Statistical Learning II

Lecture 1 - Recap of maths

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
10.09.2025

Recap of Linear Algebra

The bread of statistical learning

The Euclidean space

The Euclidean space $\mathbb{R}^d$ is the vector space of d-tuples:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d \quad (\mathbb{R}^{d \times 1})$$

"column vector"

The Euclidean space

The Euclidean space $\mathbb{R}^d$ is the vector space of d-tuples:

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} \in \mathbb{R}^d \quad (\mathbb{R}^{d \times 1})$$

"column vector"

Recall, $\mathbb{R}^d$ is a vector space of dimension d with basis:

$$\boldsymbol{e}_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$$

Position i

The Euclidean space

The Euclidean space is endowed with an inner (or scalar) product

$$\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d \qquad \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{d} u_i v_i$$

The Euclidean space

The Euclidean space is endowed with an inner (or scalar) product

$$\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d \qquad \langle \boldsymbol{u}, \boldsymbol{v} \rangle = \sum_{i=1}^{d} u_i v_i$$

Which induces a natural notion of distance and size:

$$||\boldsymbol{u}||_2^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle = \sum_{i=1}^{d} u_i^2 \qquad d(\boldsymbol{u}, \boldsymbol{v}) = ||\boldsymbol{u} - \boldsymbol{v}||_2^2$$

"Euclidean or ℓ_2 norm" "Euclidean distance"

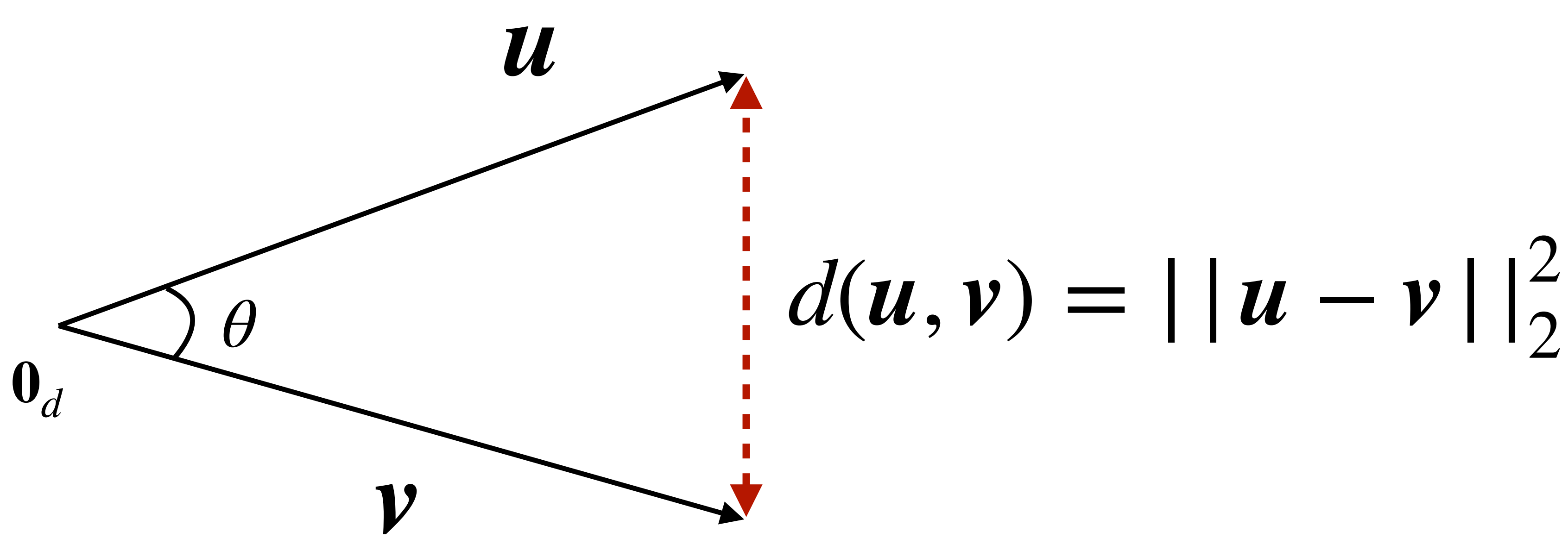
We say two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ are orthogonal if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$

Euclidean geometry

$$||\boldsymbol{u}||_2^2 = \langle \boldsymbol{u}, \boldsymbol{u} \rangle = \sum_{i=1}^{d} u_i^2 \qquad d(\boldsymbol{u}, \boldsymbol{v}) = ||\boldsymbol{u} - \boldsymbol{v}||_2^2$$

"Euclidean or ℓ_2 norm" "Euclidean distance"

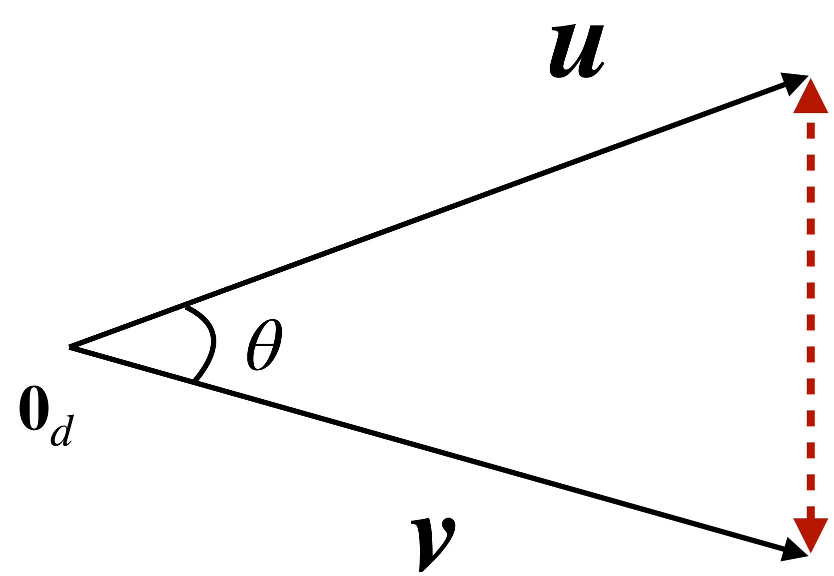
They correspond to our intuitive notion of geometry in the plane



$$\cos(\theta) = \frac{\langle \boldsymbol{u}, \boldsymbol{v} \rangle}{||\boldsymbol{u}||_2 ||\boldsymbol{v}||_2}$$

Euclidean geometry

They correspond to our intuitive notion of geometry in the plane



$$d(\boldsymbol{u},\boldsymbol{v}) = ||\boldsymbol{u}-\boldsymbol{v}||_2^2$$

$$\cos(\theta) = \frac{\langle \boldsymbol{u},\boldsymbol{v}\rangle}{||\boldsymbol{u}||_2||\boldsymbol{v}||_2}$$

In particular, we say two vectors $\boldsymbol{u},\boldsymbol{v} \in \mathbb{R}^d$ are orthogonal if

$$\langle \boldsymbol{u},\boldsymbol{v}\rangle = 0$$

Other norms

One can define other notions of size in $\mathbb{R}^d$

$$||\boldsymbol{u}||_p = \left(\sum_{i=1}^{d} u_i^p\right)^{1/p} \qquad p \geq 1$$

"ℓ_p norm"

Other norms

One can define other notions of size in $\mathbb{R}^d$

$$||\boldsymbol{u}||_p = \left(\sum_{i=1}^{d} u_i^p\right)^{1/p} \qquad p \geq 1$$

"ℓ_p norm"

⚠ $||\cdot||_p$ is not associated to an inner product for $p \neq 2$

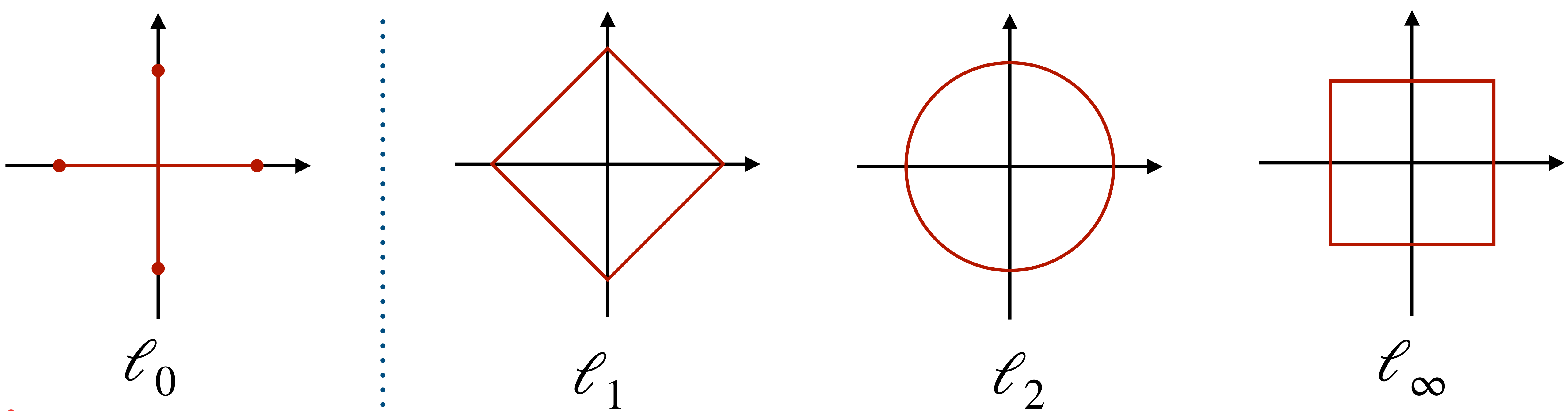
Other norms

One can define other notions of size in $\mathbb{R}^d$

$$||\boldsymbol{u}||_p = \left(\sum_{i=1}^{d} u_i^p \right)^{1/p} \qquad p \geq 1$$

"ℓ_p norm"

$||\cdot||_p$ is not associated to an inner product for $p \neq 2$



Not a norm

Matrices

A real-valued matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is a table of real numbers.

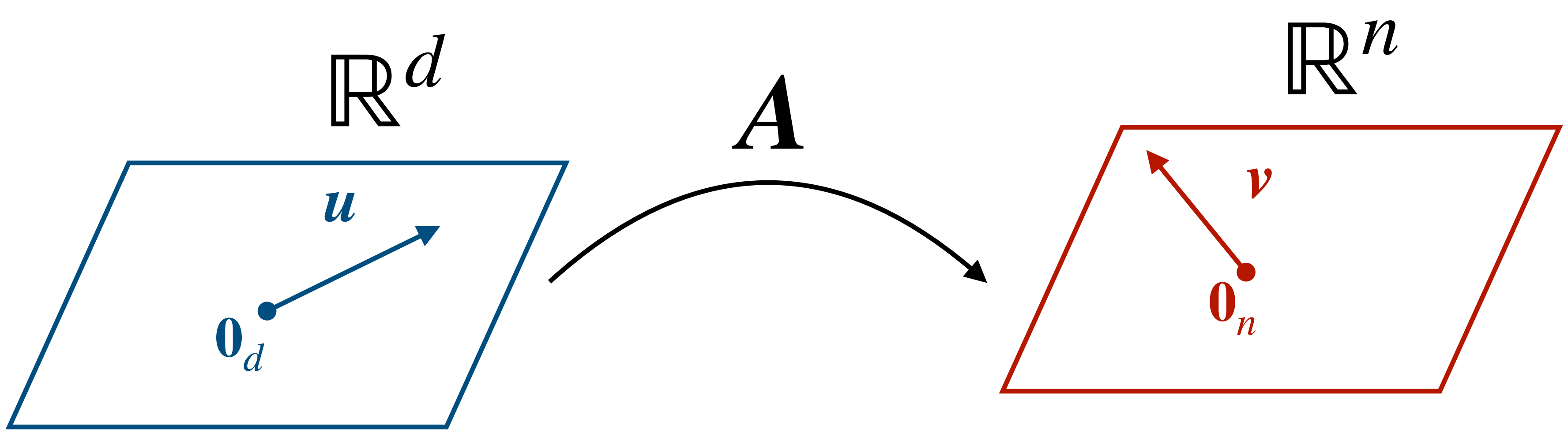
$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

Matrices

A real-valued matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is a table of real numbers.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

It is most often used to describe the coordinates of linear transformations $\boldsymbol{A} : \mathbb{R}^d \to \mathbb{R}^n$ with respect to a basis.



$$v_i = \sum_{j=1}^{d} a_{ij} u_j$$

Matrices

A real-valued matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is a table of real numbers.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

$\boldsymbol{A}_1$ $\boldsymbol{A}_2$ $\boldsymbol{A}_d$

- The columns of $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ are vectors $\boldsymbol{A}_i \in \mathbb{R}^n$ with $(\boldsymbol{A}_i)_j = a_{ij}$

  "Column space" $\mathrm{col}(\boldsymbol{A}) = \mathrm{span}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_d) \subset \mathbb{R}^n$

Matrices

A real-valued matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is a table of real numbers.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \begin{matrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_2 \\ \\ \boldsymbol{a}_n \end{matrix} \in \mathbb{R}^{n \times d}$$

- The columns of $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ are vectors $\boldsymbol{A}_i \in \mathbb{R}^n$ with $(\boldsymbol{A}_i)_j = a_{ij}$

  "Column space" $\mathrm{col}(\boldsymbol{A}) = \mathrm{span}(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_d) \subset \mathbb{R}^n$

- The rows of $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ are vectors $\boldsymbol{a}_j \in \mathbb{R}^d$ with $(\boldsymbol{a}_j)_i = a_{ij}$

  "Row space" of $\mathrm{row}(\boldsymbol{A}) = \mathrm{span}(\boldsymbol{a}_1, \cdots, \boldsymbol{a}_n) \subset \mathbb{R}^d$

Flattening matrices

The space of matrices $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is itself a vector space of dimension nd. Therefore we can identify:

$$\mathbb{R}^{n \times d} \simeq \mathbb{R}^{nd}$$

By flattening the matrices into vectors.

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \mapsto \begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_{1n} \\ a_{21} \\ \vdots \end{bmatrix} \in \mathbb{R}^{nd}$$

Rank of a matrix

- The rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the dimension of column space

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{col}(\boldsymbol{A}))$$

This is equivalent to the number of independent columns.

Rank of a matrix

- The rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the dimension of column space

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{col}(\boldsymbol{A}))$$

This is equivalent to the number of independent columns.

Proposition

$$\operatorname{rank}(\boldsymbol{A}) = \dim(\operatorname{col}(\boldsymbol{A})) = \dim(\operatorname{row}(\boldsymbol{A}))$$

Rank of a matrix

- The rank of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is the dimension of column space

$$\mathrm{rank}(\boldsymbol{A}) = \dim(\mathrm{col}(\boldsymbol{A}))$$

This is equivalent to the number of independent columns.

Proposition

$$\mathrm{rank}(\boldsymbol{A}) = \dim(\mathrm{col}(\boldsymbol{A})) = \dim(\mathrm{row}(\boldsymbol{A}))$$

- A matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is said to be full-rank if

$$\mathrm{rank}(\boldsymbol{A}) = \min(n, d)$$

Another point of view

- Alternatively, we can see the column space $\mathbf{col}(\boldsymbol{A}) \subset \mathbb{R}^n$ as
The image of the associated linear map.

$$\mathrm{im}(\boldsymbol{A}) = \mathrm{col}(\boldsymbol{A}) = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \text{ for some } \boldsymbol{u} \in \mathbb{R}^{\mathrm{d}}\}$$

Another point of view

- Alternatively, we can see the column space $\mathrm{col}(\boldsymbol{A}) \subset \mathbb{R}^n$ as The image of the associated linear map.

$$\mathrm{im}(\boldsymbol{A}) = \mathrm{col}(\boldsymbol{A}) = \{\boldsymbol{v} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \text{ for some } \boldsymbol{u} \in \mathbb{R}^{\mathrm{d}}\}$$

- The null-space or kernel of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ is defined as:

$$\ker(\boldsymbol{A}) = \{\boldsymbol{u} \in \mathbb{R}^d : \boldsymbol{A}\boldsymbol{u} = 0\}$$

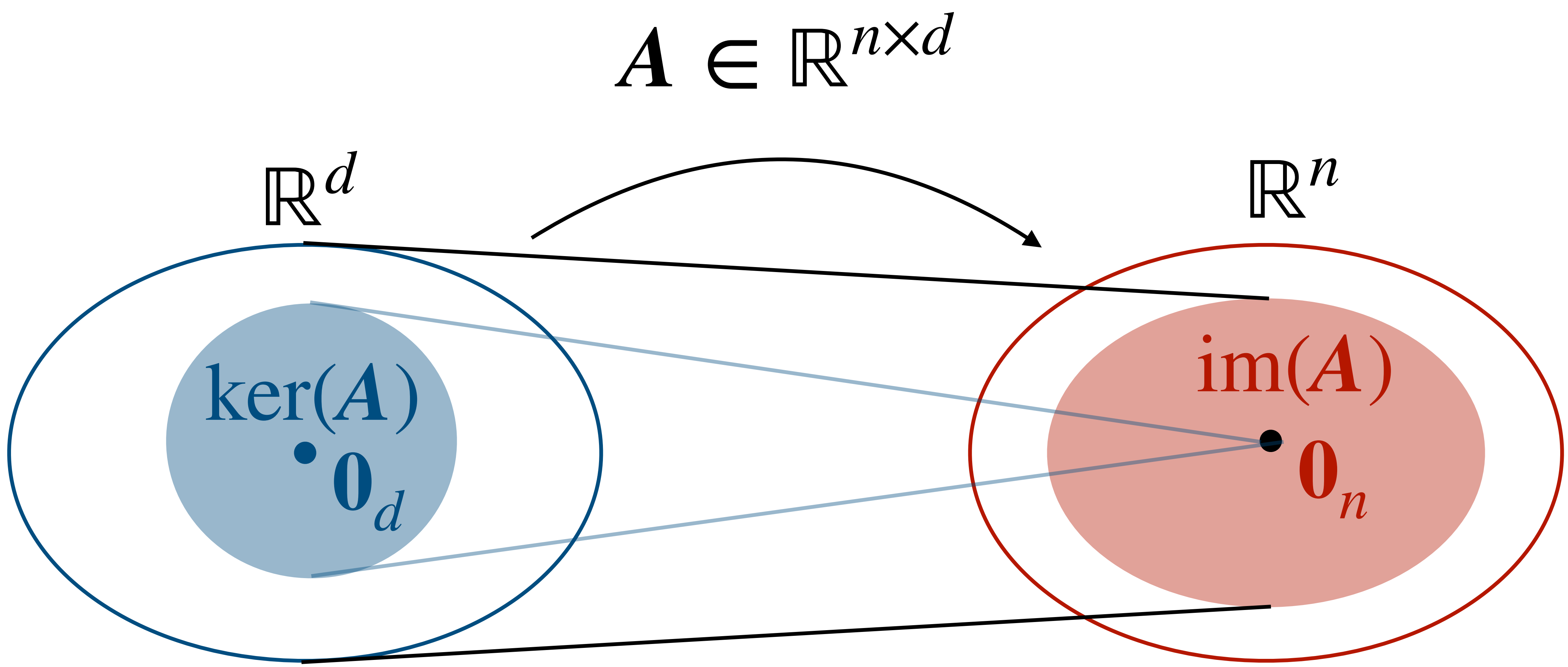
Note that $\ker(\boldsymbol{A}) \subset \mathbb{R}^d$
and $\boldsymbol{0} \in \ker(\boldsymbol{A})$

Image and null-space

Proposition

Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ denote a linear map. We have:

$$\operatorname{rank}(\boldsymbol{A}) + \dim(\ker(\boldsymbol{A})) = \mathrm{n}$$



Matrix inverse

A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is said to be invertible if there exists $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ such that:

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_d$$

In this case, we denote $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

Matrix inverse

A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is said to be invertible if there exists $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ such that:

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_d$$

In this case, we denote $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

For any invertible matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$
$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.

Matrix inverse

A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is said to be invertible if there exists $\boldsymbol{B} \in \mathbb{R}^{d \times d}$ such that:

$$\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}_d$$

In this case, we denote $\boldsymbol{B} = \boldsymbol{A}^{-1}$.

For any invertible matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$
$(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$.

Proposition

A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is invertible
if and only if it is full-rank

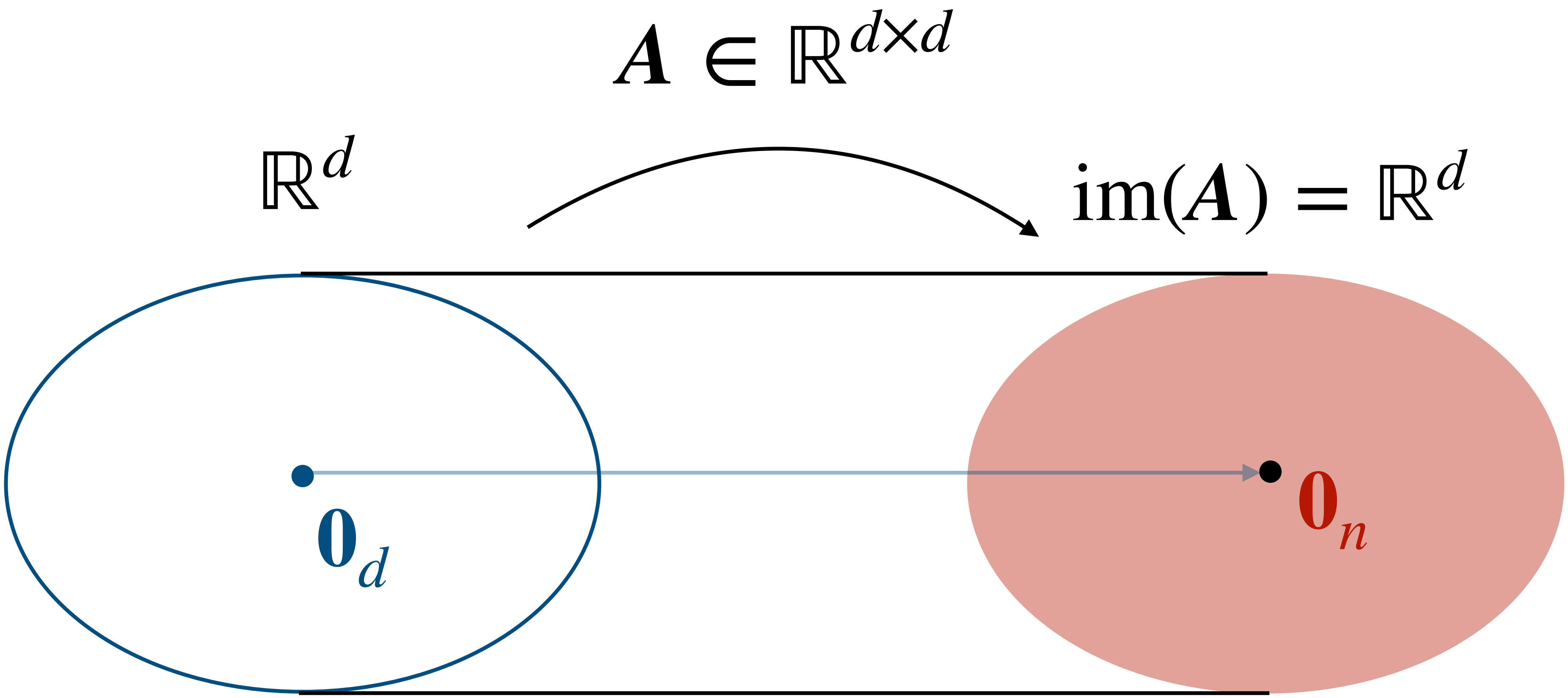
$$\text{rank}(\boldsymbol{A}) = d$$

Matrix inverse

Proposition

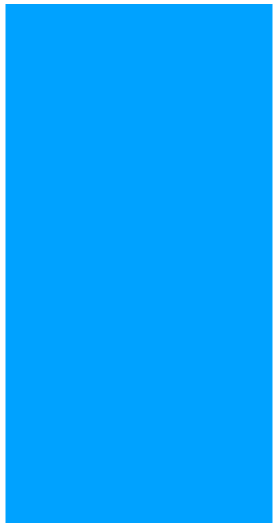
A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is invertible if and only if it is full-rank

$$\mathrm{rank}(\boldsymbol{A}) = d$$



Matrix transpose

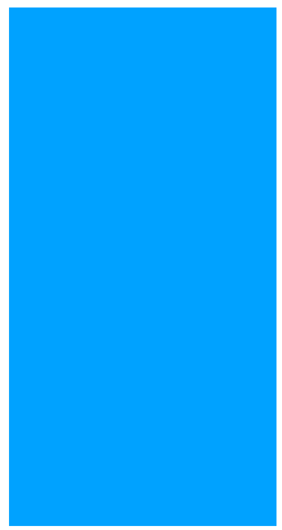
- The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with elements a_{ij} the matrix with $\boldsymbol{A}^\top \in \mathbb{R}^{d \times n}$ with elements a_{ji}

$\boldsymbol{A} =$ 

$\boldsymbol{A}^\top =$ 

Matrix transpose

- The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with elements a_{ij} the matrix with $\boldsymbol{A}^\top \in \mathbb{R}^{d \times n}$ with elements a_{ji}

$\boldsymbol{A} =$ 

$\boldsymbol{A}^\top =$ 

- We have:

$$(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$$

$$(a\boldsymbol{A} + b\boldsymbol{B})^\top = a\boldsymbol{A}^\top + b\boldsymbol{B}^\top$$

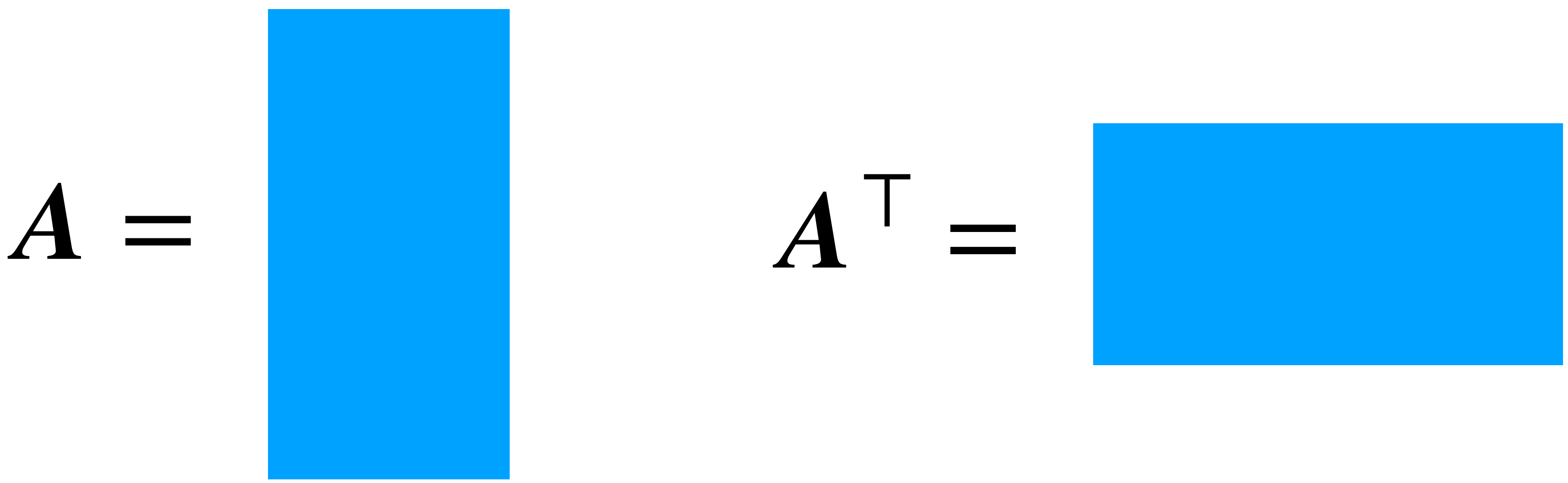
$$(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$$

$$(\boldsymbol{A}\boldsymbol{B})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$$

🧐 Exercise: check this.

Matrix transpose

- The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with elements a_{ij} the matrix with $\boldsymbol{A}^\top \in \mathbb{R}^{d \times n}$ with elements a_{ji}



- Note that by seeing $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d \times 1}$ as column vectors, we can also write the Euclidean inner product as:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle = \boldsymbol{u}^\top \boldsymbol{v}$$

Exercise: check this.

Matrix trace

- The trace of a square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is the sum of its diagonal:

$$\operatorname{Tr} \boldsymbol{A} = \sum_{i=1}^{d} a_{ii}$$

Matrix trace

- The trace of a square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is the sum of its diagonal:

$$\mathrm{Tr}\,\boldsymbol{A} = \sum_{i=1}^{d} a_{ii}$$

- It satisfies:

$$\mathrm{Tr}\,\boldsymbol{AB} = \mathrm{Tr}\,\boldsymbol{BA}$$

$$\mathrm{Tr}\,(a\boldsymbol{A} + b\boldsymbol{B}) = a\mathrm{Tr}\,\boldsymbol{A} + b\mathrm{Tr}\,\boldsymbol{B}$$

$$\mathrm{Tr}\,\boldsymbol{A}^\top = \mathrm{Tr}\,\boldsymbol{A}$$

🧐 Exercise: check this.

Symmetric matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is symmetric if $\boldsymbol{A}^\top = \boldsymbol{A}$

Symmetric matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is symmetric if $\boldsymbol{A}^\top = \boldsymbol{A}$

For any $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n \times n}$ are symmetric matrices.

Symmetric matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is symmetric if $\boldsymbol{A}^\top = \boldsymbol{A}$

⚠️ For any $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{d\times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n\times n}$ are symmetric matrices.

Letting $\boldsymbol{a}_i \in \mathbb{R}^d$ denote the rows of $\boldsymbol{A} \in \mathbb{R}^{n\times d}$, we have:

$$(\boldsymbol{A}\boldsymbol{A}^\top)_{ij} = \langle \boldsymbol{a}_i, \boldsymbol{a}_j \rangle$$

🧐 Exercise: check this.

Note: a similar representation holds for columns of $\boldsymbol{A}$

Orthogonal matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is orthogonal if $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$

Orthogonal matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is orthogonal if $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$

  Orthogonal matrices preserve the norm and distance between vectors (they are isometries):

$$||\boldsymbol{A}\boldsymbol{u}||_2 = ||\boldsymbol{u}||_2$$

Exercise: check this.

Orthogonal matrices

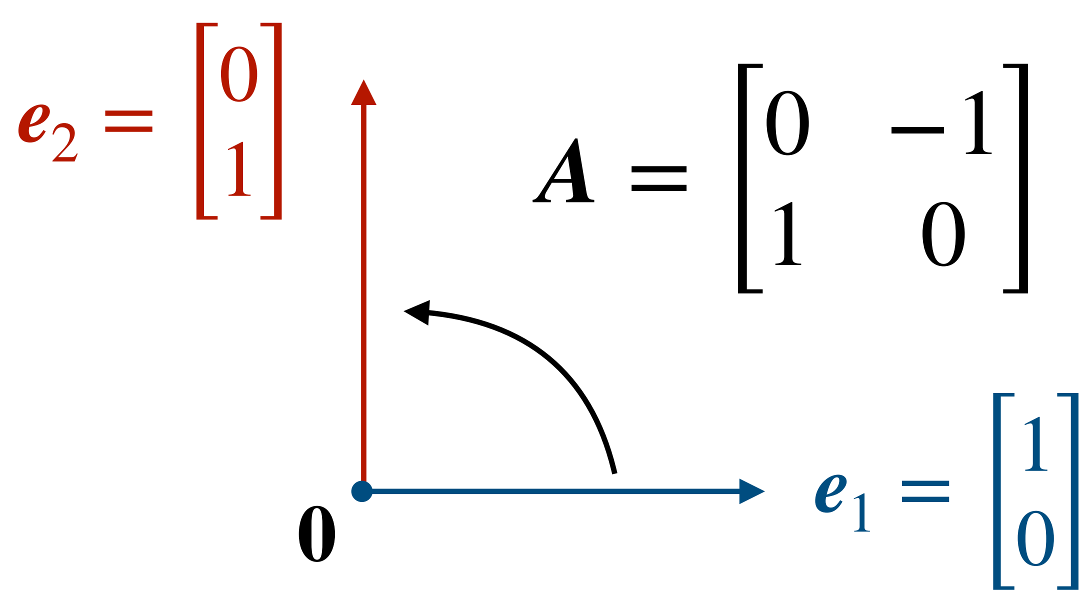
- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is orthogonal if $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$

Orthogonal matrices preserve the norm and distance between vectors (they are isometries):

$$||\boldsymbol{A}\boldsymbol{u}||_2 = ||\boldsymbol{u}||_2$$

Exercise: check this.

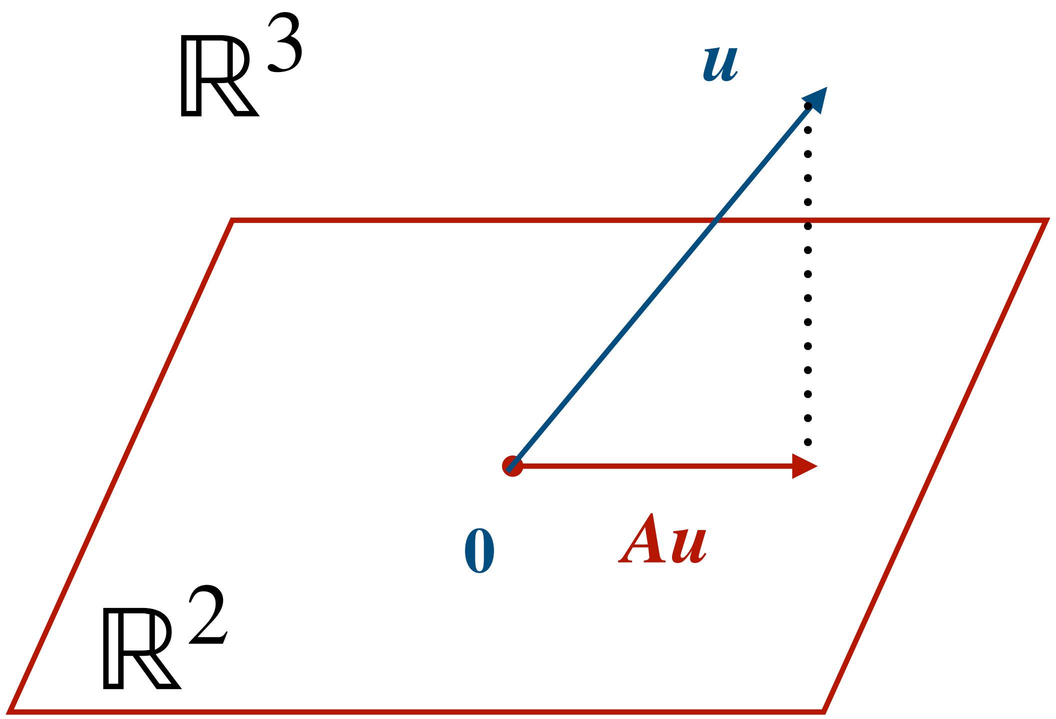
Geometrically, they define rotations



Projection matrix

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a projection if $\boldsymbol{A}^2 = \boldsymbol{A}$

  Moreover, if $\boldsymbol{A}$ is also orthogonal, we call it a orthogonal projection.



$$\boldsymbol{P} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Projection matrix

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a projection if $\boldsymbol{A}^2 = \boldsymbol{A}$

  Moreover, if $\boldsymbol{A}$ is also orthogonal, we call it a orthogonal projection.

Proposition

Any $\boldsymbol{v} \in \mathbb{R}^d$ can be uniquely written as:

$$\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{A}\boldsymbol{v} \qquad \boldsymbol{u} \in \ker(\boldsymbol{A})$$

Projection matrix

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is a projection if $\boldsymbol{A}^2 = \boldsymbol{A}$

  Moreover, if $\boldsymbol{A}$ is also orthogonal, we call it a orthogonal projection.

Proposition

Any $\boldsymbol{v} \in \mathbb{R}^d$ can be uniquely written as:

$$\boldsymbol{v} = \boldsymbol{u} + \boldsymbol{A}\boldsymbol{v} \qquad \boldsymbol{u} \in \ker(\boldsymbol{A})$$

⚠ The only projection matrix which is invertible is the identity.

Eigen-(values, vectors)

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ denote a square matrix. An eigenvector is a vector that is only re-scaled under the action of $\boldsymbol{A}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Where $\lambda \in \mathbb{R}$ is known as an eigenvalue.

Eigen-(values, vectors)

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ denote a square matrix. An eigenvector is a vector that is only re-scaled under the action of $\boldsymbol{A}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Where $\lambda \in \mathbb{R}$ is known as an eigenvalue.

We call the set of eigenvalues the spectrum of $\boldsymbol{A}$:

$$\text{spec}(\boldsymbol{A}) = \{\lambda \in \mathbb{R} : \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}\}$$

Eigen-(values, vectors)

Let $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ denote a square matrix. An eigenvector is a vector that is only re-scaled under the action of $\boldsymbol{A}$:

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Where $\lambda \in \mathbb{R}$ is known as an eigenvalue.

We call the set of eigenvalues the spectrum of $\boldsymbol{A}$:

$$\operatorname{spec}(\boldsymbol{A}) = \{\lambda \in \mathbb{R} : \boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}\}$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ can have at most d independent eigenvectors.
- An eigenvalue λ can be associated to more than one independent eigenvector.

Positive matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive definite if all eigenvalues are positive:

$$\lambda \in \operatorname{spec}(\boldsymbol{A}) \Rightarrow \lambda > 0$$

Positive matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive definite if all eigenvalues are positive:

$$\lambda \in \operatorname{spec}(\boldsymbol{A}) \Rightarrow \lambda > 0$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive semi-definite if all eigenvalues are non-negative:

$$\lambda \in \operatorname{spec}(\boldsymbol{A}) \Rightarrow \lambda \geq 0$$

Positive matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive definite if all eigenvalues are positive:

$$\lambda \in \text{spec}(\boldsymbol{A}) \Rightarrow \lambda > 0$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive semi-definite if all eigenvalues are non-negative:

$$\lambda \in \text{spec}(\boldsymbol{A}) \Rightarrow \lambda \geq 0$$

Proposition

Symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ are positive semi-definite

Exercise: prove this.

Positive matrices

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive definite if all eigenvalues are positive:

$$\lambda \in \text{spec}(\boldsymbol{A}) \Rightarrow \lambda > 0$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is called positive semi-definite if all eigenvalues are non-negative:

$$\lambda \in \text{spec}(\boldsymbol{A}) \Rightarrow \lambda \geq 0$$

Proposition

Symmetric matrices $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ are positive semi-definite

not necessarily positive definite.

Exercise: prove this.

Spectral theorem

Theorem

Any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$$

$\boldsymbol{U} \in \mathbb{R}^{d \times d}$ are orthogonal matrices and $\boldsymbol{D}$ is a diagonal matrix with elements given by the eigenvalues.

Spectral theorem

Theorem

Any symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$$

$\boldsymbol{U} \in \mathbb{R}^{d \times d}$ are orthogonal matrices and $\boldsymbol{D}$ is a diagonal matrix with elements given by the eigenvalues.

We can equivalently write the spectral decomposition as:

$$\boldsymbol{A} = \sum_{i=1}^{\text{rank}(A)} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$$

Where $\boldsymbol{v}_i \in \mathbb{R}^d$ are orthonormal eigenvectors.

Important facts

- The trace of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is the sum of its eigenvalues

$$\operatorname{Tr} \boldsymbol{A} = \sum_{i=1}^{d} \lambda_i$$

Important facts

- The trace of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is the sum of its eigenvalues

$$\operatorname{Tr} \boldsymbol{A} = \sum_{i=1}^{d} \lambda_i$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is invertible i.f.f. $0 \notin \operatorname{spec}(\boldsymbol{A})$

Important facts

- The trace of a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is the sum of its eigenvalues

$$\operatorname{Tr} \boldsymbol{A} = \sum_{i=1}^{d} \lambda_i$$

- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is invertible i.f.f. $0 \notin \operatorname{spec}(\boldsymbol{A})$

- The eigenvalues of a projection matrix $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ are 0 or 1

$$\boldsymbol{P} = \sum_{i=1}^{\operatorname{rank}(P)} \boldsymbol{v}_i \boldsymbol{v}_i^{\top}$$

Exercise: show this.

Moreover, $\boldsymbol{P} \in \mathbb{R}^{d \times d}$ is orthogonal if $\boldsymbol{v}_i$ are orthogonal vectors.

Singular value decomposition

Note that for any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n \times n}$ are a symmetric matrices.

Singular value decomposition

Note that for any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n \times n}$ are a symmetric matrices.

Therefore, $\boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ can be diagonalised:

$$\boldsymbol{A}^\top \boldsymbol{A} = \sum_{i=1}^{r} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \qquad \boldsymbol{A}\boldsymbol{A}^\top = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

Where: $r = \operatorname{rank}(\boldsymbol{A}^\top \boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}\boldsymbol{A}^\top)$

$\boldsymbol{u}_i \in \mathbb{R}^n, \boldsymbol{v}_i \in \mathbb{R}^d$ are orthonormal vectors.

$\lambda_i \geq 0$

Singular value decomposition

Note that for any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$, $\boldsymbol{A}^\top \boldsymbol{A} \in \mathbb{R}^{d \times d}$ and $\boldsymbol{A}\boldsymbol{A}^\top \in \mathbb{R}^{n \times n}$ are a symmetric matrices.

Therefore, $\boldsymbol{A}^\top \boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ can be diagonalised:

$$\boldsymbol{A}^\top \boldsymbol{A} = \sum_{i=1}^{r} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \qquad \boldsymbol{A}\boldsymbol{A}^\top = \sum_{i=1}^{r} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$$

Where: $r = \operatorname{rank}(\boldsymbol{A}^\top \boldsymbol{A}) = \operatorname{rank}(\boldsymbol{A}\boldsymbol{A}^\top)$

$\boldsymbol{u}_i \in \mathbb{R}^n$, $\boldsymbol{v}_i \in \mathbb{R}^d$ are orthonormal vectors.

$\lambda_i \geq 0$

Therefore, defining the singular values $\sigma_i = \sqrt{\lambda_i}$

Singular value decomposition

Theorem

Any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ can be decomposed as

$$\boldsymbol{A} = \sum_{i=1}^{\mathrm{rank}(\boldsymbol{A})} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

Singular value decomposition

Theorem

Any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ can be decomposed as

$$\boldsymbol{A} = \sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

This can be equivalently written as:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

With: $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ orthogonal matrices

$\boldsymbol{D} \in \mathbb{R}^{n \times d}$ a rectangular matrix with the singular values σ_i

Singular value decomposition

Theorem

Any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ can be decomposed as

$$\boldsymbol{A} = \sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

This can be equivalently written as:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$$

With: $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{d \times d}$ orthogonal matrices

$\boldsymbol{D} \in \mathbb{R}^{n \times d}$ a rectangular matrix with the singular values σ_i

Computationally, it is more efficient to define $\boldsymbol{U} \in \mathbb{R}^{n \times r}$, $\boldsymbol{V} \in \mathbb{R}^{d \times r}$ and $\boldsymbol{D} \in \mathbb{R}^{r \times r}$

Pseudo-inverse

The SVD allow us to define a generalised notion of matrix inverse. Let $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with SVD:

$$\boldsymbol{A} = \sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$$

The pseudo-inverse $\boldsymbol{A}^+ \in \mathbb{R}^{d \times n}$ is defined via its SVD:

$$\boldsymbol{A}^+ = \sum_{i=1}^{\operatorname{rank}(\boldsymbol{A})} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top$$

Pseudo-inverse

The pseudo-inverse $\boldsymbol{A}^+ \in \mathbb{R}^{d \times n}$ is defined via its SVD:

$$\boldsymbol{A}^+ = \sum_{i=1}^{\text{rank}(\boldsymbol{A})} \frac{1}{\sigma_i} \boldsymbol{v}_i \boldsymbol{u}_i^\top$$

It satisfies:

$$\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A} \qquad \boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$$

$$(\boldsymbol{A}^+)^+ = \boldsymbol{A}$$

If $\boldsymbol{A}$ is invertible, $\boldsymbol{A}^+ = \boldsymbol{A}^{-1}$

If $\boldsymbol{A}$ is full-rank, $\boldsymbol{A}^+ = \begin{cases} (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top & \text{if } n \geq d \\ \boldsymbol{A}^\top(\boldsymbol{A}\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^\top & \text{if } n < d \end{cases}$

Exercise: show this.

Pseudo-inverse

The pseudo-inverse is useful to define orthogonal projectors

For any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$:

$$\boldsymbol{A}^{+}\boldsymbol{A} \in \mathbb{R}^{d \times d} \qquad \boldsymbol{A}\boldsymbol{A}^{+} \in \mathbb{R}^{n \times n}$$

Exercise: show this.

Define orthogonal projection operators in the column and row space of $\boldsymbol{A}$, respectively.

Pseudo-inverse

The pseudo-inverse is useful to define orthogonal projectors

For any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$:

$$\boldsymbol{A}^{+}\boldsymbol{A} \in \mathbb{R}^{d \times d} \qquad \boldsymbol{A}\boldsymbol{A}^{+} \in \mathbb{R}^{n \times n}$$

Exercise: show this.

Define orthogonal projection operators in the column and row space of $\boldsymbol{A}$, respectively.

Similarly,

$$\boldsymbol{I}_d - \boldsymbol{A}^{+}\boldsymbol{A} \in \mathbb{R}^{d \times d} \qquad \boldsymbol{I}_n - \boldsymbol{A}\boldsymbol{A}^{+} \in \mathbb{R}^{n \times n}$$

Define orthogonal projection operators in the kernel of $\boldsymbol{A}$ and $\boldsymbol{A}^{\top}$, respectively.

Recap of Probability

The butter of statistical learning

Random variable

A random variable X mathematically formalises the notion of a “measurement” or “random event”.

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Discrete: when the possible outcomes are countable.

Examples:

- the outcome of tossing a coin $X \in \{\text{head}, \text{tail}\}$
- rolling a dice $X \in \{1, \ldots, 6\}$
- The number of people in France $X \in \mathbb{N}$

Random variable

A random variable X mathematically formalises the notion of a “measurement” or “random event”. It can be:

- Discrete: when the possible outcomes are countable.

Examples:
- the outcome of tossing a coin $X \in \{\text{head}, \text{tail}\}$
- rolling a dice $X \in \{1, \ldots, 6\}$
- The number of people in France $X \in \mathbb{N}$

Discrete r.v.s are described by their probability distribution

$$\mathbb{P}(X = k)$$

A positive “function” that sums to one. $\sum_{k \in \text{supp(X)}} \mathbb{P}(X = k) = 1$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Continuous r.v.s are described by their probability density function (p.d.f.), which integrates to probabilities:

$$\mathbb{P}(X \in [a,b]) = \int_a^b \mathrm{d}x \; p_X(x)$$

A "function" that integrates to one: $$\int_{\mathrm{supp}(X)} \mathrm{d}x \; p_X(x) = 1$$

Random variable

A random variable X mathematically formalises the notion of a "measurement" or "random event". It can be:

- Continuous: when the possible outcomes are uncountable.

Examples:
- The temperature in the room $X \in \mathbb{R}$
- The GDP of France next year $X \in \mathbb{R}$

Continuous r.v.s are described by their probability density function (p.d.f.), which integrates to probabilities:

$$\mathbb{P}(X \in [a,b]) = \int_a^b \mathrm{d}x \; p_X(x)$$

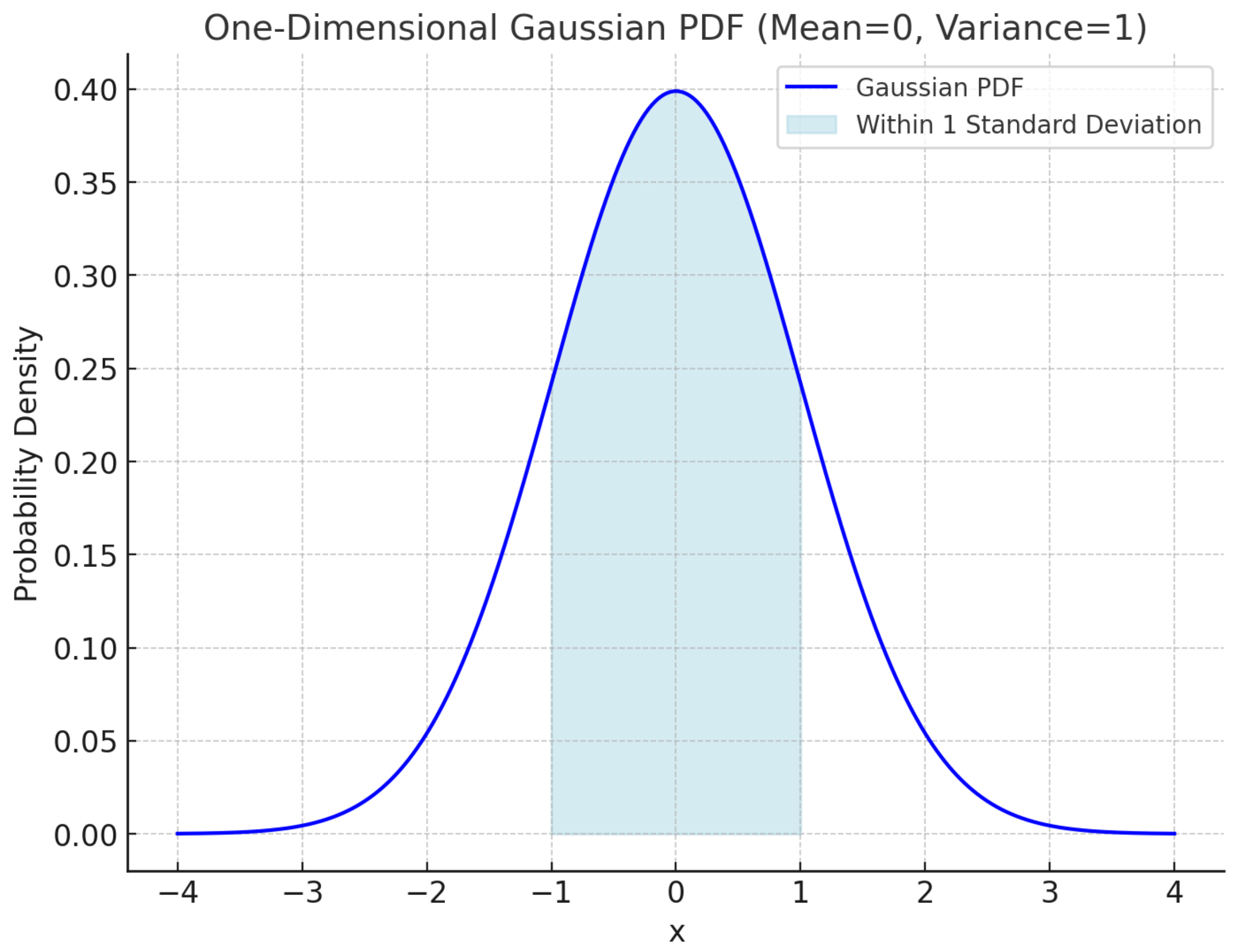
A "function" that integrates to one: $\displaystyle\int_{\mathrm{supp}(X)} \mathrm{d}x \; p_X(x) = 1$

⚠ The p.d.f. is NOT a probability. It can be negative.

Normal distribution

A Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has the following p.d.f.:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



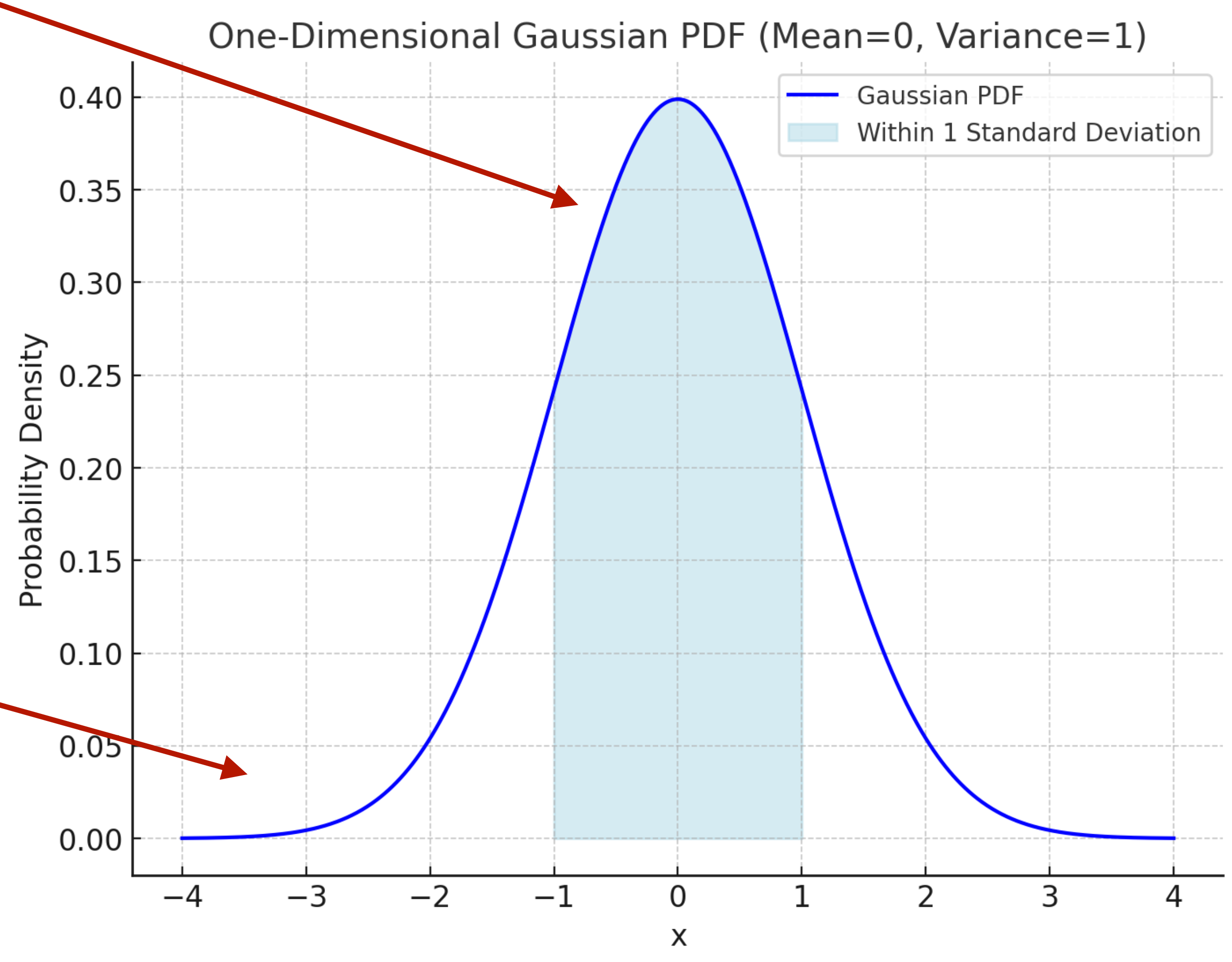
Normal distribution

A Gaussian r.v. $X \sim \mathcal{N}(\mu, \sigma^2)$ has the following p.d.f.:

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

High-probability

Low-probability



Expectation and variance

Let $X \sim p_X$ denote a continuous r.v.

- The expectation (or mean) of X is defined as

$$\mathbb{E}[X] = \int \mathrm{d}x \; p_X(x) x$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[X] = \mu$

Expectation and variance

Let $X \sim p_X$ denote a continuous r.v.

- The expectation (or mean) of X is defined as

$$\mathbb{E}[X] = \int \mathrm{d}x \; p_X(x) x$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}[X] = \mu$

- The variance of X is defined as:

$$\mathrm{Var[X]} = \mathbb{E}[(\mathrm{X} - \mathbb{E}[\mathrm{X}])^2] = \mathbb{E}[\mathrm{X}^2] - \mathbb{E}[\mathrm{X}]^2$$

For example, for $X \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathrm{Var[X]} = \sigma^2$

Change of variables

Let $X \sim p_X$ denote a continuous r.v. and $f : \mathbb{R} \to \mathbb{R}$

Change of variables

Let $X \sim p_X$ denote a continuous r.v. and $f : \mathbb{R} \to \mathbb{R}$

Then, $Y = f(X)$ is also a random variable, with p.d.f. given by

$$p_Y(y) = \int \mathrm{d}x \; p_X(x)\delta(y - f(x))$$

Where $\delta(x)$ is the "Dirac delta function":

$$\int_{\mathbb{R}} \mathrm{d}x \; \delta(x - y)f(x) = f(y)$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x, y)$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x, y)$$

The correlation between X, Y is defined by

$$\mathbb{E}[XY] = \int \mathrm{d}x \int \mathrm{d}y \; p_{X,Y}(x, y) xy$$

Joint distribution

Two random variables X, Y that concern the same random experiment are characterised by their joint p.d.f.

$$p_{X,Y}(x, y)$$

The correlation between X, Y is defined by

$$\mathbb{E}[XY] = \int \mathrm{d}x \int \mathrm{d}y \; p_{X,Y}(x, y) xy$$

We say X, Y are uncorrelated if $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad\qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

- We say the r.v.s. X, Y are independent if

$$p_{X,Y}(x, y) = p_X(x) p_Y(x)$$

Independence

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the marginal distributions

$$p_X(x) = \int \mathrm{d}y \; p_{X,Y}(x, y) \qquad\qquad p_Y(y) = \int \mathrm{d}x \; p_{X,Y}(x, y)$$

- We say the r.v.s. X, Y are independent if

$$p_{X,Y}(x, y) = p_X(x) p_Y(x)$$

Note that independence implies uncorrelated,
but not the converse!

Exercise: Construct a counter-example

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x \mid y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x \,|\, y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Note that $X, Y \sim p_{X,Y}$ are independent if and only if:

$$p_{X|Y}(x \,|\, y) = p_X(x)$$

Conditional distribution

- Given two r.v.s $X, Y \sim p_{X,Y}$, we define the conditional p.d.f.

$$p_{X|Y}(x \,|\, y) = \frac{p_{X,Y}(x, y)}{p_Y(y)}$$

Note that $X, Y \sim p_{X,Y}$ are independent if and only if:

$$p_{X|Y}(x \,|\, y) = p_X(x)$$

Theorem (Bayes theorem)

$$p_{X|Y}(x \,|\, y) = \frac{p_{Y|X}(y \,|\, x) p_X(x)}{p_Y(y)}$$

Law of large numbers

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$

Define the sample mean (note this is itself a r.v.)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Law of large numbers

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$

Define the sample mean (note this is itself a r.v.)

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Theorem (Weak LLN)

$$\bar{X}_n \xrightarrow{P} \mu \qquad \text{as} \qquad n \to \infty$$

$$\lim_{n \to \infty} \mathbb{P}(|\bar{X}_n - \mu| < \epsilon) = 1$$

⚠ Be aware there are many variations of the LLN.

Central limit theorem

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Again, consider the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Central limit theorem

Let $X_1, \ldots, X_n \sim p_X$ denote i.i.d. r.v.s. with mean $\mathbb{E}[X_i] = \mu$ and variance $\mathrm{Var}(X_i) = \sigma^2 < \infty$

Again, consider the sample mean

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Theorem (Lindeberg CLT)

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(\mu, \sigma^2)$$

$$\lim_{n\to\infty} \mathbb{P}(\sqrt{n}(\bar{X}_n - \mu) \leq z) = \mathbb{P}(Z \leq z/\sigma) \qquad Z \sim \mathcal{N}(0,1)$$

⚠ Be aware there are many variations of the CLT.