Statistical Learning II

Lecture 1 - Introduction & preliminaries

Bruno Loureiro
@ CSD, DI-ENS & CNRS

brloureiro@gmail.com

DL3 IASO, Université Paris Dauphine-PSL
11.09.2024

Course Organisation

- 12 classes of 3h, divided in:

  - 1h30 lectures

  - 1h30 exercises & lab (Python) (with Leonardo De Filippis)

Course Organisation

- 12 classes of 3h, divided in:

  - 1h30 lectures

  - 1h30 exercises & lab (Python)
  (with Leonardo De Filippis)

- Contact us: via Microsoft Teams or e-mail:

  bruno.loureiro@di.ens.fr and leonardo.defilippis99@gmail.com

Course Organisation

- 12 classes of 3h, divided in:

  - 1h30 lectures

  - 1h30 exercises & lab (Python) (with Leonardo De Filippis)

- Contact us: via Microsoft Teams or e-mail:

  bruno.loureiro@di.ens.fr and leonardo.defilippis99@gmail.com

- Evaluation: 2 x exams (midterm and final)

Menu for the semester

Goal: Develop a *mathematical* understanding of *classical* and *modern* machine learning models

Menu for the semester

Goal: Develop a *mathematical* understanding of *classical* and *modern* machine learning models

- Classical methods:
    - Ridge regression
    - LASSO
    - Generalised linear models
    - Kernel methods
    - Principal component analysis (PCA)

Menu for the semester

Goal: Develop a *mathematical* understanding of *classical* and *modern* machine learning models

- Classical methods:
    - Ridge regression
    - LASSO
    - Generalised linear models
    - Kernel methods
    - Principal component analysis (PCA)

- Modern methods:
    - Neural networks
    - Diffusion models
    - Your suggestions?

Introduction & Motivation

Or: *why should I care about Statistical Learning?*

A brief history of ML

1940
1st digital computer

W. McCulloch W. Pitts

McCulloch-Pitts neuron
1943

1947
Invention of the transistor

J. Bardeen W.H. Brattain W. Shockley

F. Rosenblat

Rosenblat's perceptron / MLP
1957

1969
ARPANET (Internet)

S. Linnainmaa

Backprop
1970

K. Fukushima

Neocognitron
1980

1982
Hopfield Model

J. Hopfield

Y. LeCun

CNNs
1989

1997
IBM Deep Blue

2016
AlphaGo

2017
Transformers

2020
AlphaFold2

DALL-E
2021

ChatGPT
2022

What about the maths?

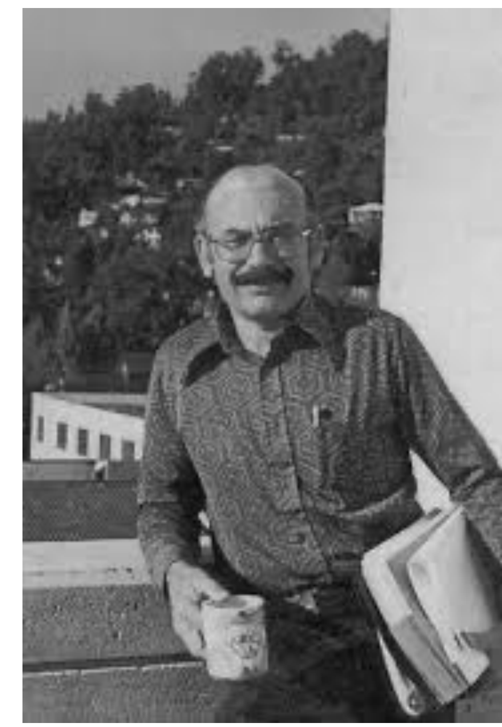
Yet, on the mathematical side...

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?



Leo Breiman
1928

What about the maths?

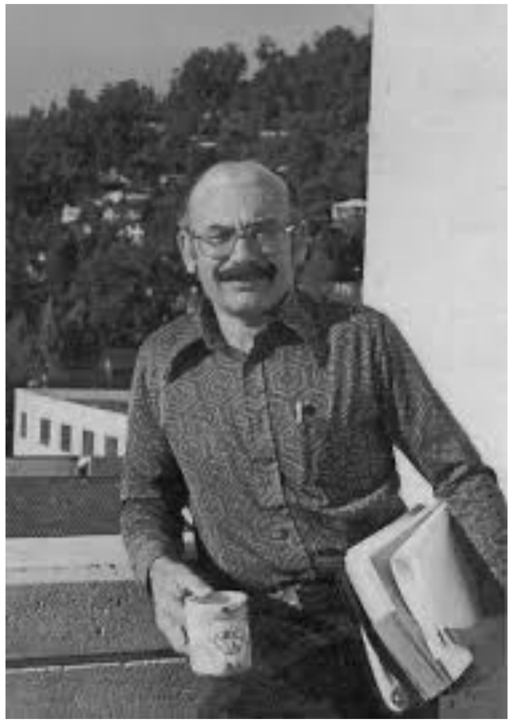
Yet, on the mathematical side…

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?



Leo Breiman
1928

This was written in 1995!!!

What about the maths?

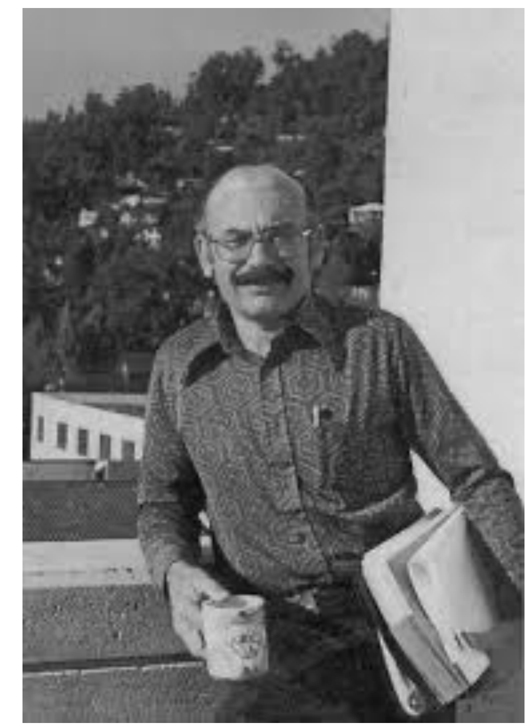
Yet, on the mathematical side…

Leo Breiman
Statistics Department, University of California, Berkeley, CA 94305;
e-mail: leo@stat.berkeley.edu

Reflections After Refereeing Papers for NIPS

For instance, there are many important questions regarding neural networks which are largely unanswered. There seem to be conflicting stories regarding the following issues:

- Why don't heavily parameterized neural networks overfit the data?
- What is the effective number of parameters?
- Why doesn't backpropagation head for a poor local minima?
- When should one stop the backpropagation and use the current parameters?



Leo Breiman
1928

This was written in 1995!!!

But why should I care?

Some good reasons to care

- Reliability and Liability

  If a model does something unexpected, who is responsible?

  Crucial in sensitive applications, e.g. medicine, law, self-driving cars/planes…

Some good reasons to care

- Reliability and Liability

  If a model does something unexpected, who is responsible?

  Crucial in sensitive applications, e.g. medicine, law, self-driving cars/planes...

- Efficient design

  Data centres are responsible for 4% of the energy consumption in the US.

  Can we design models and algorithms that learn more efficiently from data?

Some good reasons to care

- Reliability and Liability

  If a model does something unexpected, who is responsible?

  Crucial in sensitive applications, e.g. medicine, law, self-driving cars/planes...

- Efficient design

  Data centres are responsible for 4% of the energy consumption in the US.

  Can we design models and algorithms that learn more efficiently from data?

- Scientific curiosity

Some good reasons to care

- Reliability and Liability

  If a model does something unexpected, who is responsible?

  Crucial in sensitive applications, e.g. medicine, law, self-driving cars/planes...

- Efficient design

  Data centres are responsible for 4% of the energy consumption in the US.

  Can we design models and algorithms that learn more efficiently from data?

- Scientific curiosity

- And in the worst case, understanding the maths will make you a better engineer / data scientist.

Our expectations

My expectations: By the end of the term, I expect you to:

- Get acquainted with the most popular machine learning algorithms

Our expectations

My expectations: By the end of the term, I expect you to:

- Get acquainted with the most popular machine learning algorithms

- Appreciate (some) of the mathematics behind the methods.

Our expectations

My expectations: By the end of the term, I expect you to:

- Get acquainted with the most popular machine learning algorithms
- Appreciate (some) of the mathematics behind the methods.
- Be able to implement these methods from scratch.

99% of books in statistical learning:

Let $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$,
denote independently drawn samples from a probability distribution….

99% of books in statistical learning:

Let $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, $i = 1, \ldots, n$,
denote independently drawn samples from a
probability distribution....

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

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

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"

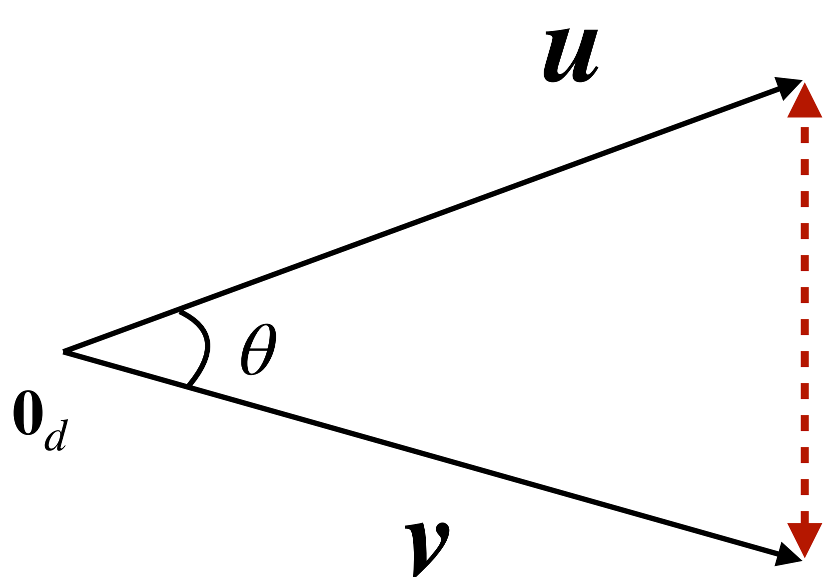
We say two vectors $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$ are orthogonal if $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$

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

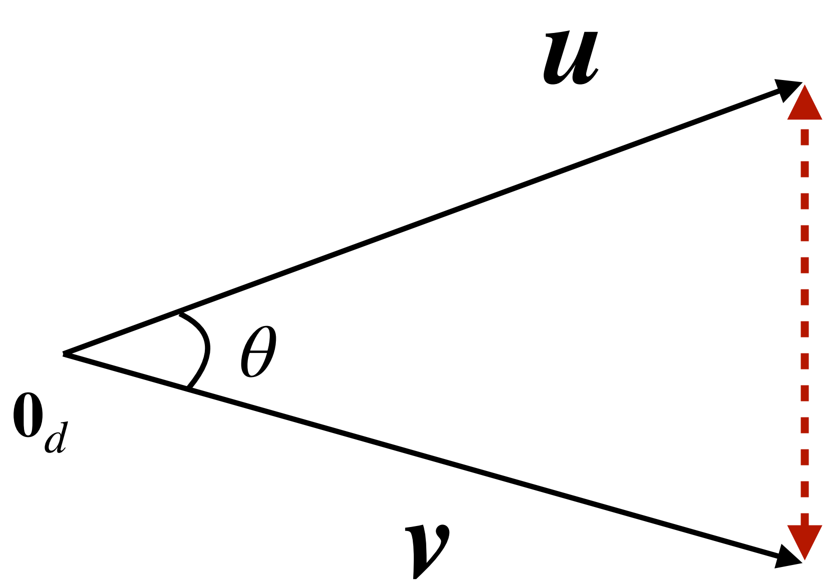


$$d(\boldsymbol{u}, \boldsymbol{v}) = ||\boldsymbol{u} - \boldsymbol{v}||_2^2$$

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

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

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

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 $p \neq 2$

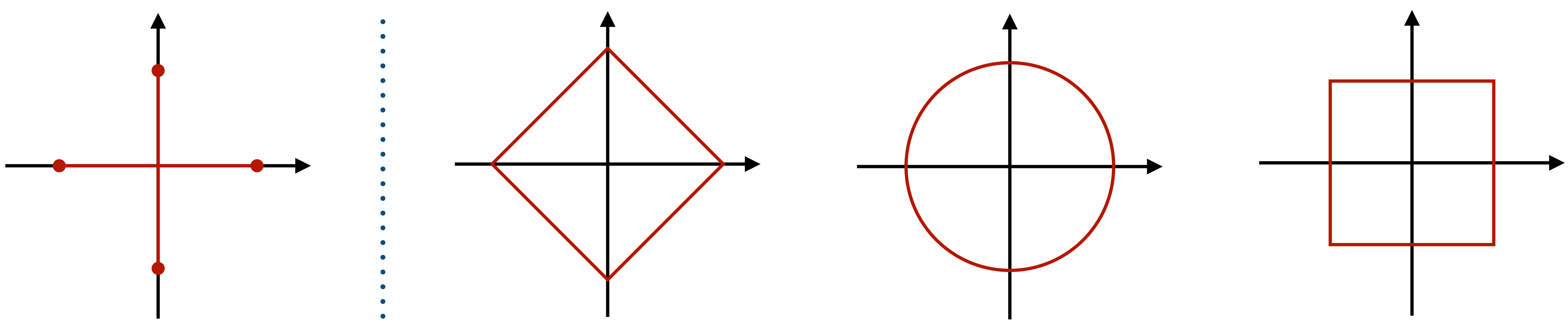
Other norms

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 $p \neq 2$



ℓ_0 ℓ_1 ℓ_2 ℓ_∞

Not a norm

Matrices

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

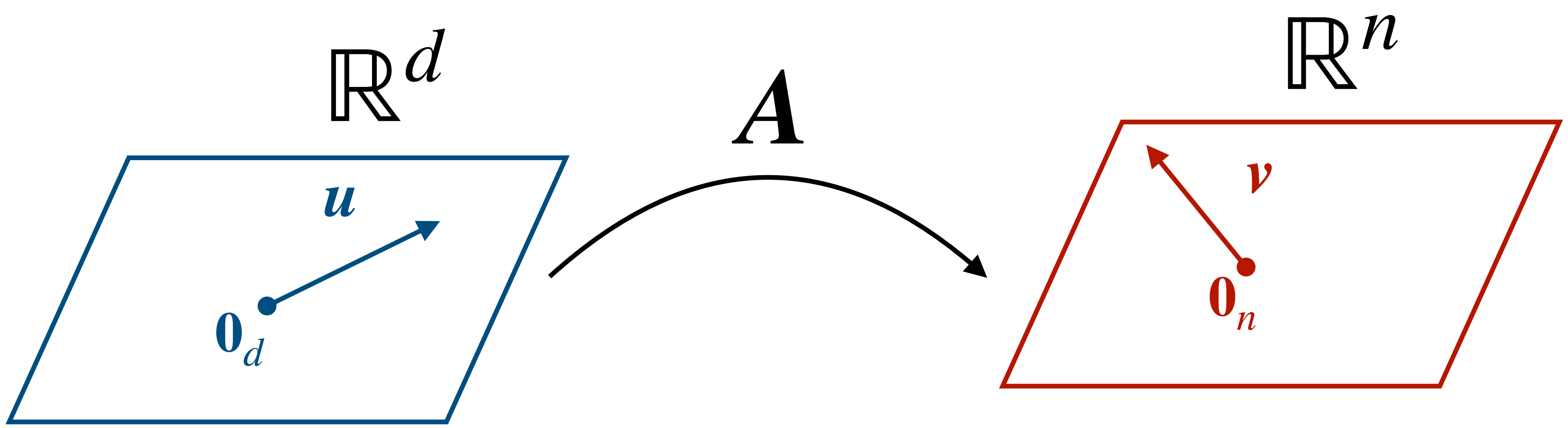
$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nd} \end{bmatrix} \in \mathbb{R}^{n \times d}$$

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.

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

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

Matrices

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

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

- 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

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

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

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

This is equivalent to the number of independent columns.

Proposition

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

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

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

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 $\mathbf{col}(\boldsymbol{A}) \subset \mathbb{R}^n$ as
The image of the associated linear map.

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

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

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

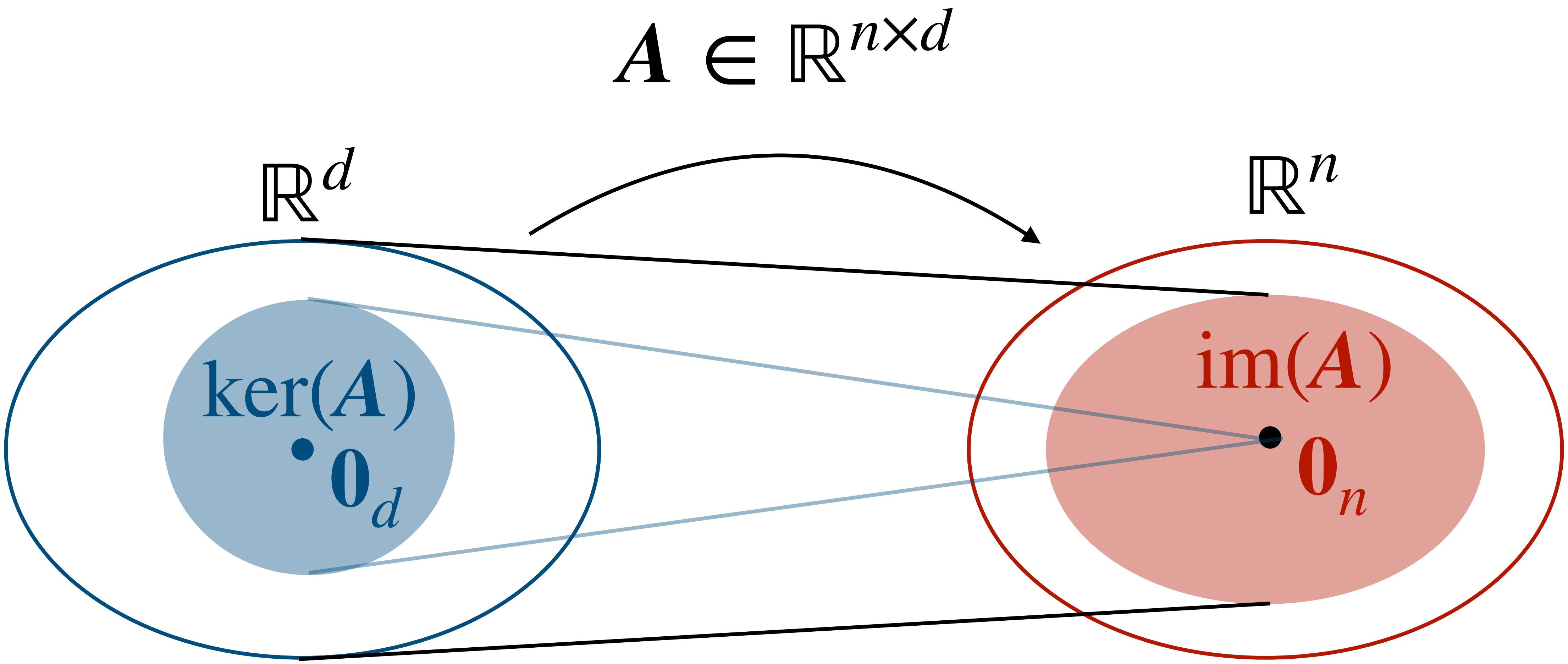
Note that $\ker(\boldsymbol{A}) \subset \mathbb{R}^d$
and $\mathbf{0} \in \ker(\boldsymbol{A})$

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

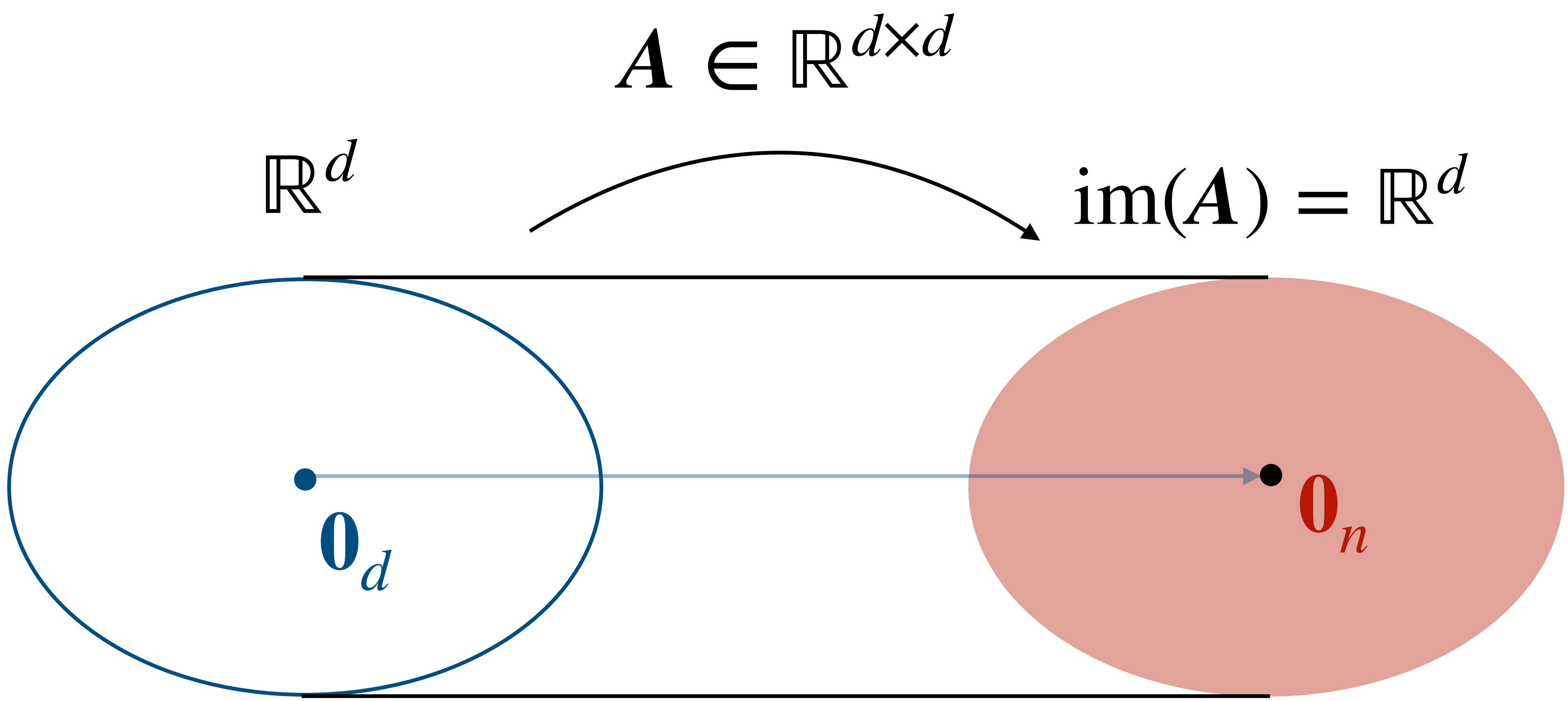
$$\operatorname{rank}(\boldsymbol{A}) = d$$

Matrix inverse

Proposition

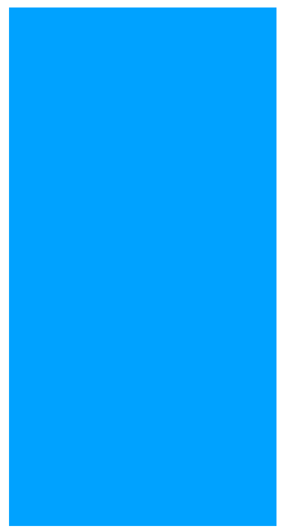
A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is invertible if and only if it is full-rank

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



Matrix transpose

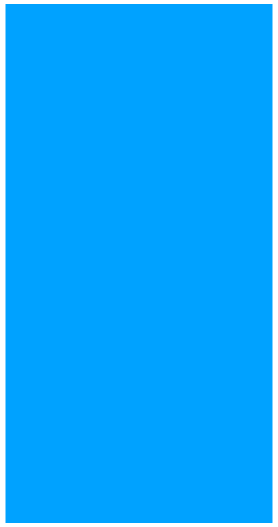
- The transpose of a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times d}$ with elements a_{ij} the matrix with $\boldsymbol{A}^\top \in \mathbb{R}^{d \times n}$ with elements a_{ji}

$\boldsymbol{A} =$ 

$\boldsymbol{A}^\top =$ 

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:

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

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

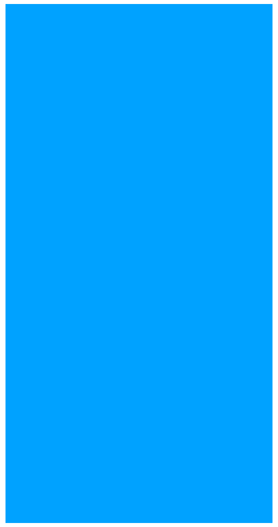
$$(\boldsymbol{A}^{-1})^\top = (\boldsymbol{A}^\top)^{-1}$$

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

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}

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

- 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:

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

- It satisfies:

$$\mathrm{Tr}\ \boldsymbol{A}\boldsymbol{B} = \mathrm{Tr}\ \boldsymbol{B}\boldsymbol{A}$$

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

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

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

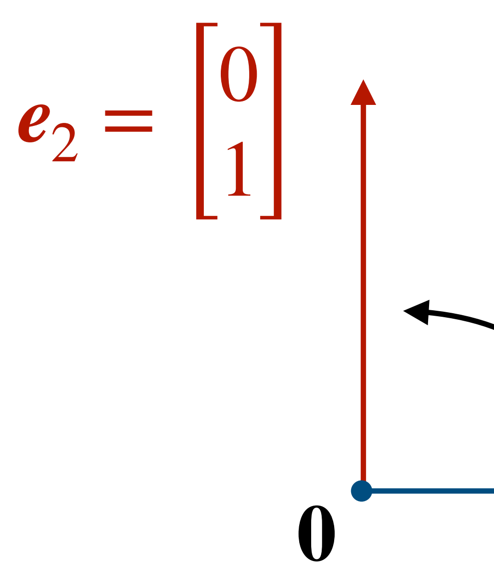
- A square matrix $\boldsymbol{A} \in \mathbb{R}^{d \times d}$ is orthogonal if $\boldsymbol{A}^\top = \boldsymbol{A}^{-1}$

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

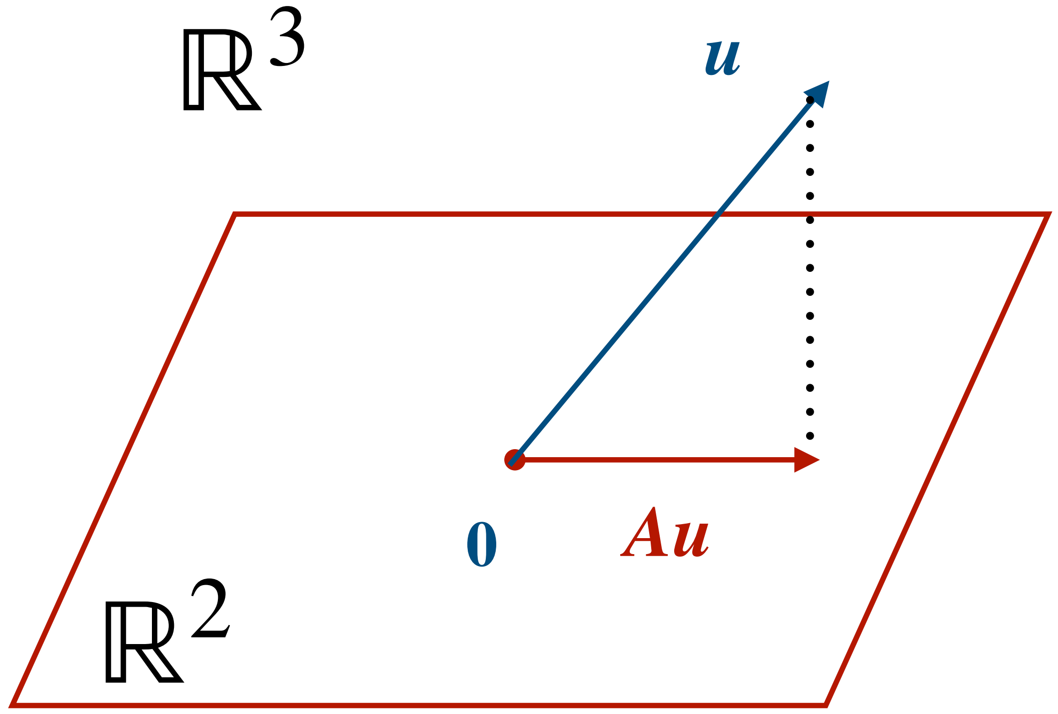


$$\boldsymbol{A} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

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.



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

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 $\boldsymbol{A}$:

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

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:

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

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

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

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:

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

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

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

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:

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

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

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

Important facts

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

$$\boxed{\mathrm{Tr}\,\boldsymbol{A} = \sum_{i=1}^{d} \lambda_i}$$

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

Important facts

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

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

- 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

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

🧐 <u>Exercise</u>: 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:

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

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

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

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

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

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:

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

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

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

Pseudo-inverse

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

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

It satisfies: $\boldsymbol{A}\boldsymbol{A}^+\boldsymbol{A} = \boldsymbol{A}$ $\quad$ $\boldsymbol{A}^+\boldsymbol{A}\boldsymbol{A}^+ = \boldsymbol{A}^+$

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

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

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

Exercise: show this.

Pseudo-inverse

The pseudo-inverse is useful to define orthogonal projectors

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

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

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.