^n$. They are said to be *orthogonal complements*, denoted as $\mathsf{im}(A^\top) = \ker(A)^\perp$.



# QR decomposition

**Theorem (QR decomposition)**

Every matrix $A \in \mathbb{R}^{m \times n}$ with linearly independent columns can be uniquely factored as

$$A = QR,$$

in which the columns of $Q \in \mathbb{R}^{m \times n}$ are an orthonormal basis for $\operatorname{im} A$ and $R \in \mathbb{R}^{n \times n}$ is an upper-triangular matrix with positive diagonal entries.

For example:

$$\begin{bmatrix} 0 & -20 & -14 \\ 3 & 27 & -4 \\ 4 & 11 & -2 \end{bmatrix} = \frac{1}{25} \begin{bmatrix} 0 & -20 & -15 \\ 15 & 12 & -16 \\ 20 & -9 & 12 \end{bmatrix} \begin{bmatrix} 5 & 25 & -4 \\ 0 & 25 & 10 \\ 0 & 0 & 10 \end{bmatrix}$$

**Definition**

Given a square matrix $A \in \mathbb{R}^{n \times n}$, we say that $\lambda \in \mathbb{C}$ is an *eigenvalue* of $A$ and $x \in \mathbb{C}^n$ is the corresponding *eigenvector* if

$$Ax = \lambda x, \quad x \neq 0. \tag{2}$$

The pair $x, \lambda$ is sometimes referred to as an *eigenpair*.

Intuitively, the above definition means that multiplying $A$ by an eigenvector $x$ results is a vector having the same direction as $x$ but scaled by a factor $\lambda$.

Eq. (2) is equivalent to

$$(\lambda I - A)x = 0, \quad x \neq 0.$$

But $(\lambda I - A)x = 0$ has a non-zero solution if and only if $(\lambda I - A)$ is not full rank, i.e. if and only if

$$\mathsf{det}(\lambda I - A) = 0.$$

Thus the eigenvalues are the roots of the $n$th degree polynomial

$$p(\lambda) = \mathsf{det}(\lambda I - A),$$

# Eigenvalues and eigenvectors (cont.)

called the *characteristic polynomial*. As a consequence, an $n \times n$ matrix has $n$ (not necessarily distinct) eigenvalues $\lambda_1, \ldots, \lambda_n$.

If $\bar{\lambda}$ is an eigenvalue, the set of the corresponding eigenvectors (whose union with $\{0\}$ is a subspace called *eigenspace*) is the set of non-zero solutions of the system

$$(\bar{\lambda} I - A)x = 0.$$

The following properties hold, for any $A \in \mathbb{R}^{n \times n}$.

- $\mathsf{tr}\, A = \sum_{i=1}^{n} \lambda_i$;
- $\det A = \prod_{i=1}^{n} \lambda_i$ (thus, a singular matrix has at least one zero eigenvalue);
- the rank of $A$ is equal to the number of non-zero eigenvalues of $A$;
- if $A$ is non-singular and $(x, \lambda)$ is en eigenpair of $A$, then $(x, 1/\lambda)$ is an eigenpair of $A^{-1}$;
- if $A$ is triangular, its eigenvalues are the entries of main diagonal;
- if $A$ is symmetric, then its eigenvalues are real;
- if $A$ is symmetric, then its eigenvectors are orthogonal.

# Eigenvalues and eigenvectors (cont.)

**Theorem (Schur decomposition)**

Any matrix $A \in \mathbb{R}^{n \times n}$ can be expressed as

$$QUQ^*,$$

where $U$ is an upper triangular matrix and $Q$ is a unitary matrix (i.e. a matrix whose conjugate transpose $Q^*$ is also its inverse).

Notice that, in general, both $U$ and $Q$ are complex valued.

Example (real eigenvalues):

$$\begin{bmatrix} 7 & -2 \\ 12 & -3 \end{bmatrix} = \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 3 & -14 \\ 0 & 1 \end{bmatrix} \frac{1}{\sqrt{5}} \begin{bmatrix} 1 & 2 \\ -2 & 1 \end{bmatrix}.$$

Example (complex eigenvalues):

$$\begin{bmatrix} 1 & 1 \\ -2 & 3 \end{bmatrix} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1-j \\ 1+j & -1 \end{bmatrix} \begin{bmatrix} 2+j & -1+2j \\ 0 & 2-j \end{bmatrix} \frac{1}{\sqrt{3}} \begin{bmatrix} 1 & 1-j \\ 1+j & -1 \end{bmatrix}.$$

**Theorem (Schur diagonalization)**

Any symmetric matrix $A \in \mathbb{R}^{n \times n}$, having eigenvalues $\lambda_1, \ldots, \lambda_n$, can be expressed as

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

where $\Lambda = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$ and $T$ is orthogonal. The columns $t_1 \ldots t_n$ of $T$ are eigenvectors of $T$ associated to, respectively, $\lambda_1, \ldots, \lambda_n$ (in other words, $(\lambda_i, t_i)$ is an eigenpair).

Example:

$$\begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

# Quadratic forms

**Definition (Quadratic form)**

Given a square matrix $A \in \mathbb{R}^{n \times n}$ and a vector $x \in \mathbb{R}^n$, the scalar value $x^\top Ax$ is called a *quadratic form*:

$$x^\top Ax = \sum_{i=1}^{n} x_i (Ax)_i = \sum_{i=1}^{n} x_i (\sum_{j=1}^{n} a_{ij} x_j)_i = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j.$$

Since the transpose of a scalar is equal to the scalar itself we have

$$x^\top Ax = (x^\top Ax)^\top = x^\top A^\top x.$$

Moreover, by averaging the first and last member, which are equal, we get

$$x^\top Ax = x^\top \underbrace{\left( \frac{A}{2} + \frac{A^\top}{2} \right)}_{\text{symmetric part of A}} x$$

thus only the symmetric part of $A$ contributes to the quadratic form.

**Definition (Positive definite matrix)**

A matrix $A \in \mathbb{R}^{n \times n}$ is *positive definite* (denoted by $A \succ 0$) if

$$x^\top A x > 0, \quad \forall x \neq 0, x \in \mathbb{R}^n.$$

**Definition (Positive semidefinite matrix)**

A matrix $A \in \mathbb{R}^{n \times n}$ is *positive semidefinite* (denoted by $A \succeq 0$) if

$$x^\top A x \geq 0, \quad \forall x \in \mathbb{R}^n.$$

The following matrices are, respectively positive definite and positive semidefinite:

$$A_1 = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 0 & 0 \\ 0 & 10 \end{bmatrix}.$$

Indeed, the respective quadratic forms are $x^\top A_1 x = 5x_1^2 + 10x_2^2 > 0, \forall x \neq 0$ and $x^\top A_2 x = 10x_2^2 \geq 0, \forall x$. Notice that the definitions above imply that a positive definite matrix is also positive semidefinite.

**Definition (Negative definite matrix)**

A matrix $A \in \mathbb{R}^{n \times n}$ is *negative definite* (denoted by $A \prec 0$) if

$$x^\top A x < 0, \quad \forall x \neq 0, x \in \mathbb{R}^n.$$

**Definition (Negative semidefinite matrix)**

A matrix $A \in \mathbb{R}^{n \times n}$ is *negative semidefinite* (denoted by $A \preceq 0$) if

$$x^\top A x \leq 0, \quad \forall x \in \mathbb{R}^n.$$

# Quadratic forms (cont.)

Recall that all eigenvalues of a symmetric matrix are real. The following results are useful.

**Lemma**

A symmetric matrix $A$ is positive (negative) definite if and only if all its eigenvalues are strictly positive (negative):

$$A \succ 0 \quad \iff \quad \lambda_i \underset{(<)}{>} 0, \quad i = 1, \dots, n$$

where $\lambda_i$ denotes the $i$-th eigenvalue of $A$.

**Lemma**

A symmetric matrix $A$ is positive (negative) semidefinite if and only if all its eigenvalues are non-negative (non-positive):

$$A \succeq 0 \quad \iff \quad \lambda_i \underset{(\leq)}{\geq} 0, \quad i = 1, \dots, n$$

Observe that, for non-symmetric matrices, the eigenvalue check must be carried out on the symmetric part.

# Quadratic forms (cont.)

The following properties are useful:

- for any $A \in \mathbb{R}^{m \times n}$, the matrices $(A^\top A) \in \mathbb{R}^{n \times n}$ and $(AA^\top) \in \mathbb{R}^{m \times m}$ are both positive semidefinite;
- the gradient $\nabla$ (regarded as a column vector) of a quadratic form is:

$$\nabla x^\top Ax = 2Ax \qquad \text{(if } A \text{ is symmetric);}$$

- the Hessian matrix $H(\cdot)$ of a quadratic form is:

$$H(x^\top Ax) = 2A \qquad \text{(if } A \text{ is symmetric).}$$

**Theorem (Cholesky)**

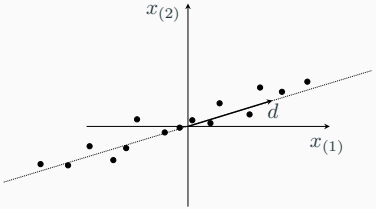
Any real symmetric and positive definite matrix $A$ can be decomposed uniquely as the product of lower triangular matrix having strictly positive eigenvalues and its transpose:

$$A = LL^\top.$$

For example:

$$\begin{bmatrix} 4 & 12 & -16 \\ 12 & 37 & -43 \\ -16 & -43 & 98 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 6 & 1 & 0 \\ -8 & 5 & 3 \end{bmatrix} \begin{bmatrix} 2 & 6 & -8 \\ 0 & 1 & 5 \\ 0 & 0 & 3 \end{bmatrix}.$$

# Principal Component Analysis (PCA)




An important application of Schur diagonalization is the Principal Component Analysis (PCA). The PCA has many motivations and interpretations. The most common is "find the direction along which the data varies the most".

Suppose we are given a set of $n$ points $x_1, \dots, x_n \in \mathbb{R}^m$. Assume that the set has zero mean:

$$\mu \doteq \frac{1}{n} \sum_{i=1}^{n} x_i = 0.$$

(If the set has nonzero mean, we can subtract $\mu$ to all the vectors). Let a direction in $\mathbb{R}^m$ be represented by a unit vector $d \in \mathbb{R}^m$. The component of $x_i$ along the direction $d$ is thus given by the dot product $d^\top x_i$. The amount of "variation of the data set along $d$" can be quantified as the empirical variance of the components, i.e.

$$\frac{1}{n} \sum_{i=1}^{n} \left( d^\top x_i \right)^2 .$$

# Principal Component Analysis (PCA) (cont.)

Finding the direction of maximal variance amounts to solving the following optimization problem:

$$\underset{\|d\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{n} \left(d^\top x_i\right)^2 = \underset{\|d\|_2=1}{\operatorname{argmax}} \sum_{i=1}^{n} \left(d^\top x_i\right)\left(x_i^\top d\right),$$

where the division by $n$ has been omitted since the minimizer is the same.

By collecting the data in the matrix

$$X = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

we have

$$d^\top X = \begin{bmatrix} d^\top x_1 & d^\top x_2 & \dots & d^\top x_n \end{bmatrix} \quad \text{and} \quad X^\top d = \begin{bmatrix} x_1^\top d \\ x_2^\top d \\ \vdots \\ x_n^\top d \end{bmatrix}$$

thus the objective function can be written more compactly and the problem becomes

$$\underset{\|d\|_2=1}{\operatorname{argmax}} \; d^\top X X^\top d.$$

# Principal Component Analysis (PCA) (cont.)

The matrix $XX^\top$ is called *covariance matrix* and is symmetric and positive semidefinite by construction. Thus, according to the Schur's diagonalization theorem, it can be diagonalized by an orthogonal transform:

$$XX^\top = T\Lambda T^\top = T \begin{bmatrix} \lambda_1 & 0 & \dots & \\ 0 & \lambda_2 & 0 & \dots \\ & & \ddots & \\ 0 & 0 & & 0 \\ 0 & 0 & 0 & \lambda_m \end{bmatrix} T^\top$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$. Thus, the objective function becomes

$$d^\top T\Lambda T^\top d,$$

and we search for the maximizing unit vector $d$. By letting

$$y = T^\top d,$$

and observing that $\|y\|_2 = 1 \Leftrightarrow \|d\|_2 = 1$, the problem becomes

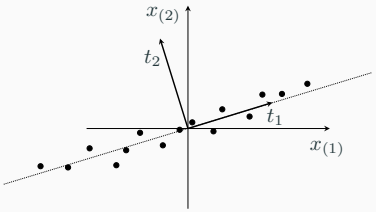
$$\underset{\|y\|_2=1}{\text{argmax}}\, y^\top \Lambda y,$$

whose solution, since the eigenvalues appear in decreasing order, is $y = [1 \;\; 0 \;\; \dots \;\; 0]^\top$. Thus, the maximizing $d$ is

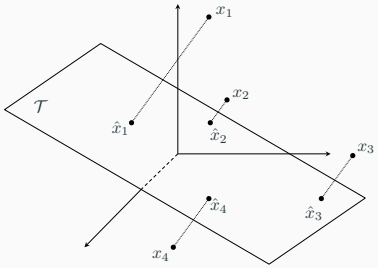
$$d = Ty = t_1,$$

i.e. the first column of $T$, corresponding to an eigenvector associated to the largest eigenvalue.

From the above, it should be clear that the subsequent columns of $T$, i.e. $t_2, t_3 \dots$ represent directions, orthogonal to each other and to $t_1$, exhibiting a decreasing variance.



# Principal Component Analysis (PCA) (cont.)




The PCA can be used for dimensionality reduction, by projecting the original data in the subspace spanned by the first $k < m$ columns of $T$. If $x \in \mathbb{R}^m$ is a vector of the original set, its projection onto the subspace $\mathcal{T}$ spanned by $\{t_1, \ t_2, \ \ldots \ , t_k\}$ is the vector $\hat{x} \in \mathbb{R}^m$:

$$\hat{x} = DD^\top x, \qquad \text{where} \quad D = [t_1 \ \ t_2 \ \ \ldots \ \ t_k].$$

The coordinates of $\hat{x}$ with respect to the basis of $\{t_1, \ , t_2 \ \ \ldots \ \ , t_k\}$ of $\mathcal{T}$ are given by the vector

$$z = D^\top x$$

and are called the first $k$ *principal components* of $x$.

# Solutions to linear systems of equations

# Solutions to linear systems of equations

Consider the linear system of equations:

$$Ax = b. \tag{3}$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$.

For $b = 0$ the system is said to be *homogeneous* and (3) becomes

$$Ax = 0.$$

From the definition of nullspace, the set of the solutions is thus $\mathsf{ker}(A)$. Two cases may occur:

1. $\mathsf{rank}\, A = n$ (i.e. $A$ is full column rank): in that case by the rank-nullity theorem, $\mathsf{dim}(\mathsf{ker}(A)) = 0$. Thus $\mathsf{ker}(A) = \{0\}$ and the trivial solution $x = 0$ is the only solution;
2. $\mathsf{rank}\, A < n$: there exist infinite solutions, precisely the set of solutions is a subspace of $\mathbb{R}^n$ of dimension

$$\mathsf{dim}(\mathsf{ker}(A)) = n - \mathsf{rank}(A) \geq 1.$$

# Solutions to linear systems of equations (cont.)

For $b \neq 0$ the system $Ax = b$ is said to be *non-homogeneous*. The following theorem provides a necessary and sufficient condition for the existence of solutions.

**Theorem**

Consider the system

$$Ax = b,$$

where $A \in \mathbb{R}^{m \times n}$, $x \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$. Then,

$$\text{the system admits a solution} \quad \iff \quad \mathsf{rank}(A) = \mathsf{rank}\left(\begin{bmatrix} A & b \end{bmatrix}\right).$$

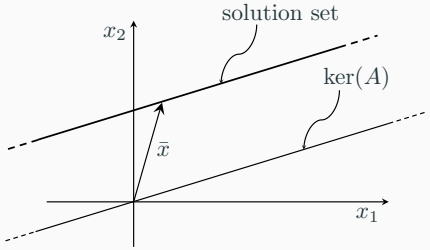
Moreover, if it admits a solution, and denoting by $\bar{x}$ any specific solution to $Ax = b$, the entire solution set can be described as

$$\{x : x = x_0 + \bar{x} \quad \text{where} \quad x_0 \in \mathsf{ker}(A)\}. \tag{4}$$

# Solutions to linear systems of equations (cont.)

Observe that:

- the condition $\mathsf{rank}(A) = \mathsf{rank}\left(\begin{bmatrix} A & b \end{bmatrix}\right)$ is equivalent to $b \in \mathsf{im}(A)$;
- if $\mathsf{rank}(A) = n$ and the system admits a solution, then the solution is unique;
- if $A$ is square and full rank, the condition is certainly satisfied and there exists a unique solution, which is $x = A^{-1}b$;
- the set of solutions (4) is a linear variety, as represented in figure.



# A useful property

In the following we will use the following property.

**Property**

For any $A \in \mathbb{R}^{m \times n}$:

$$\mathsf{rank}(A) = \mathsf{rank}(A^\top A) = \mathsf{rank}(AA^\top).$$

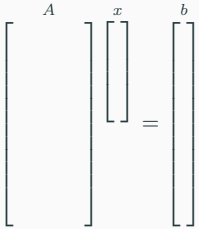
**Proof.**
We first prove the first equality. Let $x \in \ker(A^\top A)$. Then $(A^\top A)x = 0$ and multiplying by $x^\top$ to the left we get $x^\top A^\top Ax = 0$ which implies that $\|Ax\| = 0$, thus $x \in \ker(A)$. We have proven that $x \in \ker(A^\top A) \implies x \in \ker(A)$. The opposite implication is obvious, thus

$$x \in \ker(A^\top A) \iff x \in \ker(A).$$

In other words, $A$ and $A^\top A$ have the same nullspace. Since they have the same number of columns $n$, by the rank-nullity theorem, the dimension of their columnspaces must be the same. The second equality follows by the fact that $\mathsf{rank}\, M = \mathsf{rank}\, M^\top$. □

# Least-squares (approximate) solution to overdetermined systems

When trying to fit a model with real, noisy, data, *overdetermined* systems of equations are frequently encountered.




We want to solve $Ax = b$ for $x$, but $b \notin \mathbf{im}\, A$. In that case the system admits no solution, but still one may want to find the $x$ such that $Ax$ is the closest possible to $b$. If we measure the distance using the Euclidean norm we can state the problem:

$$\min_x \|Ax - b\|_2^2 .$$

# Least-squares (approximate) solution to overdetermined systems (cont.)

In the frequent case when rank $A = n$ (full column rank), the problem admits a unique solution. Indeed:

$$\|Ax - b\|_2^2 = (Ax - b)^\top (Ax - b) = (x^\top A^\top - b^\top)(Ax - b) = x^\top A^\top Ax - 2x^\top A^\top b + b^\top b.$$

By taking the gradient with respect to $x$ and equating to zero, we get

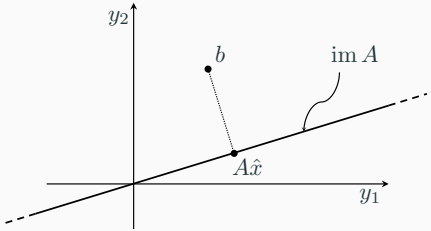
$$2A^\top Ax - 2A^\top b = 0 \qquad \text{or} \qquad A^\top Ax = A^\top b.$$

Notice that $A^\top A \in \mathbb{R}^{n \times n}$ and its rank is $n$. Thus it is invertible, and, as a consequence, the unique solution is

$$\hat{x} = (A^\top A)^{-1} A^\top b. \tag{5}$$

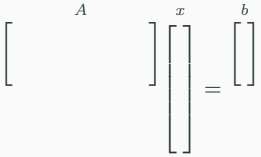
# Least-squares (approximate) solution to overdetermined systems (cont.)

The least-squares approximate solution admits a geometric interpretation. Indeed, $A\hat{x}$ is (by definition) the point in $\mathbf{im}\ A$ that is closest to $b$, i.e. is the projection of $b$ onto $\mathbf{im}\ A$.



# Minimum-norm solution to underdetermined systems

Another situation that may occur is that of an *underdetermined* system (more variables than independent equations), typically having the form




Assuming that $m > n$ and $\text{rank}\, A = m$ (full row rank) infinite solutions exist.

It can be useful to find the *minimum norm solution*, i.e. solve the problem:

$$\begin{aligned} &\min \|x\|_2 \\ &\text{subject to} \quad Ax = b \end{aligned}.$$

The minimum norm solution $\hat{x}$ can be shown to be:

$$\hat{x} = A^\top (AA^\top)^{-1} b.$$

# Minimum-norm solution to underdetermined systems (cont.)

Indeed, the minimization problem can be stated as:

$$\begin{aligned} &\textbf{min } x^\top x \\ &\text{subject to} \quad Ax = b \end{aligned} .$$

By introducing the Lagrange multipliers vector $\lambda \in \mathbb{R}^m$, we get the Lagrangian

$$\mathcal{L}(x, \lambda) = x^\top x + \lambda^\top (Ax - b).$$

The stationarity conditions are

$$\nabla_x \mathcal{L}(x, \lambda) = 2x + A^\top \lambda = 0 \qquad \text{and} \qquad \nabla_\lambda \mathcal{L}(x, \lambda) = Ax - b = 0.$$

From the first we get $x = -A^\top \lambda / 2$ thus, from the second, we have $\lambda = -2(AA^\top)^{-1} b$. Notice that $AA^\top \in \mathbb{R}^{m \times m}$ is certainly invertible, being rank $m$. Finally, by substituting in the first we get

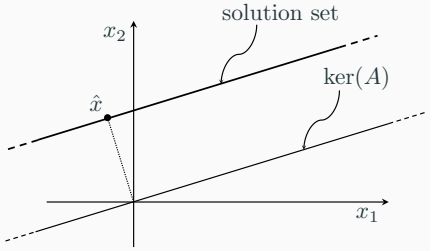
$$\hat{x} = A^\top (AA^\top)^{-1} b. \tag{6}$$

# Minimum-norm solution to underdetermined systems (cont.)

The minimum norm solution is orthogonal to $\ker A$, indeed, for any $y \in \ker A$ we have $Ay = 0$ and also $y^\top A^\top = 0$. Thus

$$y^\top \hat{x} = y^\top A^\top (AA^\top)^{-1} b = 0.$$

The minimum norm solution admits a geometric interpretation. Indeed, $\hat{x}$ is the projection of the origin of $\mathbb{R}^n$ onto the solution set of $Ax = b$.



# Singular value decomposition

# Singular value decomposition

Any matrix $A \in \mathbb{R}^{m \times n}$ can be seen as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$ (it associates $Ax \in \mathbb{R}^m$ to $x \in \mathbb{R}^n$).

The *singular value decomposition* of $A$ reveals a lot about this map.

**Theorem (Singular Value Decomposition)**

Any matrix $A \in \mathbb{R}^{m \times n}$ can be written as

$$A = U \mathbf{\Sigma} V^\top,$$

where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices and $\mathbf{\Sigma} \in \mathbb{R}^{m \times n}$ has the form:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad \mathbf{\Sigma}_1 = \begin{bmatrix} \sigma_1 & 0 & \dots & \\ 0 & \sigma_2 & 0 & \dots \\ \vdots & \vdots & \ddots & \\ 0 & \dots & 0 & \sigma_p \end{bmatrix} \in \mathbb{R}^{p \times p}$$

where $\sigma_1 \geq \sigma_2 \geq \dots \sigma_p \geq 0$, and $p = \min\{m, n\}$.

# Singular value decomposition (cont.)

**Definition**

The real values $\sigma_1, \sigma_2, \ldots, \sigma_p$ are called the *singular values* of $A$.

**Definition**

The columns $u_1, u_2, \ldots, u_m$ of $U$ are called the *left singular vectors* of $A$.

The columns $v_1, v_2, \ldots, v_n$ of $V$ are called the *right singular vectors* of $A$.

Observe that, for $1 \leq i \leq p$:

$$Av_i = U\Sigma V^\top v_i = U\Sigma \begin{bmatrix} 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \end{bmatrix} \leftarrow i\text{th row} = U \begin{bmatrix} 0 \\ \vdots \\ \sigma_i \\ 0 \\ \vdots \end{bmatrix} = \sigma_i u_i. \tag{7}$$

Similarly, we get

$$A^\top u_i = \sigma_i v_i. \tag{8}$$

# Singular value decomposition (cont.)

Multiplying (7) to the left by $A^\top$ and substituting (8) we get:

$$(A^\top A)v_i = \sigma_i^2 u_i, \quad i = 1, \dots, p,$$

thus $(v_i, \sigma_i^2)$ is an eigenpair of $A^\top A$. Similarly we obtain:

$$(AA^\top)u_i = \sigma_i^2 v_i, \quad i = 1, \dots, p,$$

thus $(u_i, \sigma_i^2)$ is an eigenpair of $AA^\top$.

Indeed, the following property holds true.

**Property**

Let $A \in \mathbb{R}^{m \times n}$.
If $n \geq m$, the singular values of $A$ are the square root of the eigenvalues of $AA^\top$.
If $m \geq n$, the singular values of $A$ are the square root of the eigenvalues of $A^\top A$.

# Singular value decomposition (cont.)

**Property**

The rank of $A$ equals the number of non-zero singular values of $A$.

**Proof.**
It is sufficient to observe that, since $U$ and $V^\top$ are nonsingular:

$$\mathsf{rank}(A) = \mathsf{rank}(U\Sigma V^\top) = \mathsf{rank}(\Sigma).$$

□

# Singular value decomposition (cont.)

The singular value decomposition provides bases for $\mathbf{im}\, A$ and $\mathbf{ker}\, A$, as shown by the next proposition.

**Proposition**

If $A = U\Sigma V^\top$ is a singular value decomposition of $A$, and $\mathbf{rank}(A) = r$, then:

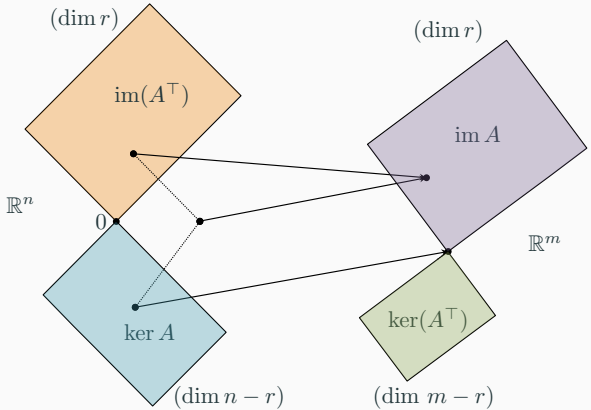
- the first $r$ columns of $U$ are a basis for $\mathbf{im}\, A$;
- the last $n - r$ columns of $V$ are a basis for $\mathbf{ker}\, A$.

Moreover, the Frobenius norm and the $2$-norm of $A$ can be characterized in terms of singular value decomposition:

$$\|A\|_F^2 = \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_p^2$$
$$\|A\|_2 = \sigma_1.$$

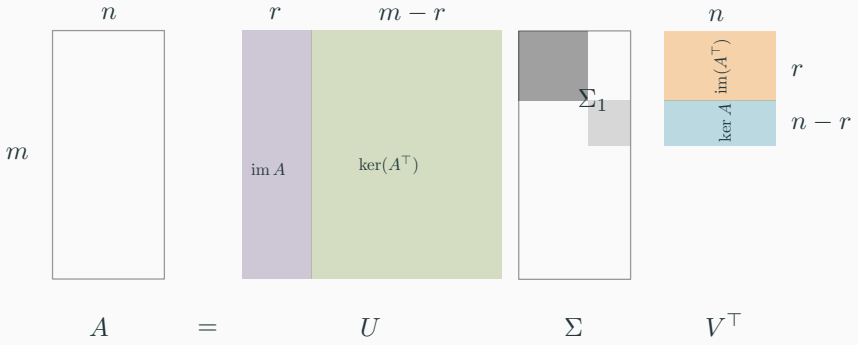
# Four subspaces

The diagram below shows the four subspaces associated to $A$ (sometimes called the four fundamental subspaces) and their relationship to the linear map $y = Ax$.



# Four subspaces (cont.)

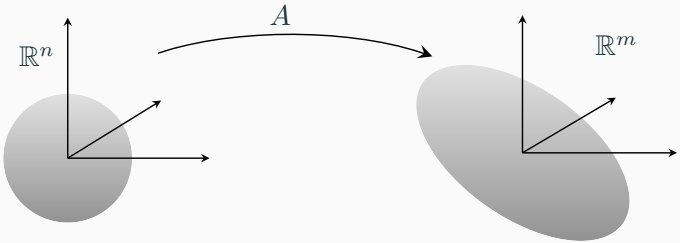
The singular value decomposition provides orthogonal bases for the four subspaces, as represented in the figure below. The colors of the partitions of $U$ and $V$ correspond to the subspaces of the previous slide.



# Geometrical interpretation

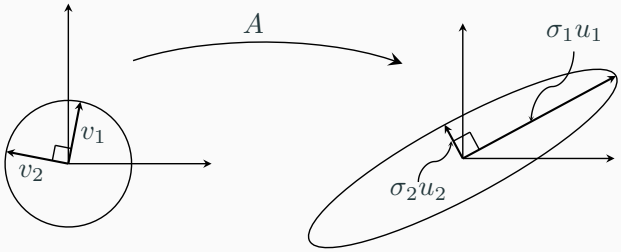
The singular values admit the following geometrical interpretation.

The singular values of $A \in \mathbb{R}^{m \times n}$ represent the length of the semiaxes of the hyperellipse in $\mathbb{R}^m$ obtained by applying the linear map $A$ to the unit hypersphere of $\mathbb{R}^n$ (centered in the origin).



# Geometrical interpretation (cont.)

As an example, consider the case $\mathbb{R}^2 \longrightarrow \mathbb{R}^2$:




We have

$$\|Av_1\| = \|\sigma_1 u_1\| = \sigma_1 \|u_1\| = \sigma_1$$

$$\|Av_2\| = \|\sigma_2 u_2\| = \sigma_2 \|u_2\| = \sigma_2$$

# Singular values and eigenvalues

The singular values are different from the eigenvalues. In particular:

- the singular values are defined for any matrix, while the eigenvalues exist only for square matrices;
- the singular values are always real;
- the singular values are always non negative;
- if $A$ is square, then the singular values can be computed by taking the square root of the eigenvalues of either $A^\top A$ or $AA^\top$:

$$\sigma_i = \sqrt{\lambda_i(A^\top A)} = \sqrt{\lambda_i(AA^\top)}.$$

# Example

Let

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

thus the map is from $\mathbb{R}^3$ to $\mathbb{R}^2$. We can compute the singular values as $\sqrt{\lambda_i(AA^\top)}$:

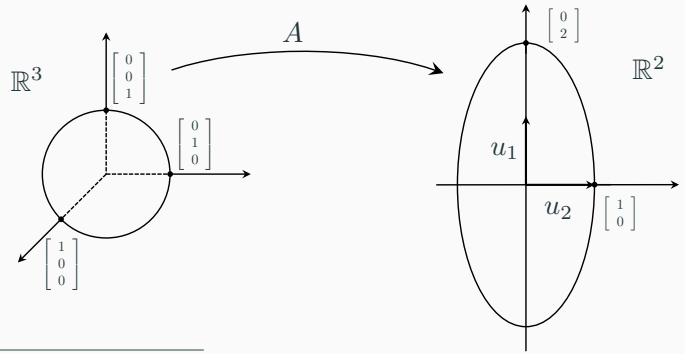
$$AA^\top = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix} \qquad \Longrightarrow \qquad \lambda_1 = 4, \lambda_2 = 1$$

thus $\sigma_1 = 2, \sigma_2 = 1$.

# Example (cont.)

It can be easily checked that a singular value decomposition is[3]

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \left[\begin{array}{cc|c} 2 & 0 & 0 \\ 0 & 1 & 0 \end{array}\right] \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$




[3] the singular value decomposition is not unique: for instance one can change the sign of both $u_i$ and $v_i$ and get a different SVD, but this is not the only source of ambiguity.

# Compact SVD and dyadic expansion

**Proposition (Compact SVD)**

Let $A \in \mathbb{R}^{m\times n}$ and let $\mathsf{rank}(A) = r$. If $A = U\Sigma V^\top$ is the singular value decomposition of $A$, it can be shown that:

$$A = U_r \Sigma_r V_r^\top, \tag{9}$$

where $U_r = [u_1 \;\; u_2 \;\; \dots u_r]$, $V_r = [v_1 \;\; v_2 \;\; \dots v_r]$, and $\Sigma_r = \mathsf{diag}\{\sigma_1, \sigma_2, \dots, \sigma_r\}$.

Moreover, $A$ admits the following *dyadic expansion*:

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top = \sum_{i=1}^{r} \sigma_i Z_i, \qquad \text{(where, clearly, } Z_i \in \mathbb{R}^{m\times n} \text{ and } \mathsf{rank}(Z_i) = 1\text{)}$$

For instance, the matrix of the previous example has the following compact SVD and dyadic expansion:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} = 2 \underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{u_1 v_1^\top} + 1 \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{u_2 v_2^\top}$$

# Low rank approximation

**Theorem**

Let $k < r = \mathsf{rank}(A)$ and $A_{(k)} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. Then:

$$\min_{\mathsf{rank}(B) \leq k} \|A - B\|_2 = \left\|A - A_{(k)}\right\|_2 = \sigma_{k+1}.$$

In other words, $A_{(k)}$ is the best approximation of rank $\leq k$ of $A$, with respect to the 2-norm.

**Theorem**

Let $k < r = \mathsf{rank}(A)$ and $A_{(k)} = \sum_{i=1}^{k} \sigma_i u_i v_i^\top$. Then:

$$\min_{\mathsf{rank}(B) \leq k} \|A - B\|_F = \left\|A - A_{(k)}\right\|_F = \sum_{i=k+1}^{r} \sigma_i.$$

In other words, $A_{(k)}$ is the best approximation of rank $\leq k$ of $A$, with respect to the Frobenius norm.

Notice that $A_{(k)}$ can be written as:

$$A_{(k)} = U_k \mathbf{\Sigma}_k V_k^\top$$

where $U_k = [u_1 \;\; u_2 \;\; \dots u_k]$, $V_k = [v_1 \;\; v_2 \;\; \dots v_k]$, and $\mathbf{\Sigma}_r = \mathsf{diag}\{\sigma_1, \sigma_2, \dots, \sigma_k\}$..

# Non-trivial approximate solutions to homogeneous systems

Consider the homogeneous system $Ax = 0$, where $A \in \mathbb{R}^{m \times n}$. If $A$ is full column rank, the only solution is $x = 0$. Otherwise, the system admits non-trivial solutions (in other words, $A$ has non-trivial nullspace).

When trying to fit a model with real, noisy, data, it may well happen that $A$ is full column rank due to noise or measurement errors, although from a theoretical point of view it should not. Still, one may be interested in non-trivial approximate solutions. In particular, an approximate solution can be found by solving:

$$\min_{\|x\|=1} \|Ax\|^2 , \tag{10}$$

where we state the constraint $\|x\| = 1$ to exclude the trivial solution $x = 0$ (the choice of the value $1$ is arbitrary, basically because we are interested in the direction of $x$).

The solution of (10) is easily proven to be the right singular vector associated to the smallest singular value.

**Theorem**

Let $A \in \mathbb{R}^{m \times n}$ and let $A = U\Sigma V^\top$ be its singular value decomposition. The solution of the following constrained minimization problem:

$$\min_{\|x\|=1} \|Ax\|^2,$$

is $x = v_n$, where $v_n$ is the $n$th column of $V$, i.e. the right singular vector associated to the smallest singular value.

**Proof.**
Recalling that $U$ and $V$ are orthogonal, we have:

$$\min_{\|x\|=1} \|Ax\|^2 = \min_{\|x\|=1} \left\| U\Sigma V^\top x \right\|^2 = \min_{\|x\|=1} \left\| \Sigma V^\top x \right\|^2 = \min_{\|y\|=1} \|\Sigma y\|^2 = \min_{\|y\|=1} \sum_i \sigma_i^2 y_i.$$

Since the singular values appear in $\Sigma$ in decreasing order, the minimum is achieved for $y = [0 \quad \dots \quad 0 \quad 1]^\top$, thus the solution is $x = v_n$. □

# Orthogonal Procrustes problem

The Orthogonal Procrustes[4] problem amounts to finding the orthogonal transformation $W$ that renders the transformed matrix $WB$ as close as possible to $A$ (in Frobenius norm). Its solution can be expressed in terms of an SVD.

**Theorem (Orthogonal Procrustes problem)**

Given two matrices $A$ and $B$, the solution to the problem

$$\min_{W^\top W = I} \|A - WB\|_F^2$$

is $W = VU^\top$, where $BA^\top = U\Sigma V^\top$ is the singular value decomposition of $BA^\top$.

An important special case, frequently encountered in Computer Vision and in Robotics, is $B = I$, when the aim is to find the orthogonal matrix $W$ which is closest to a given $A$:

$$\min_{W^\top W = I} \|A - W\|_F^2 .$$

The solution corresponds to substituting $\Sigma$ with the identity matrix in the SVD of $A$.

[4] *"He killed Damastes, surnamed Procrustes, by compelling him to make his own body fit his bed, as he had been wont to do with those of strangers."* (Plutarch, Life of Theseus).

We have seen that the PCA of matrix $X$ amounts to performing an eigenvalue decomposition of the (symmetric and positive semidefinite) matrix $XX^\top$:

$$XX^\top = T\Lambda T^\top. \tag{11}$$

Now, considering the SVD of $X$:

$$X = U\Sigma V^\top,$$

we can express the matrix $XX^\top$ as follows:

$$XX^\top = \left(U\Sigma V^\top\right)\left(U\Sigma V^\top\right)^\top = U\Sigma \underbrace{V^\top V}_{I}\Sigma U^\top = U\Sigma^2 U^\top \tag{12}$$

Since $\Sigma^2$ is diagonal, (12) is an eigenvalue decomposition as well as (11) is, and the columns $u_i$ of $U$ are eigenvectors of $XX^\top$ associated to the eigenvalues $\sigma_i^2$.

Thus PCA reduces to computing the SVD of $X$ (*without forming* $XX^\top$) and, denoting by $U_k$ the matrix containing the first $k$ columns of $U$, the vector of the first $k$ principal components of $x$ is $U_k^\top x$.

Suppose we are given $n$ points in $\mathbb{R}^m$

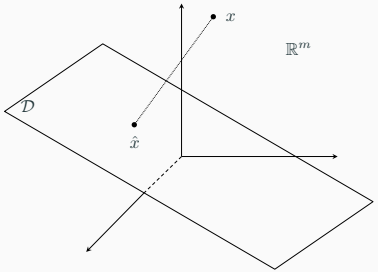
$$X = \left[ \begin{array}{cccc} x_1 & x_2 & \dots & x_n \end{array} \right] \in R^{m \times n}$$

Let

$$D = \left[ \begin{array}{cccc} d_1 & d_2 & \dots & d_k \end{array} \right] \in R^{m \times k}$$

where $k < m$ be an orthonormal basis for a subspace $\mathcal{D}$ of $\mathbb{R}^m$. The projection onto $\mathcal{D}$ of a vector $x \in \mathbb{R}^m$ is

$$\hat{x} = DD^\top x.$$



We can *reduce the dimensionality* of the set $\{x_1, \dots, x_n\}$ by encoding each $x_i$ as the vector $z_i \in \mathbb{R}^k$ of the components of $\hat{x}_i$ along the basis $\{d_1, \dots, d_k\}$:

$$z_i = D^\top x_i.$$

For a given $k < m$ (i.e for a given target dimensionality) what is a reasonable criterion for the choice of $D$? From $z_i \in \mathbb{R}^k$ we can reconstruct $\hat{x}_i = Dz_i \in \mathbb{R}^m$; thus, a reasonable criterion is the *minimization of the sum of the squared reconstruction errors*, i.e.

$$D_{\text{opt}} = \underset{D^\top D = I}{\text{argmin}} \sum_{i=1}^{n} \left\| x_i - \underbrace{DD^\top x_i}_{\hat{x}_i} \right\|_2^2, \tag{13}$$

where the constraint guarantees that the columns of $D$ form an orthonormal basis. Recalling the Frobenius matrix norm, Eq. (13) can be written in compact form as

$$D_{\text{opt}} = \underset{D^\top D = I}{\text{argmin}} \left\| X - \underbrace{DD^\top X}_{\doteq B} \right\|_F^2. \tag{14}$$

Now observe that:

- since $D$ is rank $k$ we have:

$$\text{rank } B \leq k$$

- hence, problem (14) can be seen as a constrained low-rank approximation problem (it is constrained because $B$ must be of the form $DD^\top X$ where $D$ has orthogonal, unit norm columns).

If $X = U\Sigma V^\top$ is the SVD of $X$, the solution of the unconstrained problem is well-known to be

$$B = U_k \Sigma_k V_k^\top,$$

where $U_k = [u_1 \;\; u_2 \;\; \dots u_k]$, $V_k = [v_1 \;\; v_2 \;\; \dots v_k]$, and $\Sigma_k = \text{diag}\{\sigma_1, \sigma_2, \dots, \sigma_k\}$ (see the second theorem of slide 80).

We now show that the unconstrained solution satisfies the constraint and, as a consequence, is the solution of the constrained problem. More precisely, we show that the unconstrained solution has the form $DD^\top X$ with $D = U_k$.

Indeed we have:

$$
\begin{aligned}
U_k U_k^\top X &= U_k U_k^\top U \Sigma V^\top \\
&= U_k \begin{bmatrix} I_k & 0 \end{bmatrix} \begin{bmatrix} \Sigma_k & 0 \\ 0 & \star \end{bmatrix} \begin{bmatrix} V_k^\top \\ \star \end{bmatrix} \\
&= U_k \begin{bmatrix} \Sigma_k & 0 \end{bmatrix} \begin{bmatrix} V_k^\top \\ \star \end{bmatrix} \\
&= U_k \Sigma_k V_k^\top
\end{aligned}
$$

where $\star$ denotes a submatrix that does not affect the result.

# References

Kolter, Z. (2019). Linear Algebra Review and Reference. Accessed: October 13, 2019.

Meyer, C. D. (2000). *Matrix analysis and applied linear algebra*. SIAM.

Strang, G. (2016). *Introduction to Linear Algebra, 5th Edition*. Wellesley - Cambridge Press.

554SM –Fall 2020

Linear Algebra Review

END