Robotics Research Center Summer School 2023

IIIT Hyderabad

May 16, 2023

Bayes, Eigen and Beyond!

Naren Akash
Center for Visual Information Technology

Robotics Research Center
Summer School 2023

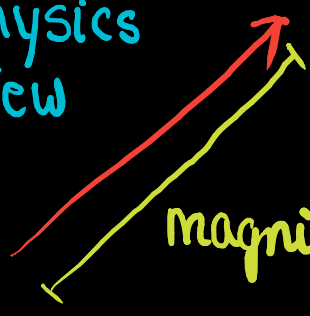
PART ONE
LINEAR ALGEBRA

VECTORS

Computer Science View

$a = [1, 2, 3]$ Ordered list, 3 dim

Physics View

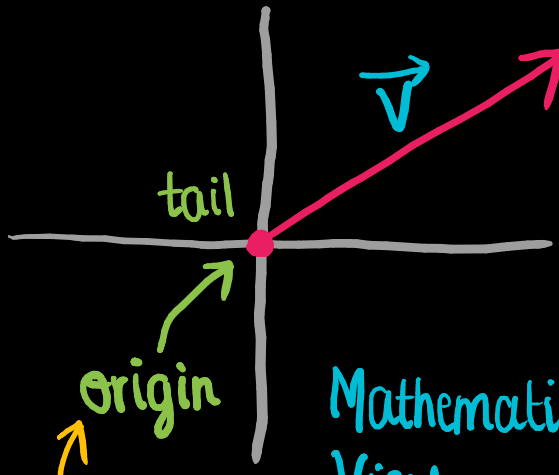


direction

2D or 3D

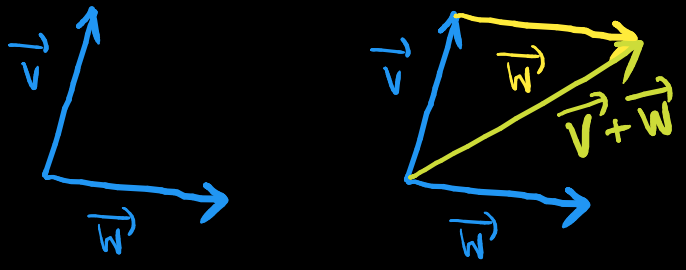
magnitude

Mathematics View



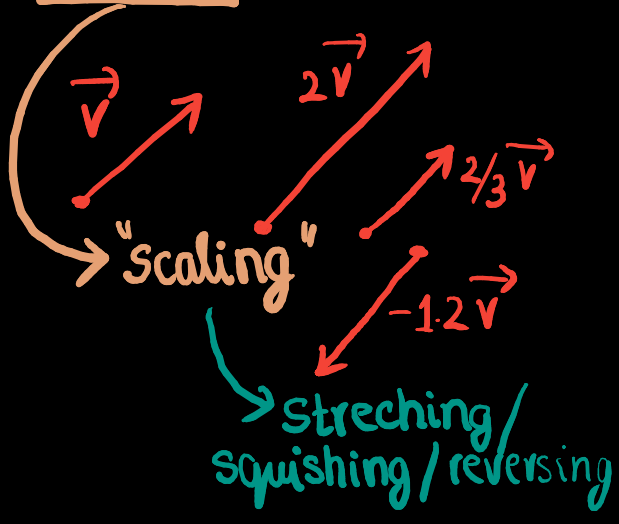
root of all vectors

(i) Vector Addition



each vector
certain movement
↳ dist + dir in space

(ii) Scalar Multiplication



Linear Algebra tend to revolve around these two fundamental topics.

(i) Vector Addition

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ y_1 + y_2 \end{bmatrix}$$

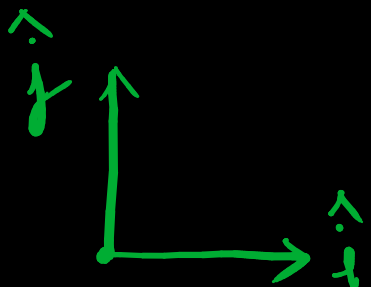
(ii) Scalar Multiplication

$$c \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} cx \\ cy \end{bmatrix}$$

Vector Coordinates

[ ]

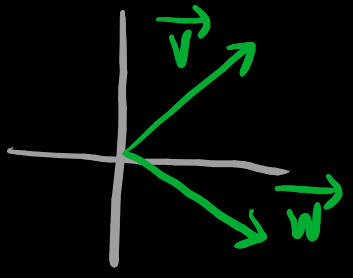
each coordinate (scalar) streches/squishes vectors



Basis of co-ordinate system

Vectors that these coordinates describe is the sum of two scaled vectors

What if we choose diff. basis vectors?



Linear Combination of $\vec{v}, \vec{w}$

$$a\vec{v} + b\vec{w}$$

scalars

Any time we describe vectors numerically, it depends on an implicit choice of basis vectors we are using.

Span of $\vec{v}$ and $\vec{w}$

the set of all the linear combinations

$$a\vec{v} + b\vec{w}$$

a, b vary over all real numbers

Span of <u>most</u> 2D vectors is <u>all</u> vectors of 2D space.

$\vec{v}$ and $\vec{w}$ are ...

Linearly dependent

One vector can be expressed as linear combination of the other

↓

It is in the span of the other

Linearly independent

$$\vec{w} \neq a\vec{v} \quad \forall a$$

$$\vec{u} \neq a\vec{v} + b\vec{w} \quad \forall a,b$$

Linear Transformation → function

vector → vector

kind of like 'movement'

(2D) Transformation is linear

- Lines remain lines
- Origin remains fixed

⇒ Grid lines are ||^{el} and evenly spaced.

Numerical description of linear transformation...

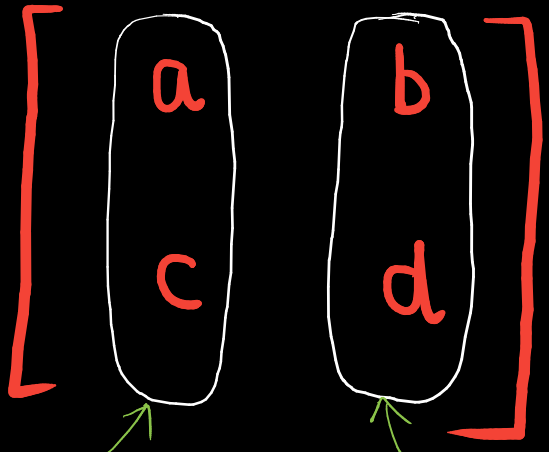
- We need to record only where the basis vectors, $\hat{i}$ and $\hat{j}$ land.

$$\vec{v} = a\hat{i} + b\hat{j}$$

$$\text{Transformed } \vec{v} = a\left(\text{Transformed } \hat{i}\right) + b\left(\text{Transformed } \hat{j}\right)$$

2D linear transformation

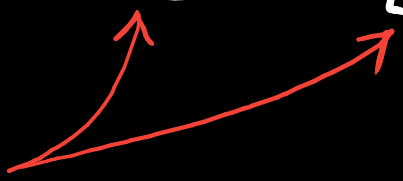
2 x 2 matrix



where $\hat{i}$ lands

where $\hat{j}$ lands

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = x \begin{bmatrix} a \\ c \end{bmatrix} + y \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix}$$

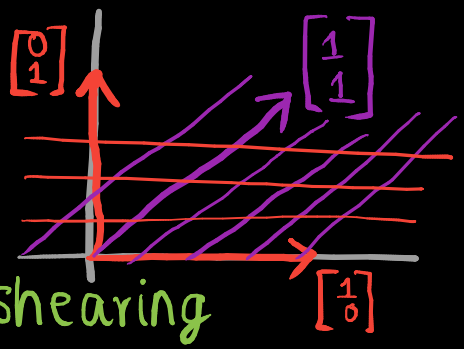


transformed version of the basis vectors.

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

90° rotation counterclockwise

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 2 & -2 \\ 1 & -1 \end{bmatrix}$$

linearly dependent columns.

Linear transformations are a way to move around space s.t.

the grid lines remain ll^{el} and evenly spaced.

MATRICES

- a language to describe linear transformations

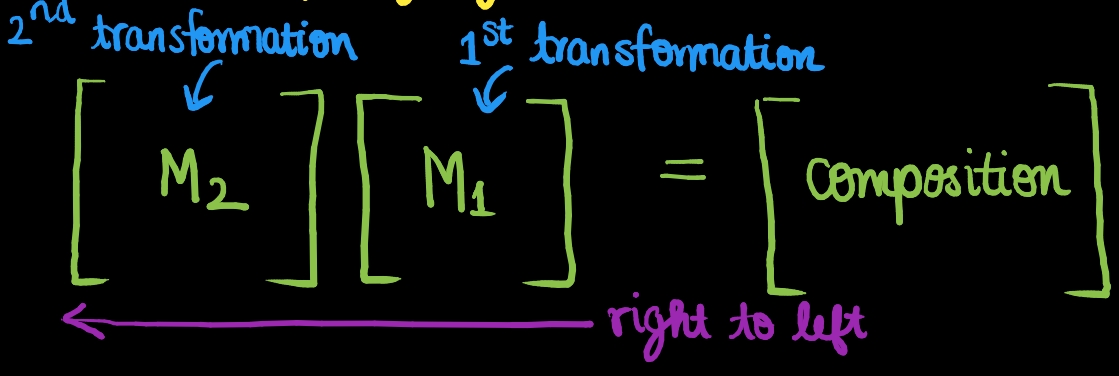
Any matrix can be interpreted as a certain transformation of the space.

- cols. represent those coordinates
- matrix-vector multiplication is a way to compute what that transformation does to a given vector.

Matrix multiplication as COMPOSITION

→ Geometric meaning

- applying one transformation after another



M_2 M_1

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae + bg & af + bh \\ ce + dg & cf + dh \end{bmatrix}$$

Where does $\hat{i}$ go?

Where does $\hat{j}$ go?

Where does $\hat{i}$ go?

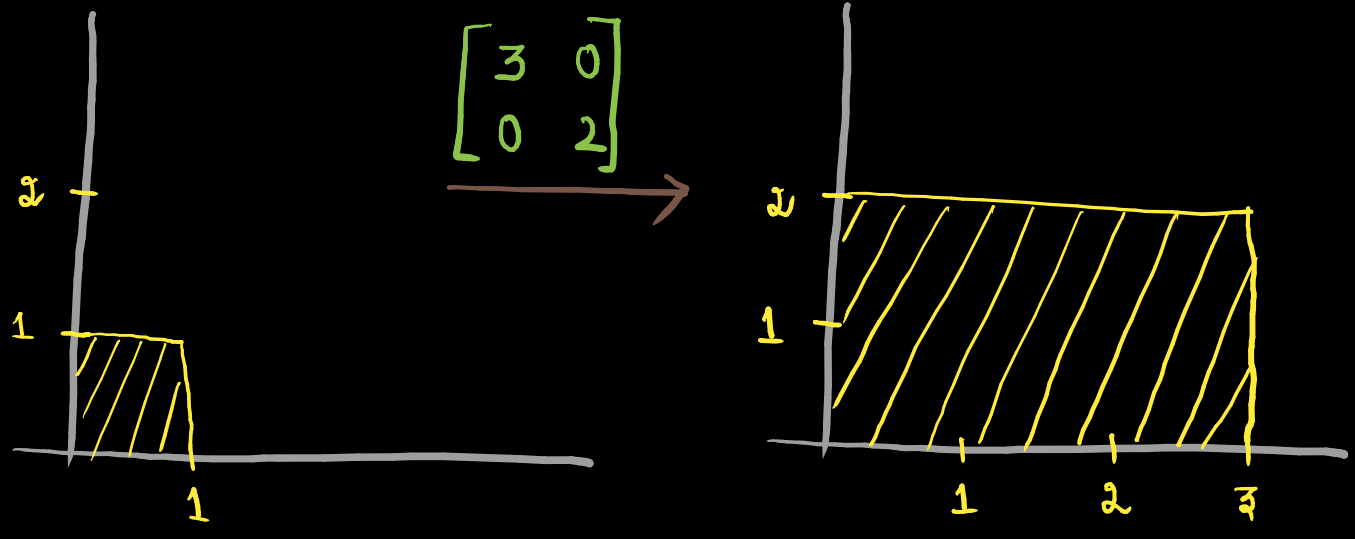
$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} e \\ g \end{bmatrix} = e \begin{bmatrix} a \\ c \end{bmatrix} + g \begin{bmatrix} b \\ d \end{bmatrix} = \begin{bmatrix} ae + bg \\ ce + dg \end{bmatrix}$$

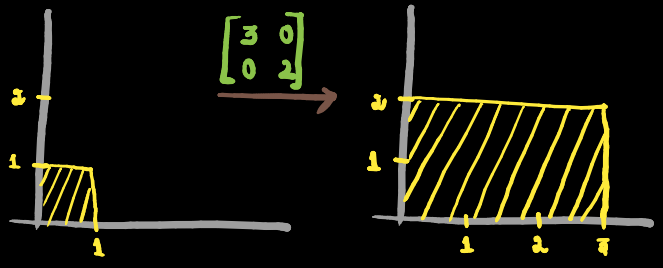
$\hat{i}$ lands here after the 1st transformation

$M_1 M_2 \stackrel{?}{=} M_2 M_1$

By thinking in terms of transformations, we can say $M_1 M_2 \neq M_2 M_1$ w/o any computation.

DETERMINANT





Area = 1×1
$= 1$

this sq. gives us how any possible region in space changes.

Area = 3×2
$= 6$

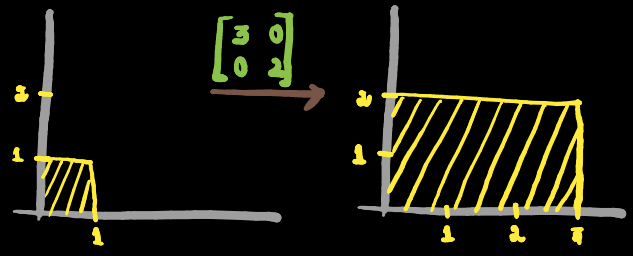
How much are areas scaled?

the factor by which the gn. region increases or decreases.

This scaling factor

the factor by which a linear transformation changes any area

is called DETERMINANT of the transformation.



Area = 1 x 1
= 1

Area = 3 x 2
= 6

$$\det\left(\begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}\right) = 6$$

$$\det\left(\begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}\right) = 0$$

squishes all of the 2D space onto a line or onto a single pt.

checking if det. of a gn. matrix is 0 will tell us if computing transformation with the matrix squishes everything into a smaller dimension.

$$\det\left(\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}\right) = -2$$

How do you scale area by a (−)ive number?

This has to do with the idea of orientation.

↳ flipping the orientation of the space.

↑ $\hat{j}$ to the left or right of $\hat{i}$?

In 3D, det· tells us how much volumes get scale

1 x 1 x 1 cube w/ edges resting on $\hat{i}$, $\hat{j}$ and $\hat{k}$

parallelepiped

$\det\left(\begin{bmatrix} & \\ & \end{bmatrix}\right)$ = vol· of parallelepiped

3x3

$$\det\left(\underset{3\times3}{\begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}}\right) = 0$$

All of the space squished into something w/ zero volume.

either a flat plane, a line or in extreme case - onto a single point.

Usefulness of matrices

Example: manipulation of spaces

→ computer graphics, robotics, etc.

More generally, lets us solve certain systems of equations.

linear system of eqns.

$A\vec{x} = \vec{v}$

unknown

linear transformation

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

Identity transformation

A^{-1} Inverse Transformation

transformation in reverse

(If $\det(A) \neq 0$)

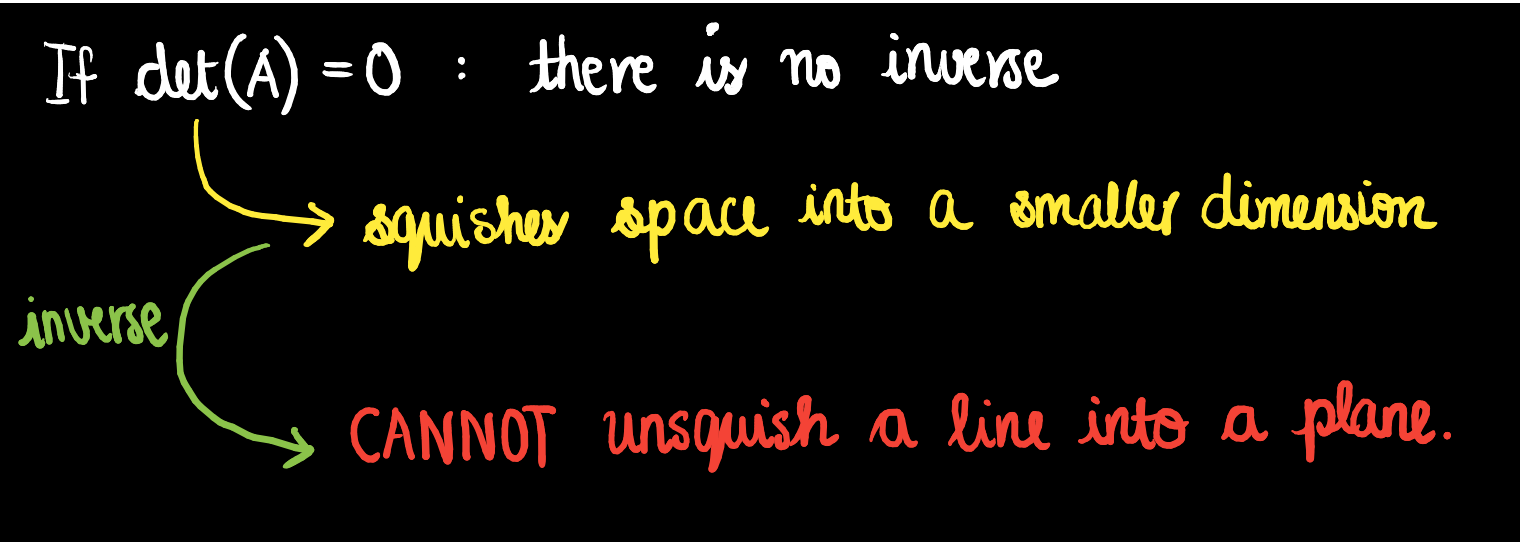
then A^{-1} exists.

$A^{-1}A\vec{x} = A^{-1}\vec{v}$

"do-nothing"

$\vec{x} = A^{-1}\vec{v}$

probably unique soln. exists.


If det(A) = 0 : there is no inverse
squishes space into a smaller dimension
inverse
CANNOT unsquish a line into a plane.

Solution can still exist when $\det(A) = 0$.

$A\vec{x} = \vec{v}$

Ex: Say, a transformation squishes space onto a line, if $\vec{v}$ lives somewhere on that line, then solution exists.

Gin. 3×3 matrix, it is lot harder for a solution to exist when it squishes space onto a line compared to a plane

- even though both of them are zero det.

RANK

Number of dimensions in the output of the transformation.

Example: When o/p of a linear transformation is 1D, we say the transformation of rank 1.

Set of all possible outputs $A\vec{v}$

"column space" of A

Cols. of matrix tells us where the basis vector land.

the span of all these transformed basis vectors gives us all possible outputs.

Rank: number of dimensions in column space

If rank = number of columns,
"full rank" matrix

$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ is always in the column space.

linear transformation must keeps origin fixed in the plane.

For full rank transformation,

only $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ lands on $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

The set of vectors that lands on the origin is called NULL SPACE ← aka. KERNEL of the matrix.

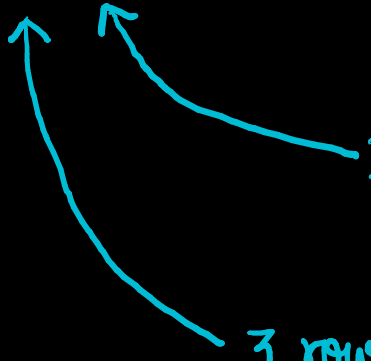
In terms of linear system of equations,

when $\vec{v}$ is zero vector,

null space gives us all possible solution to equation.

Non-square matrices as transformations b/w dimensions

3 x 2 matrix : geometric interpretation of mapping 2D to 3D.



2 columns indicate that the i/p space has 2 basis vectors

3 rows indicate that the landing spot for each of these basis vectors is described w/ 3 separate coordinates.

3×2 matrix

Column space is a 2D space slicing through the origin of the 3D space.

matrix is still full rank

#dim in col. space same as #dim in i/p space

2×3 matrix

3 basis vectors (3D) ⟶ 2 coordinates for each landing spots (2D)

Cramer's Rule

Solving linear system of equations $A\vec{x} = \vec{v}$

The type of answer depends on:

whether or not the transformation squishes all of the space into a lower dim. (i.e., it has zero det.)

If $\det(A) \neq 0$

Output of this transformation still spans the full n-dim space it started in.

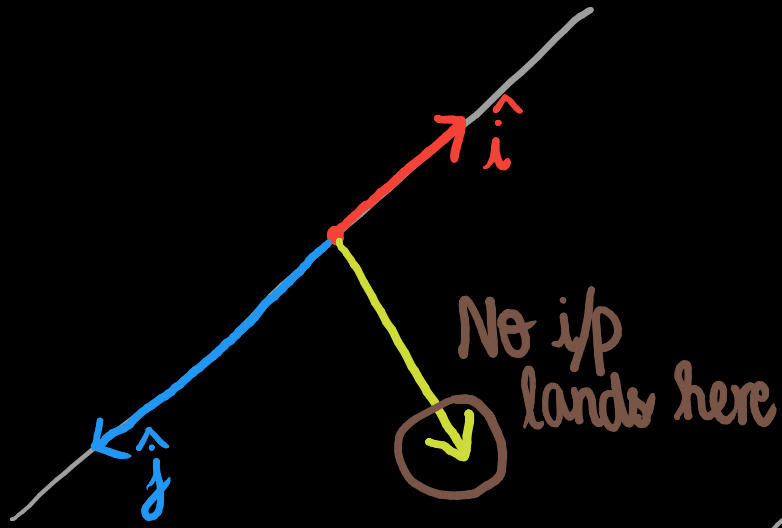
Every i/p lands on one and only one o/p.

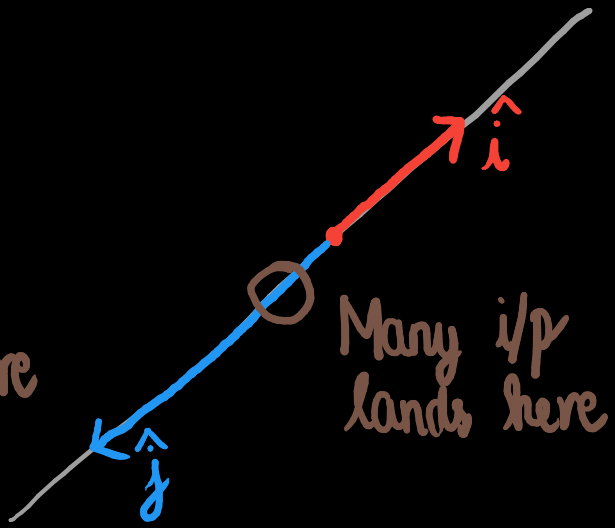
Every o/p has only one and only i/p.

$\longrightarrow$ unique solution

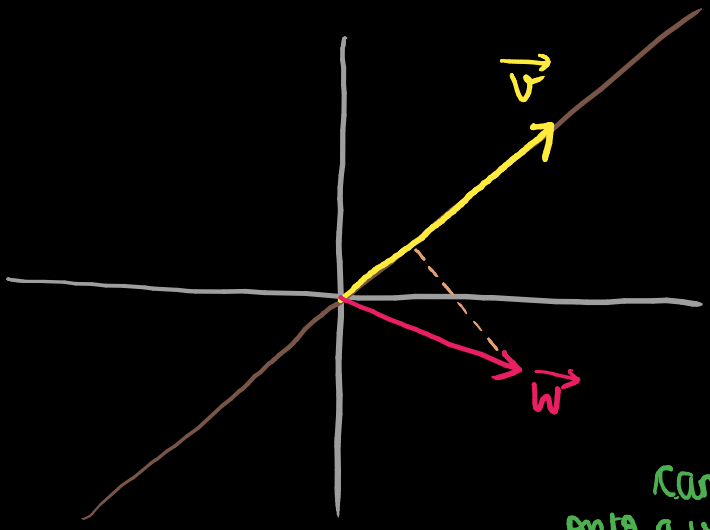
If $\det(A) = 0$

2D space $\longrightarrow$ line





Dot Product



Two vectors of same dimension

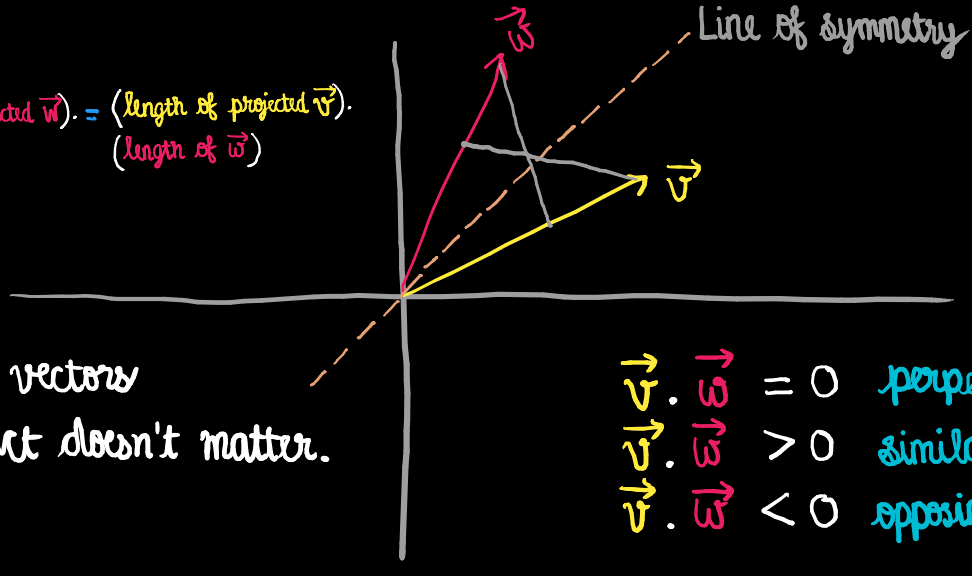
$$\begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix} \cdot \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = a_1 b_1 + \cdots + a_n b_n$$

$$\vec{v} \cdot \vec{w} = (\text{length of projected } \vec{w}) \cdot (\text{length of } \vec{v})$$

Can be negative if the projection of a vector onto a vector is in the opposite direction.

Dot product is commutative.

$\vec{v} \cdot \vec{w}$ = (length of projected $\vec{w}$) · (length of $\vec{v}$) = (length of projected $\vec{v}$) · (length of $\vec{w}$)



The order of vectors in dot product doesn't matter.

$\vec{v} \cdot \vec{w} = 0$ perpendicular

$\vec{v} \cdot \vec{w} > 0$ similar directions

$\vec{v} \cdot \vec{w} < 0$ opposing directions

What does the dot product mean?

$$\vec{a} \cdot \vec{b} = \|\vec{a}\| \, \|\vec{b}\| \cos\theta$$

cosine similarity

Dot product encodes information about the angle between the two vectors.

- measures how similar two vectors are, or, how well they travel together.

Duality

some space

number line

Any time we have a nD-to-1D linear transformation, it is associated with a unique vector in that space.

Performing the linear transformation is the same as taking the dot product with that vector.

Example:

Where $\hat{i}$ lands

Where $\hat{j}$ lands

Transform $\vec{v}$

$$\begin{bmatrix} 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix}$$

$\vec{v}$

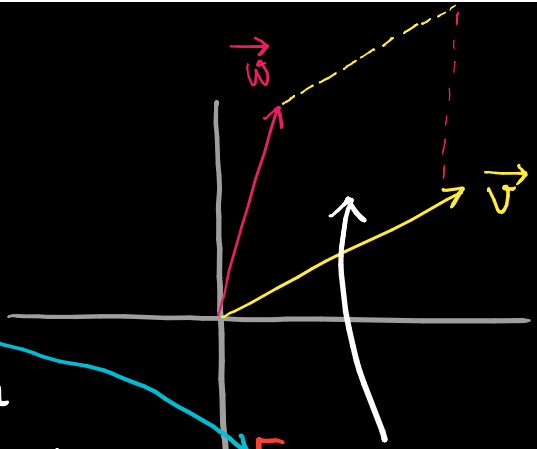
Dual vector of the transform

Dot Product

Cross Product

a vector

$\vec{v} \times \vec{w} = \vec{p}$



with length 2·5

perpendicular to the $\parallel^{gm}$ ($\vec{v}$ and $\vec{w}$)

area of parallelogram = 2·5

direction given by Right Hand Thumb rule

recall determinant

can be positive or negative

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \det\left(\begin{bmatrix} \hat{i} & v_1 & w_1 \\ \hat{j} & v_2 & w_2 \\ \hat{k} & v_3 & w_3 \end{bmatrix}\right)$$

BUT WHY?

we get some linear combination of the basis vectors.

$$\hat{i}\underbrace{(v_2 w_3 - v_3 w_2)}_{\text{some number}} + \hat{j}\underbrace{(v_3 w_1 - v_1 w_3)}_{\text{some number}} + \hat{k}\underbrace{(v_1 w_2 - v_2 w_1)}_{\text{some number}}$$

Geometry of Cross Product

$$\vec{v} \times \vec{w} = \vec{p}$$

binary operation on 2 vectors in n-dimensional space

Example: 3D cross product

$$\text{Let } f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

volume of a parallelepiped

→ a function from 3D to number line

variable: $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$

fixed: $\vec{v}$, $\vec{w}$

$$f\left(\begin{bmatrix} x \\ y \\ z \end{bmatrix}\right) = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

variable

$\vec{v}$ $\vec{w}$

This function is linear. → properties of determinant

⇒ There is some way to describe the function as matrix multiplication.

Here, from 3-dim ⟶ 1-dim.

$$\begin{bmatrix} ? & ? & ? \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

$\vec{v}$ $\vec{w}$

1 x 3 matrix encoding the 3D-to-1D linear transformation.

$$\begin{bmatrix} ? \\ ? \\ ? \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

$\vec{v}$ $\vec{w}$

From the idea of DUALITY.

Interpret the transformation as a dot product with a certain vector.

$\vec{p}$

$$p_1 \cdot x + p_2 \cdot y + p_3 \cdot z = x(v_2 \cdot w_3 - v_3 \cdot w_2) + y(v_3 \cdot w_1 - v_1 \cdot w_3) + z(v_1 \cdot w_2 - v_2 \cdot w_1)$$

$$\begin{bmatrix} ? \\ ? \\ ? \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

$\vec{p}$ $\qquad$ $\vec{v}$ $\quad$ $\vec{w}$

$$p_1 = v_2 \cdot w_3 - v_3 \cdot w_2$$
$$p_2 = v_3 \cdot w_1 - v_1 \cdot w_3$$
$$p_3 = v_1 \cdot w_2 - v_2 \cdot w_1$$

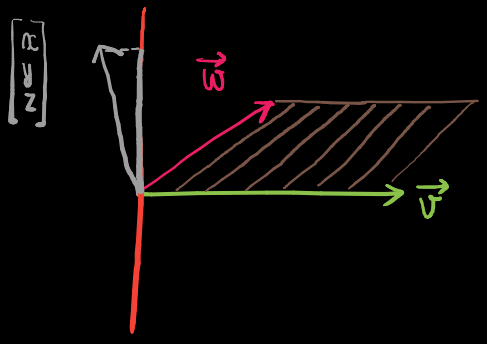
These particular combinations of the coordinates of $\vec{v}$ and $\vec{w}$ are going to be the coordinates of $\vec{p}$.

$$p_1 \cdot x + p_2 \cdot y + p_3 \cdot z = x\left(v_2 \cdot w_3 - v_3 \cdot w_2\right) + y\left(v_3 \cdot w_1 - v_1 \cdot w_3\right) + z\left(v_1 \cdot w_2 - v_2 \cdot w_1\right)$$

Collecting the constant terms that are multiplied by x, y and z is no different from plugging in the symbols $\hat{i}$, $\hat{j}$ and $\hat{k}$ to that first column.

→ signals to interpret these coefficients as the coordinates of a vector.

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} \times \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \det\left(\begin{bmatrix} \hat{i} & v_1 & w_1 \\ \hat{j} & v_2 & w_2 \\ \hat{k} & v_3 & w_3 \end{bmatrix}\right)$$



Volume of parallelepiped

$$= \left(\text{Area of parallelogram}\right) \times \left(\text{Component of } \begin{bmatrix} x \\ y \\ z \end{bmatrix} \text{ perpendicular to } \vec{v} \text{ and } \vec{w}\right)$$

Same as taking dot product of $\begin{bmatrix} x \\ y \\ z \end{bmatrix}$ and a vector $\perp r$ to $\vec{v}$ and $\vec{w}$ with a length equal to the area of the parallelogram.

→ its direction takes care of signed volume

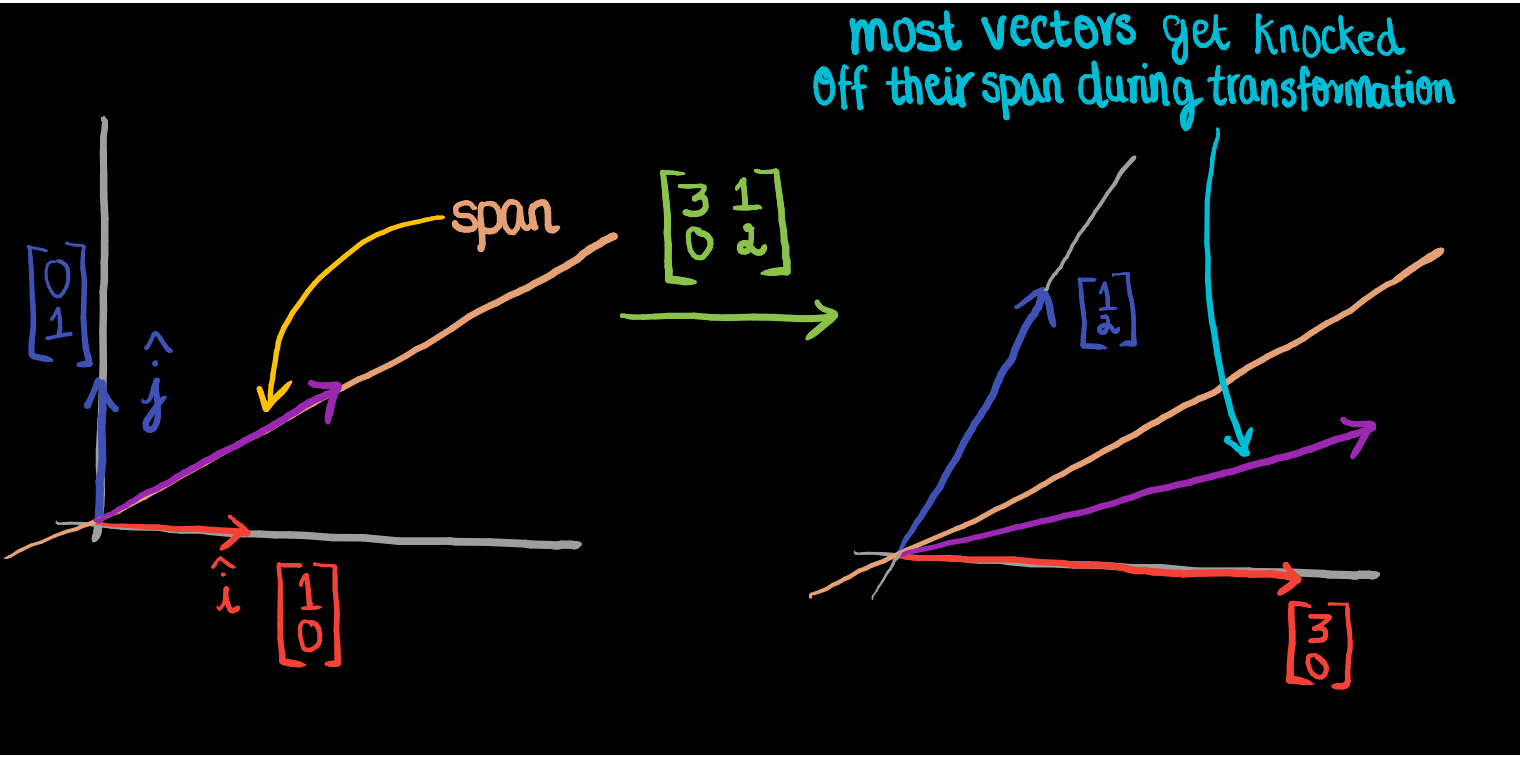
What vector $\vec{p}$ has the property that:

$$\underbrace{\begin{bmatrix} ? \\ ? \\ ? \end{bmatrix}}_{\vec{p}} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \det\left(\begin{bmatrix} x & v_1 & w_1 \\ y & v_2 & w_2 \\ z & v_3 & w_3 \end{bmatrix}\right)$$

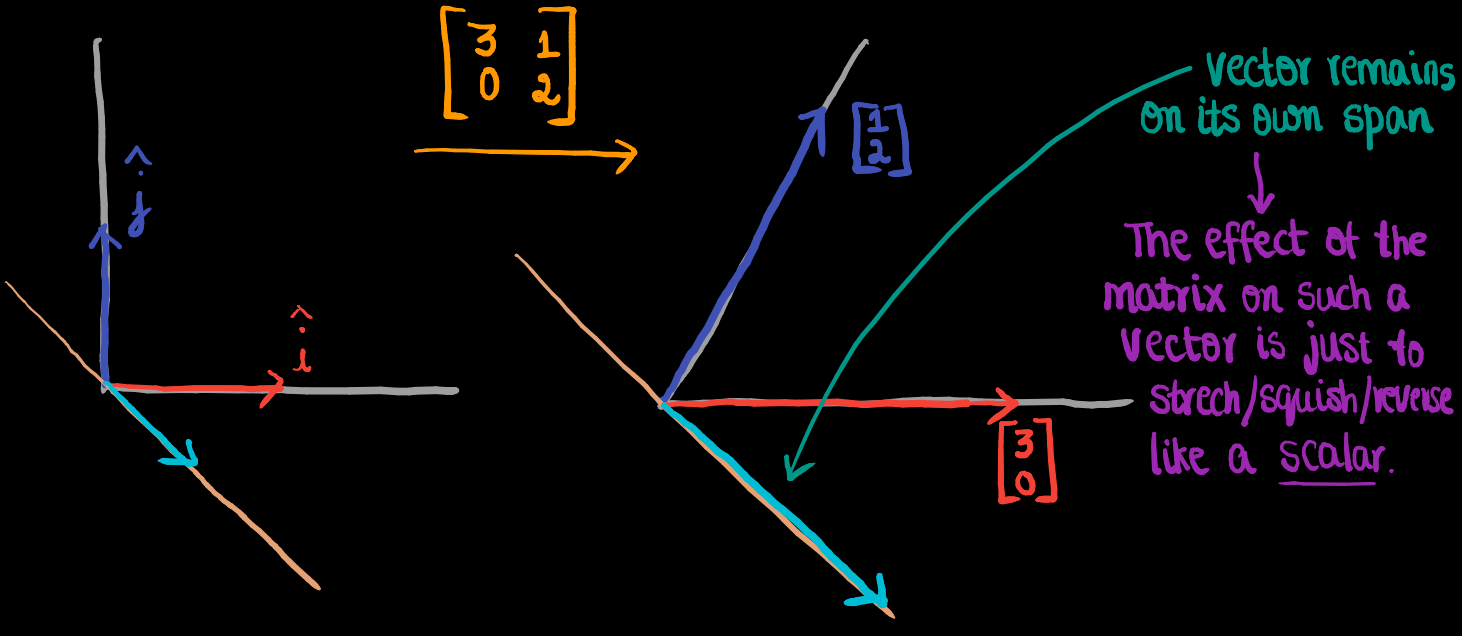
(v_1, v_2, v_3 = $\vec{v}$; w_1, w_2, w_3 = $\vec{w}$)

Geometric Interpretation to Computational Approach

EIGENVECTORS AND EIGENVALUES


most vectors get knocked
off their span during transformation
span
$\begin{bmatrix} 3 & 1 \\ 0 & 2 \end{bmatrix}$
$\begin{bmatrix} 0 \\ 1 \end{bmatrix}$
$\hat{j}$
$\hat{i}$
$\begin{bmatrix} 1 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$
$\begin{bmatrix} 3 \\ 0 \end{bmatrix}$

Some special vectors do remain on their own span.



These special vectors are called:

EIGENVECTORS of the transformation.

Each eigenvector has an associated EIGENVALUE which is: the factor by which it is streched or squished during the transformation.

Why could it be useful?

Consider 3D rotation.

Eigenvector ⟶ vector that remain on its own span

here, eigenvalue = 1.
Rotation doesn't strech/squish.

Axis of rotation!

It is easier to think about 3D rotation interms of axis of rotation and an angle, instead of a 3x3 matrix.

Can we solve $Ax = \lambda x$ for the eigenvectors and eigenvalues of A?

$$Ax = \lambda x$$
$$(A - \lambda I)x = 0$$

Both λ and x are unknown.
Need to be clever to solve the problem.

In order for λ to be an eigenvector, $A - \lambda I$ must be singular.

$$\det(A - \lambda I) = 0$$

characteristic equation of A

An $n \times n$ matrix A can have at most n distinct eigenvalues.

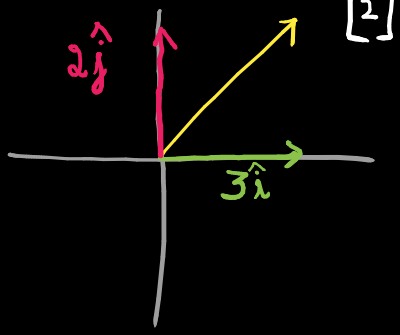
Once we've found an eigenvalue λ,
we can use elimination to find the null space of $A - \lambda I$.

The vectors in the null space are eigenvectors of A
with eigenvalue λ.

Note: For triangular matrices, the eigenvalues are exactly
the entries on the diagonal.

Change of Basis

$$\begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

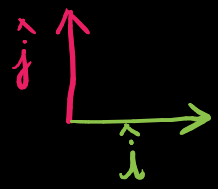


Think of coordinates as scalars.

The basis vectors encapsulates all of the implicit assumptions of our coordinate system.

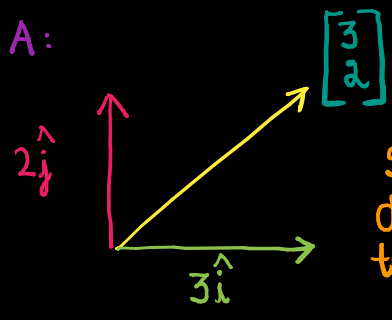
→ here, first coordinate → →
second coordinate →
unit of distance

Anyway to translate from a vector and sets of numbers is called coordinate system.



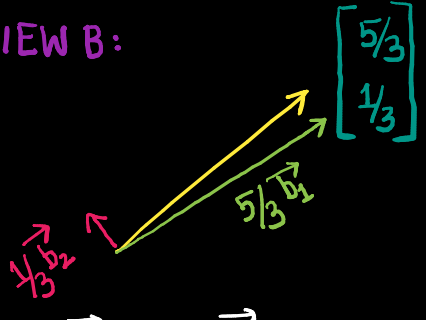
Basis vectors of standard coordinate system

VIEW A:



Same vector, but different numbers to describe it.

VIEW B:

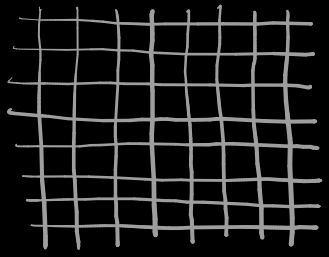


$\hat{i}$ and $\hat{j}$ are the basis vectors.

$\vec{b}_1 = \begin{bmatrix}2\\1\end{bmatrix}$, $\vec{b}_2 = \begin{bmatrix}-1\\1\end{bmatrix}$ in this view.

$\vec{b}_1$ and $\vec{b}_2$ are the basis vectors.

In this view, $\vec{b}_1 = \begin{bmatrix}1\\0\end{bmatrix}$ and $\vec{b}_2 = \begin{bmatrix}0\\1\end{bmatrix}$.



Grid is a construct and it depends on our choice of basis. Space itself has no intrinsic grid.

→ But, origin remains the same.

How do you translate between coordinate systems?

$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$

$\begin{bmatrix} -1 \\ 1 \end{bmatrix}$ in our coordinate system.

$\vec{b}_2$ $\vec{b}_1$ $2\vec{b}_2$ $-1\vec{b}_1$

$\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$-1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$ in our standard coordinate system.

$$\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ 2 \end{bmatrix} = -1 \begin{bmatrix} 2 \\ 1 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 1 \end{bmatrix} = \begin{bmatrix} -4 \\ 1 \end{bmatrix}$$

change of basis matrix

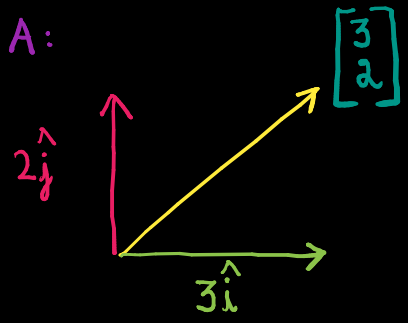
Columns represent view B's basis vector in view A's language

A transformation that moves the standard basis vectors $\hat{i}$ and $\hat{j}$ of view A to view B's basis vectors.

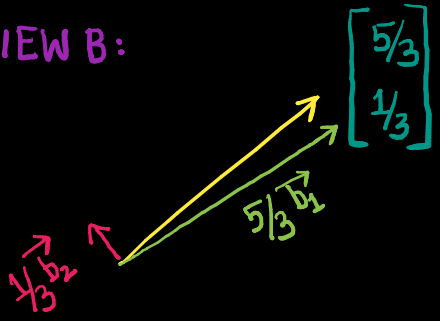
view A's grid → view B's grid $\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}$ view A's language ← view B's language

view A's grid ← view B's grid $\begin{bmatrix} 2 & -1 \\ 1 & 1 \end{bmatrix}^{-1}$ view A's language → view B's language

VIEW A:



VIEW B:



$$\begin{bmatrix} 1/3 & 1/3 \\ -1/3 & 2/3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = \begin{bmatrix} 5/3 \\ 1/3 \end{bmatrix}$$

Inverse change of basis matrix

In view A's language

In view B's language

$$A = \begin{bmatrix} | & | \\ | & | \end{bmatrix}$$

View B's basis vectors written in View A's coordinates

vector in view B's coordinates

$$\begin{bmatrix} x_i \\ y_i \end{bmatrix} = A^{-1} \begin{bmatrix} x_o \\ y_o \end{bmatrix}$$

same vector in view A's coordinates

vector in view B's coordinates

$$A \begin{bmatrix} x_i \\ y_i \end{bmatrix} = \begin{bmatrix} x_o \\ y_o \end{bmatrix}$$

same vector in view A's coordinates

$A^{-1}MA$ expression

90° rotation

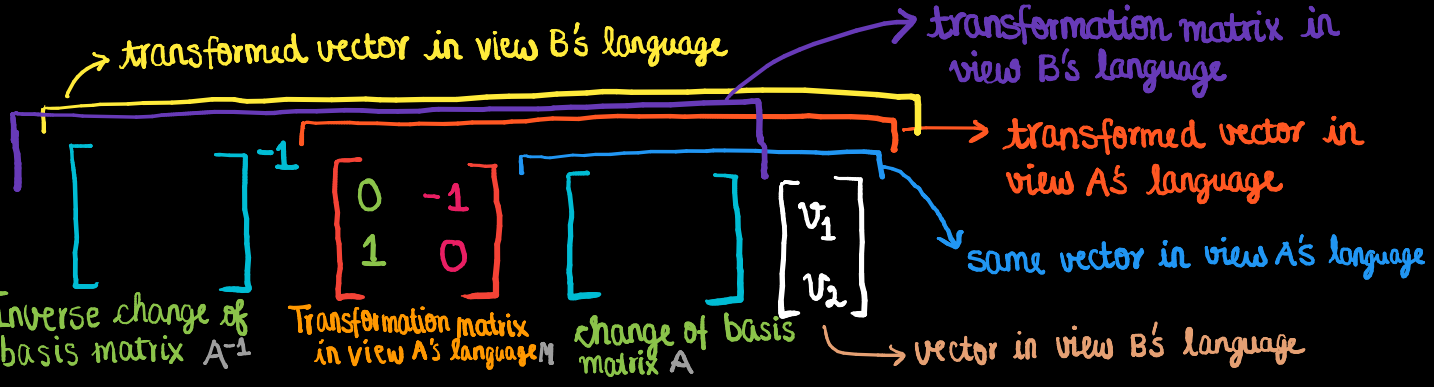
Follows our choice of basis vectors

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

where $\hat{i}$ and $\hat{j}$ lands recorded in view A's standard coordinate system.

How would the same 90° rotation of space described in view B?

→ should tell the landing spots of its basis vector in its language.



$$a_1x + b_1y + c_1z = d_1$$
$$a_2x + b_2y + c_2z = d_2$$
$$a_3x + b_3y + c_3z = d_3$$

$$A = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{bmatrix}$$

coefficient matrix

$$x = \frac{\det(A_1)}{\det(A)} \qquad y = \frac{\det(A_2)}{\det(A)} \qquad z = \frac{\det(A_3)}{\det(A)}$$

Cramer's Rule

complexity: $O(n!)$

A_i is A with i^{th} column replaced with $\begin{bmatrix} d_1 \\ d_2 \\ d_3 \end{bmatrix}$.

PART TWO
MATRIX FACTORIZATION

Row Echelon Form (ref)

→ French: staircase-like appearance

$$\begin{bmatrix} a & * & * & * & * & * \\ 0 & b & * & * & * & * \\ 0 & 0 & 0 & c & * & * \\ 0 & 0 & 0 & 0 & 0 & d \end{bmatrix}$$

1st non-zero entry of any row

All entries of a column lying below the leading entry of some row are 0.

not every row has a leading entry.

A square matrix in ref is upper triangular matrix.

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 6 & 7 \\ 0 & 0 & 8 & 9 \\ 0 & 0 & 0 & 10 \end{bmatrix}_{4\times4}$$

Reduced Row Echelon Form (rref)

Row Echelon Form +

- In any non-zero row, leading entry $\longrightarrow$ 1
- Leading entries are the only non-zero entries in a column

Elementary Row Operations

* Multiply a row by a non-zero scalar.
* Multiply a row by a non-zero scalar and add the result to another row.
* Swap two rows.

Two matrices A_1 and A_2 are row-equivalent if we can get from A_1 to A_2 by a (finite) sequence of elementary row operations.

These elementary row operations can be interpreted as multiplying the augmented matrix by a special non-singular matrix.

Exchange and Permutation Matrices

permuting the columns of identity matrix

permuting any two columns of identity matrix

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

Multiplying any $4 \times n$ matrix on the left will exchange 2^{nd} and 4^{th} rows of the matrix.

Multiplying any $m \times 4$ matrix on the right will exchange 2^{nd} and 4^{th} columns of the matrix.

Gaussian Elimination Matrices

→ Key step: transform vector $\begin{bmatrix} a \\ \alpha \\ b \end{bmatrix} \longrightarrow \begin{bmatrix} a \\ \alpha \\ 0 \end{bmatrix}$ where $a \in \mathbb{R}^k$, $0 \neq \alpha \in \mathbb{R}$, $b \in \mathbb{R}^{n-k-1}$

This can be accomplished by left matrix multiplication.

$$\begin{bmatrix} I_{k\times k} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\alpha^{-1}b & I_{(n-k-1)\times(n-k-1)} \end{bmatrix} \begin{bmatrix} a \\ \alpha \\ b \end{bmatrix} = \begin{bmatrix} a \\ \alpha \\ 0 \end{bmatrix}$$

Both G_i and G_i^{-1} are lower triangular matrices.

Gaussian elimination matrix G_i

→ invertible with inverse

$$G_i^{-1} = \begin{bmatrix} I_{k\times k} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \alpha^{-1}b & I_{(n-k-1)\times(n-k-1)} \end{bmatrix}$$

System of linear equations

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= b_2 \\
\dots\dots\dots & \\
a_{m1}x_1 + a_{m2}x_2 + \dots + a_{mn}x_n &= b_m
\end{aligned}
$$

Coefficient matrix

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

Column vector of R.H.S.

$$
\begin{bmatrix}
b_1 \\
b_2 \\
\vdots \\
b_m
\end{bmatrix}
$$

Augmented Matrix

$$
\left[\begin{array}{cccc|c}
a_{11} & a_{12} & \dots & a_{1n} & b_1 \\
a_{21} & a_{22} & \dots & a_{2n} & b_2 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
a_{m1} & a_{m2} & \dots & a_{mn} & b_m
\end{array}\right]
$$

Let $C_1 = (A_1 | b_1)$ and $C_2 = (A_2 | b_2)$ be two row equivalent augmented matrices, then the system of equations $A_1 v = b_1$ and $A_2 v = b_2$ have the same solution set.

The goal of GAUSSIAN ELIMINATION is to convert the augmented matrix into row echelon form (ref).

Every matrix is row equivalent to a matrix in ref.

GAUSSIAN ELIMINATION

Let A be a $m \times n$ matrix.

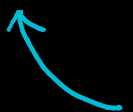
1. Let $l = 0$.
2. Set $i = l+1$. If there is no row below the l^{th} row which contains a non-zero entry, then STOP.
3. Otherwise, pick the smallest j such that there is an index $i' > i$ such that $a_{i'j} \neq 0$.
4. Swap rows i and i'.
   (a) Let $\lambda = 1/a_{ij}$ and multiply the i^{th} row by $\lambda = 1/\mu$.
   (b) If there are no non-zero elements below the i^{th} row in the j^{th} column, then increase l by 1 and then return to (2).
   (c) Pick the smallest j s.t. there is an index $i' > i$ with $\lambda = a_{i'j} \neq 0$. Multiply the i^{th} row by $-\lambda$ and add it to row i' and return to (b).

GAUSS-JORDAN ELIMINATION

The goal of Gauss-Jordan reduction is to convert the augmented matrix into reduce ow echelon form (rref).

First apply Gaussian elimination until A is in echelon form, then:

Starting from the rightmost pivot, iterate through each pivot column in reverse order. We have to eliminate the entries above each pivot.



additional to Gaussian elimination.

Matrix Decomposition

↑ transformation of a given matrix into a given canonical form.

Two major purposes

- Computational convenience
- analytic simplicity

Decompositions are like factorization of numbers... or polynomials

$36 = 2 \times 2 \times 3 \times 3$

→ prime factorization

$36 = 12 \times 3$

$36 = 18 \times 2$

$36 = 9 \times 4$

$36 = 6 \times 3 \times 2$

$p(x) = x^2 + 5x + 6$

$= (x+3)(x+2)$

→ quadratic polynomial as product of 2 linear polynomials

Most matrix computation not feasible in real world to be calculated in an optimal explicit way.

Ex. matrix inversion, matrix determinant, solving linear system, least square fitting

convert difficult matrix computation problem into several easier tasks.

The Big Six Matrix Factorizations

- Cholesky Decomposition ← Solving positive definite linear systems
- LU Decomposition ← Solving general linear systems
- QR Decomposition ← solving least squares problem; orthogonalization
- Schur Decomposition ← Computing eigenvalues, eigenvectors; Computing invariant subspaces
- Spectral / Eigen Decomposition ← problems involving eigenvalues of Hermitian matrices
- Singular Value Decomposition ← determining matrix rank, solving rank-deficient least squares

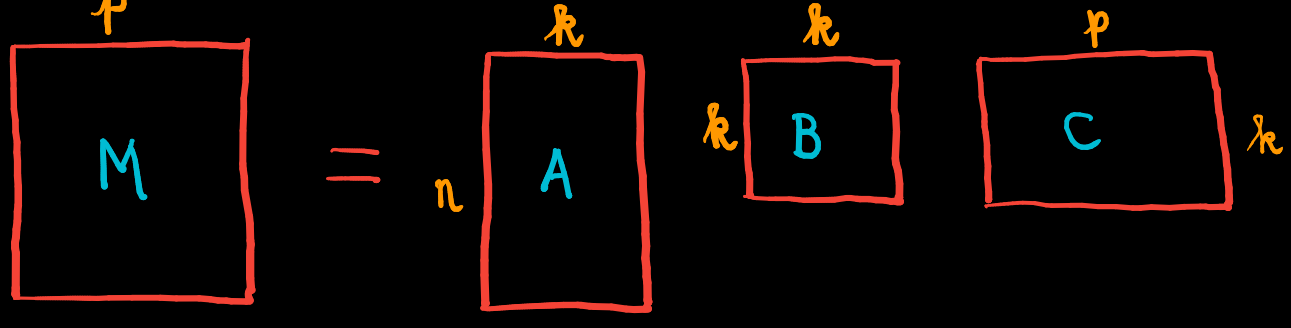
Matrix M as a product of a set of (typically) 2-3 simpler matrices.

→ gives us an idea of the inherent structures in M

ex. compact, w/ less dimensions, w/ less rows/cols, triangular, diagonal

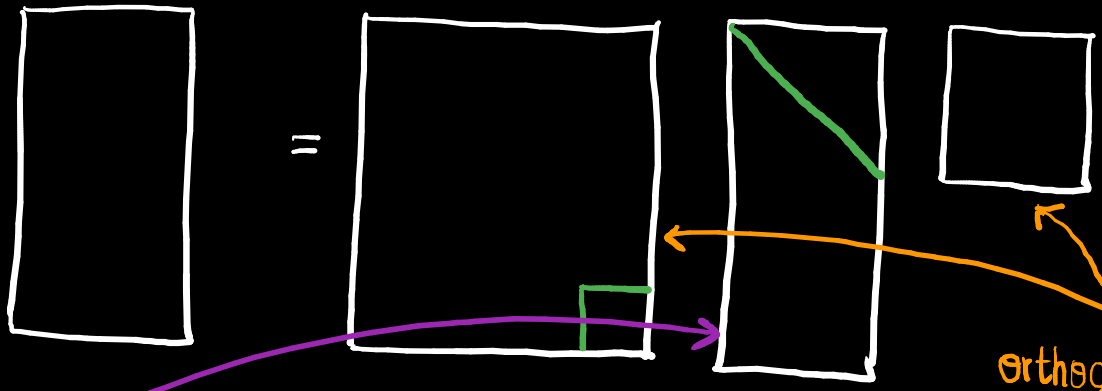
$$M = ABC$$

Assume $n > p$, $k < p$



SINGULAR VALUE DECOMPOSITION

Singular Value Decomposition (SVD)



diagonal matrix
(scaling operation)

$$A = U\Sigma V^T$$

orthogonal factors on both sides

SVD says that we can replace any transformation by:

- rotation from 'input' coordinates to convenient coordinates.
- followed by a simple scaling operation.
- followed by rotation into 'output' coordinates.

Diagonal Σ comes w/ elements sorted in descending order.

$$A = U\Sigma V^T$$

Columns of U ← left singular vectors u_i

rows of V^T ← right singular vectors v_i

entries on diagonal of Σ ← singular values

Geometric Interpretation of SVD

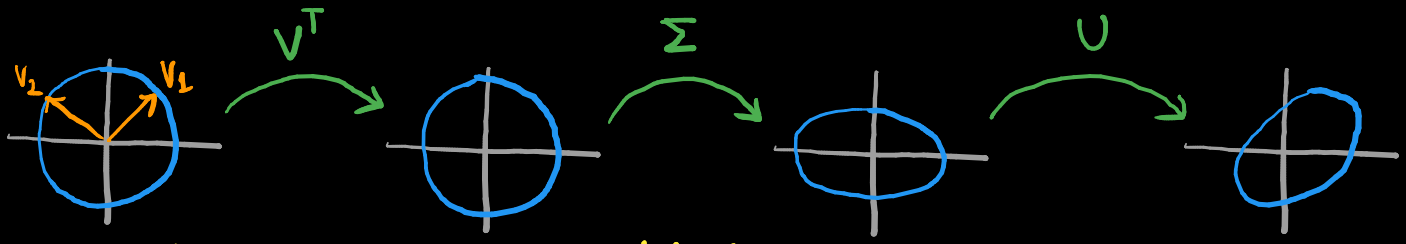
Let $U = [u_1 \; u_2]$ and $V^T = \begin{bmatrix} v_1^T \\ v_2^T \end{bmatrix}$.

If we take the unit circle and transform it by A, we get an ellipse.

Left singular vector
↳ major and minor axes of ellipse

Right singular vectors get mapped to major and minor axes.

If we break the transformation down into three stages.



rotated to align v_i with coordinate axes

scaled along these axes

rotated to align the ellipse with u_i

What happens when we have a singular matrix?

rank-deficient

dim. of the span of columns

Consider 3x3 matrix of rank 2.

- range is a subspace of $\mathbb{R}^3$, not all of $\mathbb{R}^3$.
- maps the sphere to a flat ellipse rather than an ellipsoid.

One of the singular values (last) is zero.

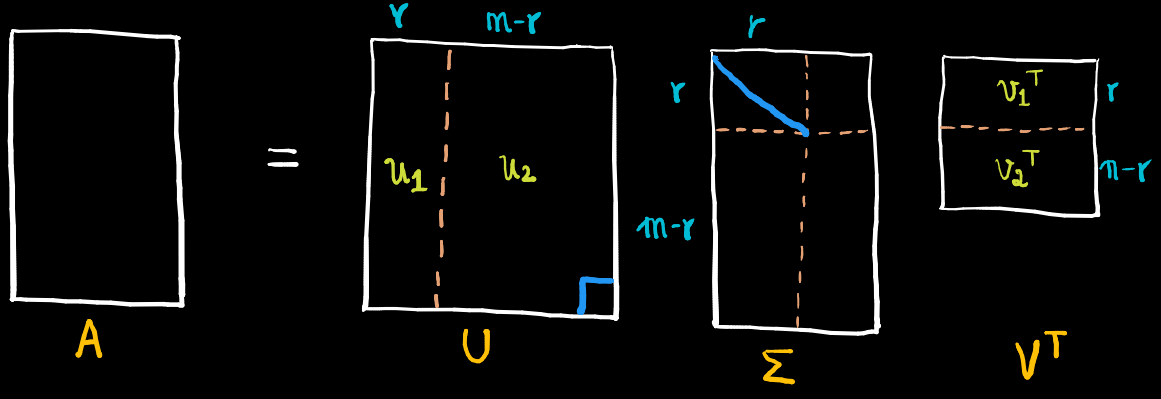
$$A = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} \begin{bmatrix} a & & \\ & b & \\ & & 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ v_2^T \\ v_3^T \end{bmatrix}$$

decreasing order

The last left singular vector is normal to that ellipse.

SVD of a rank-r matrix

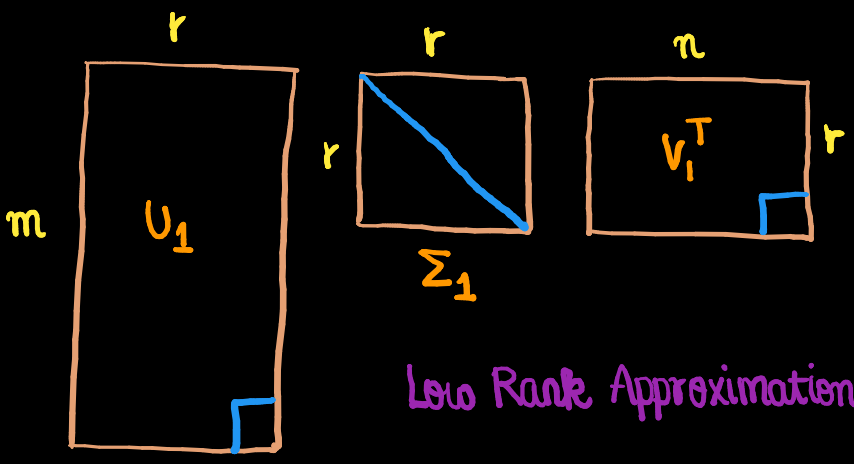
$r < n < m$



The matrix is rank-deficient ($r < n$) with an r-dimensional range and an $(n-r)$-dimensional null space. ← rank-deficient matrix has non-trivial null space.

If we are just interested in a factorization of A from which we can reconstruct A, we need only the first r singular vectors and singular values.

$$A_r = U_1 \Sigma_1 V_1^T$$



singular values σ_1 to σ_r will be positive.

$$A_r v_i = \sigma_i u_i$$

where $i = 1, \ldots, r$

Low Rank Approximation of A

Singular Vectors of A and Eigenvectors of $S = A^TA$

$$A^TA = (U\Sigma V^T)^T(U\Sigma V^T) = V\Sigma^TU^TU\Sigma V^T = V\Sigma^T\Sigma V^T$$

V is the eigenvector matrix for the symmetric positive (semi) definite A^TA.

$\Sigma^T\Sigma$ must be the eigenvalue matrix of A^TA.

The singular vectors v_i are the eigenvectors q_i of $S = A^TA$.

The eigenvalues λ_i of S are the same as σ_i^2 for A.

The rank r of S equals the rank r of A.

SPECTRAL/EIGENVALUE DECOMPOSITION

Recasts a matrix interms of its eigenvalues and eigenvectors.

A be a real, symmetric $d \times d$ matrix with eigenvalues $\lambda_1, \ldots, \lambda_d$ and corresponding orthonormal eigenvectors $u_1, \ldots, u_d$.

$$A = \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ u_1 & u_2 & \cdots & u_d \\ \downarrow & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_d \end{bmatrix} \begin{bmatrix} \leftarrow & u_1 & \rightarrow \\ \leftarrow & u_2 & \rightarrow \\ & \vdots & \\ \leftarrow & u_d & \rightarrow \end{bmatrix}$$

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

Q Λ Q^T

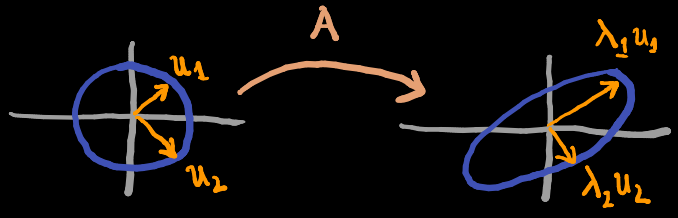
Every symmetric matrix of dimension n has a set of (not necessarily unique) n orthogonal eigenvectors. All eigenvalues are real.

Every real, symmetric matrix A can be decomposed into real-valued eigenvectors and eigenvalues: $A = Q \Lambda Q^T$

Q: orthogonal matrix of eigenvectors of A

Λ: diagonal matrix of eigenvalues

Think of A as scaling space by λ_i in the direction of u_i.



Eigendecomposition is not unique.

↑ when two eigenvalues are the same.

By convention, order entries of Λ in descending order.

If any eigenvalue is zero, then the matrix is singular.

↳ the corresponding eigenvector $Au = 0u = 0$.

Positive Definite Matrix

If a symmetric matrix A has the property:

$$x^T A x > 0 \quad \text{for any nonzero vector } x$$

then A is called **positive definite**.

If the above inequality is not strict,

then A is called **positive semidefinite**.

For positive (semi) definite matrices,

all eigenvalues are positive (non-negative).

If A is not square, eigendecomposition is undefined.

→ Every real matrix has a SVD.

SVD is more general than eigendecomposition.

SVD can be interpreted interms of eigendecomposition.

- Left singular vectors are eigenvectors of AA^T.
- Right singular vectors are eigenvectors of A^TA.
- Nonzero singular values of A are square roots of eigenvalues of A^TA and AA^T.

Orthogonal matrix

columns are unit length and mutually perpendicular.

$$Q = \begin{bmatrix} \uparrow & & \uparrow \\ q_1 & \cdots & q_n \\ \downarrow & & \downarrow \end{bmatrix}$$

$$\|q_j\| = 1 \quad \text{for all } j$$

$$q_i \cdot q_j = 0 \quad \text{when } i \neq j$$

Orthogonal transformations correspond to the geometric ideas of rotation and mirror reflection.

For a square orthogonal matrix Q,

$Q^T Q = I$ ← Columns are orthonormal

$Q Q^T = I$ ← rows are orthonormal

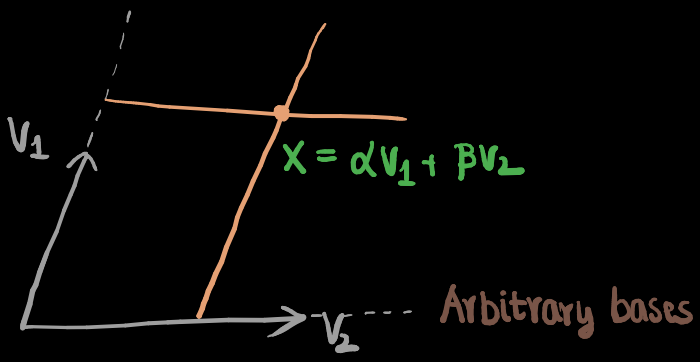
$Q^{-1} = Q^T$ ← from previous two, transpose is inverse

$\det Q = \pm 1$

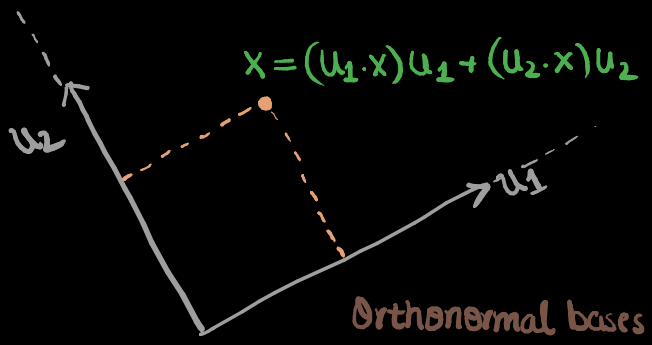
$\|Qx\|_2 = \|x\|_2$ ← orthogonal transformation preserves lengths in Euclidean norm

$(Qx).(Qy) = x.y$ ← orthogonal transformation preserves angles

Computing coordinates in arbitrary vs orthonormal bases

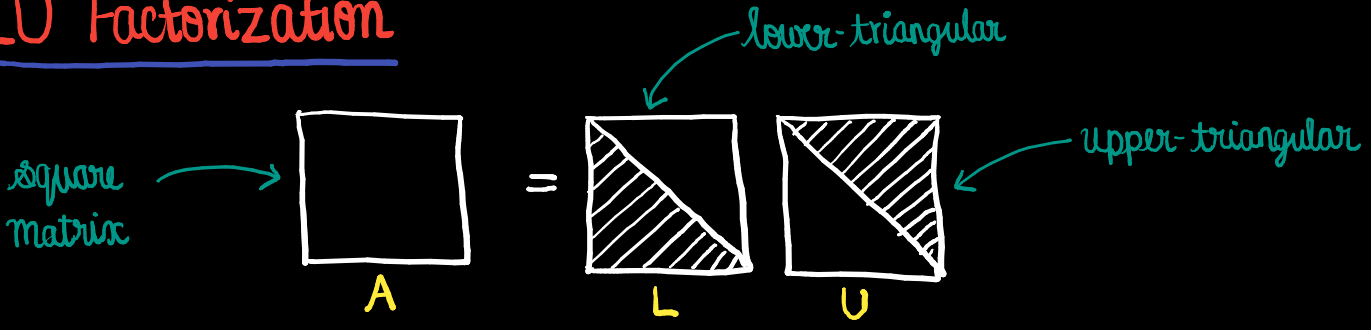


Finding α, β requires solving a linear system.



No solve required, just dot products.

LU Factorization



Think of this equation as L changing the coordinate system in which we express the "output" of A.

In the new coordinate system, transformation defined by A is represented by an upper triangular matrix.

$A = LU$

Think of this equation as L changing the coordinate system in which we express the "output" of A.

In the new coordinate system, transformation defined by A is represented by an upper triangular matrix.

U 'does the same thing as' A but just returns its result in a different coordinate system.

When we solve $Ax = b$ by computing: $y = L \backslash b$ and $x = U \backslash y$,

(matrix division)

the vector y is just x expressed in a coordinate system where A becomes convenient to work with.

Using LU decomposition to solve linear systems

Suppose $A = LU$ and want to solve: $Ax = LUx = V$.

1. Set $W = \begin{bmatrix} u \\ v \\ w \end{bmatrix} = Ux$.

2. Solve the system $LW = V$.
   Solution is W_0.

   forward substitution, since L is lower triangular.

3. Solve the system $UX = W_0$.

   backward substitution, since U is upper triangular.

Finding an LU decomposition

For a given matrix, there are many different LU decompositions.

There is a unique LU decomposition in which L matrix has ones on the diagonal ⟶ L is called lower unit triangular matrix.

Read more:

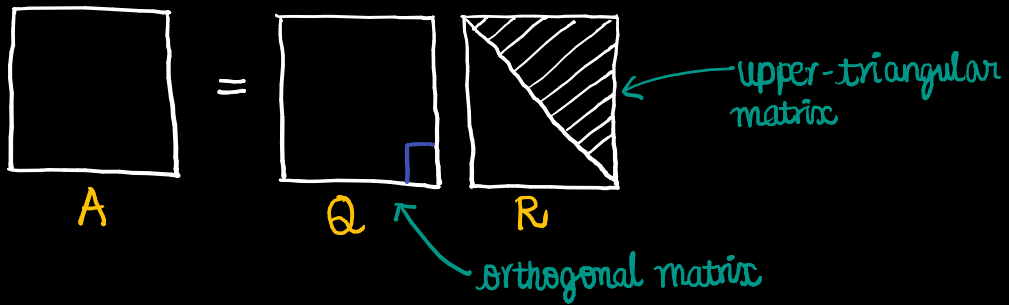
1. Recursive leading-row-column LU algorithm
2. Pivoting and LU decomposition with partial pivoting

QR Factorization

better than LU by finding not just a coordinate system in which our transformation becomes upper-triangular,

but finding an orthogonal coordinate system in which our transformation becomes upper-triangular.

$A = QR$



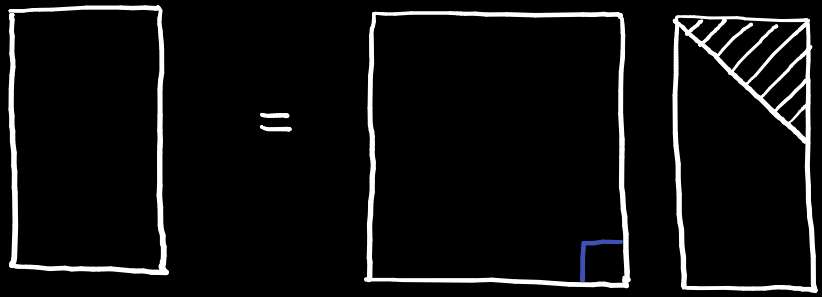
This factorization has some similarities to LU.

- L is easy to invert; so is Q.
- U is upper-triangular; so is R.

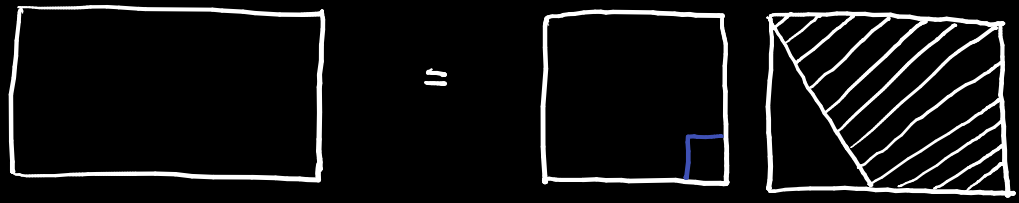
→ leads to better (more accurate) methods for least squares problem.

But, Q will preserve norm, dot products, etc. at the same time!

QR can be applied to non-square matrices.



(or)



Computing the QR factorization.

We will reduce A to the upper triangular matrix R by applying a sequence of special orthogonal transformations with simple structure.

Householder reflections

The Q factor is then the inverse product of all these reflections.

Read more: (i) Gauss transforms for LU factorization
(ii) Householder reflections for QR factorization.

PART THREE

PROBABILITY

Probability Theory

provides a consistent framework for quantification and manipulation of uncertainty.

Uncertainty in pattern recognition

(i) noise on measurements (ii) finite size datasets

Frequentist Interpretation

Fraction of times event occurs in experiment.

Bayesian Approach

Quantification of plausibility or strength of belief of an event.

Random Variable X

Stochastic variable sampled from a set of possible outcomes $x \in X$

- discrete or continuous

Throwing a dice
$X = \{1, 2, \ldots, 6\}$

Probability dist.
$p(x) \geq 0, \quad x \in X$

2 random variables X and Y

$X = \{x_1, \ldots, x_n\}$

$Y = \{y_1, \ldots, y_m\}$

Joint Probability $p(X = x_i, Y = y_j)$

Marginal Probability

$$p(X = x_i) = \sum_{j=1}^{m} p(X = x_i, Y = y_j)$$

Conditional probability of Y given X

$$p(Y = y_j \mid X = x_i) = \frac{P(X = x_i, Y = y_j)}{P(X = x_i)}$$

Product Rule

Continuous Random Variables

A random variable is continuous if there is a probability density function $f(x) \geq 0$ s.t. for $-\infty < x < \infty$:

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx$$

Cumulative Distribution Function (CDF)

For a random variable X, the CDF is defined as:

$$F(a) = F_X(a) = P(X \leq a), \text{ where } -\infty < a < \infty.$$

BAYES THEOREM

$$p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)}$$

likelihood of $X = x$ given $Y = y$

posterior probability of $Y = y$ (after observing x)

prior probability of $Y = y$ (before observing x)

evidence for $X = x$

Expectations

$x \sim p(X)$ random variable $x \in X$ and fn. $f : X \to \mathbb{R}$

$$\mathbb{E}[f] = \mathbb{E}_{x \sim p(X)}[f(x)]$$

$$= \begin{cases} \sum_{x} f(x)\, p(x) & \rightarrow \text{discrete} \\ \int_{x} f(x)\, p(x)\, dx & \rightarrow \text{continuous} \end{cases}$$

Conditional Expectation

$$\mathbb{E}[f|y] = \mathbb{E}_{x \sim p(X|Y=y)}[f(x)]$$

$$= \begin{cases} \sum_x f(x)\ p(X|Y=y) \rightarrow \text{discrete} \\ \int_x f(x)\ p(X|Y=y)\,dx \rightarrow \text{continuous} \end{cases}$$

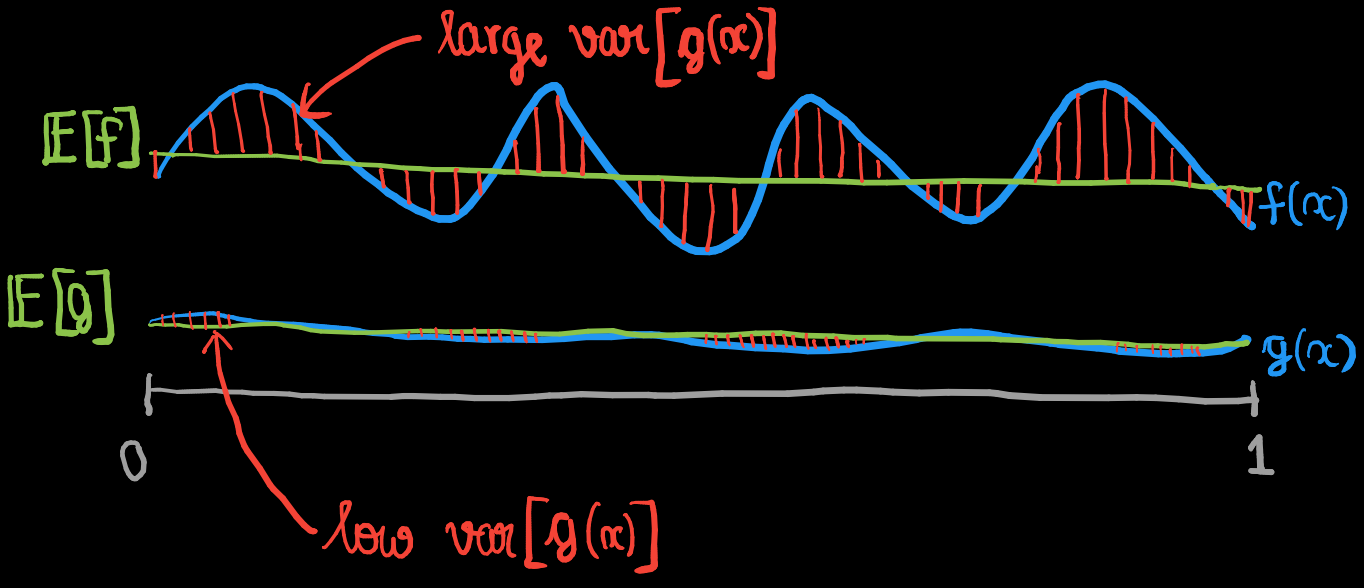
Variance

The expected quadratic distance between f and its mean $\mathbb{E}[f]$.

$$\operatorname{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2 \right]$$

$$= \mathbb{E}\left[f(x)^2\right] - \mathbb{E}\left[f(x)\right]^2$$

Example: $x \sim \text{uniform}[0,1]$ $f(x)$, $f: X \to \mathbb{R}$



Covariance between 2 random variables

Measures the extent to which X and Y vary together.

$$\mathrm{cov}[x, y] = \mathbb{E}_{x,y \sim p(x,y)}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

$$= \mathbb{E}[X, Y] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

Covariance Matrix

Vectors of random variables x and y. $\in \mathbb{R}^D$

$$\operatorname{cov}[x, y] = \mathbb{E}_{x,y \sim p(x,y)}\Big[\underbrace{(x - \mathbb{E}[x])}_{\mathbb{R}^{D\times 1}}\underbrace{(y - \mathbb{E}[y])^T}_{\mathbb{R}^{1\times D}}\Big] \in \mathbb{R}^{D\times D}$$

$$= \mathbb{E}[xy^T] - \mathbb{E}[x]\,\mathbb{E}[y]^T$$

Covariance b/w independent variables

$$p(x, y) = p(x) \cdot p(y)$$

$$\text{cov}[x, y] = \mathbb{E}[x, y] - \mathbb{E}[x]\,\mathbb{E}[y] = 0$$

Note: $\text{cov}[x, y] = 0$ doesn't imply x, y independent.

Some Important Distributions

Bernoulli

r.v. can take two possible values $\{0,1\}$ specified by parameter p

$$P(X) = p^{x}(1-p)^{1-x}$$

Some Important Distributions

parameterized by average arrival rate λ

Poisson

It measures probability of number of events happening over a fixed period of time, given a fixed average rate of occurence, and that events take place independently of the time since the last event.

$$P(X=k) = \frac{\lambda^x}{x!} e^{-\lambda}$$

mean of Poisson r.v. is λ and its variance is also λ.

Some Important Distributions

Gaussian aka Normal Distribution

- one of the most versatile distributions in probability theory

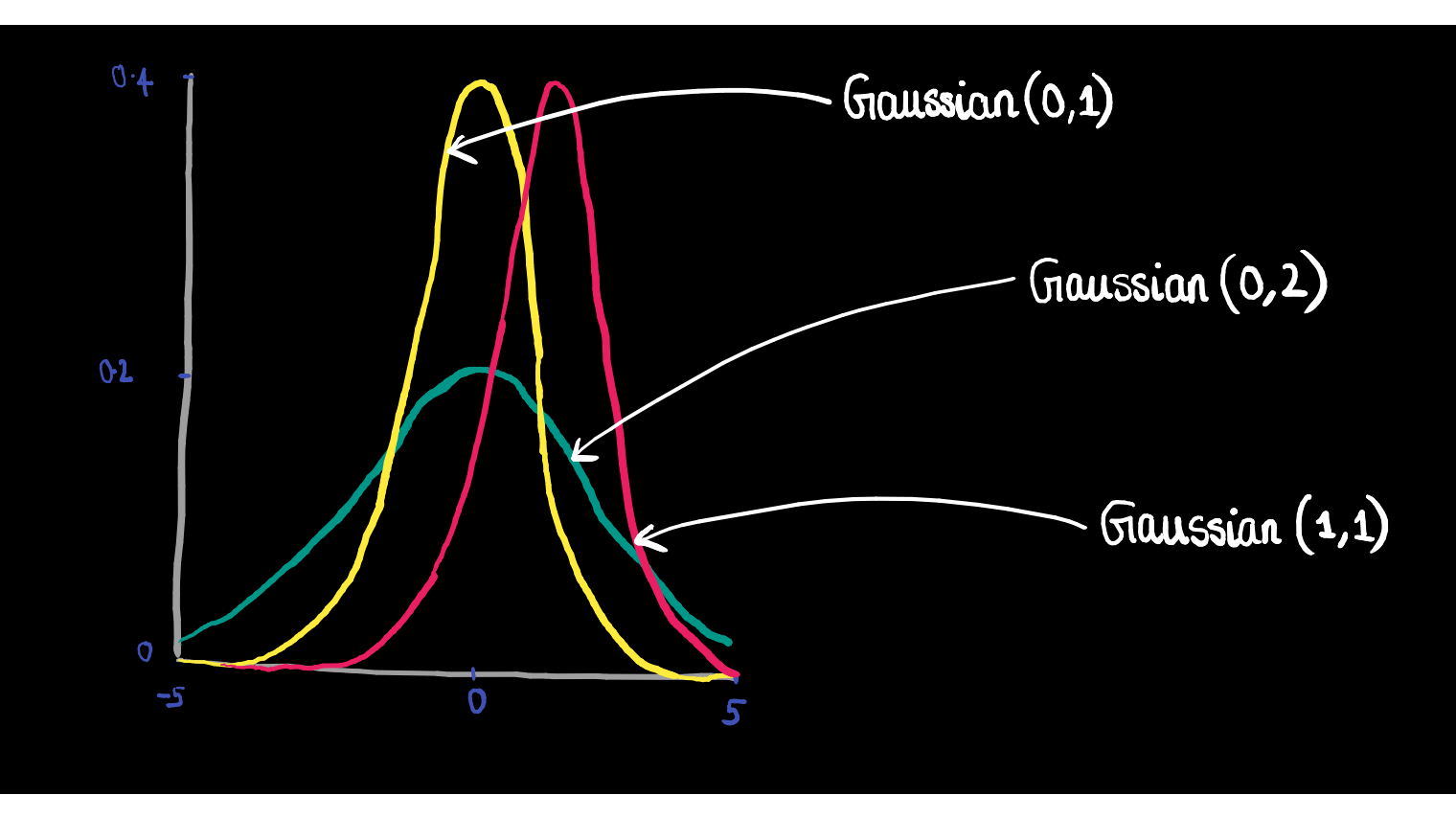
- can be used to approximate
  - binomial, when # experiments is large
  - Poisson, when average arrival rate is high

Gaussian distribution $\longrightarrow$ determined by 2 parameters

(i) mean μ (ii) variance σ^2

The probability density fn. is given by:

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


0·4
0·2
0
-5
0
5
Gaussian (0,1)
Gaussian (0,2)
Gaussian (1,1)

Multi-variate Gaussian distribution

parameterized by (μ, Σ)

For a k-dim multi-variate Gaussian.

vector of means in $\mathbb{R}^k$

covariance matrix in $\mathbb{R}^{k \times k}$

$\Sigma_{ii} = \text{Var}(X_i)$

$\Sigma_{ij} = \text{Cov}(X_i, X_j)$

PDF is defined over vectors of input.

$$f(x) = \frac{1}{\sqrt{2\pi^k |\Sigma|}} e^{\frac{-1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$

$\det(A) \equiv |A|$

$\text{inverse}(A) \equiv A^{-1}$

Confidence Intervals

If a, b be 2 numbers, where interval $[a,b]$ contains the parameter θ with CI of $1-\alpha$, then $[a,b]$ is $1-\alpha$ confidence interval for θ.

$$P(\theta \in [a,b]) = 1-\alpha$$

θ has 95% CI of $[a,b]$ means:

If we were to repeat the experiment, then 95% of the time, the true parameter would be in that interval.

doesn't mean that we believe it is in that interval given, the actual data we have observed.

PART FOUR

STATISTICAL ESTIMATION

Probability versus Statistics

Given a model

Ber(p=0.5)

Predict data

P(THHTHH) = ?

Probability versus Statistics

Predict model

Given data

Ber(p=?)

THHTHH

We focus the opposite way.

Likelihood

probability of seeing the data,

given parameter θ of the assumed model

$$L(x|\theta) = P(\text{seeing data} \mid \theta)$$

$$= P(x_1, \ldots, x_n \mid \theta)$$

$$L(x|\theta) = \prod_{i=1}^{n} P_x(x_i|\theta)$$

independence

Example: Given iid samples from $\text{Ber}(\theta)$

$x = (x_1, x_2, x_3) = (1, 0, 1)$

θ: probability of success

$$L(x|\theta) = \prod_{i=1}^{3} p_X(x_i|\theta) = p_x(1|\theta) \cdot p_x(0|\theta) \cdot p_x(1|\theta)$$

$$= \theta(1-\theta)\theta$$

$$L(x|\theta) = \theta^2(1-\theta)$$

In the previous example,

θ ⟵ probability of success.

Guess which value of θ maximizes likelihood.

We observed 2 successes out of 3 trials.

Guess for maximum likelihood estimate $\hat{\theta} = \frac{2}{3}$

estimate

MAXIMUM LIKELIHOOD ESTIMATION

The value of θ that maximizes the 'probability' of seeing the data $L(x|\theta)$.

CALCULUS

To optimize a function:

local optima → derivative is 0

$$\hat{\theta}_{MLE} = \text{argmax}_{\theta} \, L(x|\theta)$$

In MLE, we are trying to find the θ that maximizes the likelihood.

or equivalently, the log likelihood

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \, L(x|\theta) = \arg\max_{\theta} \, \ln L(x|\theta)$$

The (usual) recipie to find the MLE.

(i) Compute the likelihood or log-likelihood of data.

(ii) Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).

(iii) Optionally, verify $\hat{\theta}$ is indeed a (local) maximizer.

- second derivative of $\hat{\theta}_{MLE}$ is negative ← if θ is a single parameter
- Hessian is negative semi-definite ← if θ is a vector of parameters.

MLE: 'best' θ that maximized likelihood $L(x|\theta)$

Shouldn't we be trying to maximize $P(\theta|x)$ instead?

It does not make sense unless θ is a random variable!

In MLE, we assumed θ was fixed but unknown.

In MLE, we assumed θ was fixed but unknown.

↑ Frequentist framework

- we estimate our parameter based on data alone
- θ is not a random variable

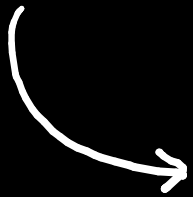
In Bayesian framework,

- θ (unknown parameter) is a random variable

We'll have some BELIEF distribution $\pi_\theta(\theta)$ ← density fn. over all possible values of the parameter.

After observing data x,
we'll have updated belief distribution $\pi_\theta(\theta|x)$

MAXIMUM A POSTERIORI ESTIMATION



allows us to incorporate
prior knowledge into our estimate.

MAXIMUM A POSTERIORI ESTIMATION

Unknown parameter(s) is a random variable Θ

Prior distribution $\pi_\Theta(\theta)$ → prior belief on Θ before seeing data

Posterior distribution $\pi_\Theta(\theta|x)$ → given data, updated belief on Θ after observing some data

By Bayes Theorem,

$$\pi_\Theta(\theta|x) = \frac{L(x|\theta) \cdot \pi_\Theta(\theta)}{P(x)} \propto L(x|\theta) \cdot \pi_\Theta(\theta)$$

π_θ is a PMF or PDF over possible values of θ.

→ In MAP, we are maximizing the posterior distribution $\pi_\theta(\theta|x)$.

→ we are finding the mode of the density/mass function.

$P(x)$ does not depend on θ, we can just maximize the numerator.

$$\hat{\theta}_{MAP} = \arg\max_\theta \pi_\theta(\theta|x) = \arg\max_\theta L(x|\theta) \cdot \pi_\theta(\theta)$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \pi_\theta(\theta|x) = \arg\max_{\theta} L(x|\theta) \cdot \pi_\theta(\theta)$$

This is exactly same as maximum likelihood,

except instead of just maximizing the likelihood,

we are maximizing the likelihood multiplied by the prior.

Read more: (i) Conjugate distributions as priors for MAP.

Examples

Samples $x = (0,0,1,1,0)$ from $\text{Ber}(\theta)$. (θ: unknown)

θ is unrestricted $\Rightarrow \theta \in (0,1)$. MLE for θ?

$$L(x|\theta) = \theta^2(1-\theta)^3$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \theta^2(1-\theta)^3 = \frac{2}{5}$$

Examples

Samples $x = (0,0,1,1,0)$ from $\text{Ber}(\theta)$. (θ: unknown)

$\theta \in \{0.2, 0.5, 0.7\}$. MLE for θ?

$$L(x|0.2) = 0.2^2 0.8^3 = 0.02048$$
$$L(x|0.5) = 0.5^2 0.5^3 = 0.03125$$
$$L(x|0.7) = 0.7^2 0.3^3 = 0.01323$$

$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta) = 0.5$$

We need to find which of the three acceptable θ values maximizes likelihood.

Examples

Samples $x = (0,0,1,1,0)$ from $\text{Ber}(\theta)$. (θ: unknown)

$\theta \in \{0.2, 0.5, 0.7\}$, but a random variable. Discrete prior $\pi_\theta(0.2) = 0.1$, $\pi_\theta(0.5) = 0.01$, $\pi_\theta(0.7) = 0.89$. MAP for θ?

$$\pi_\theta(0.2|x) = L(x|0.2)\,\pi_\theta(0.2) = 0.2^2\,0.8^3(0.1) = 0.0020480$$

$$\pi_\theta(0.5|x) = L(x|0.5)\,\pi_\theta(0.5) = 0.5^2\,0.5^3(0.01) = 0.0003125$$

$$\pi_\theta(0.7|x) = L(x|0.7)\,\pi_\theta(0.7) = 0.7^2\,0.3^3(0.89) = 0.0117747$$

$$\hat{\theta}_{MAP} = \arg\max_{\theta \in \{0.2, 0.5, 0.7\}} L(x|\theta)\,\pi_\theta(\theta) = 0.7$$

Bayes, Eigen and Beyond!

Naren Akash
Center for Visual Information Technology

Robotics Research Center
Summer School 2023

REFERENCES

1. Essence of Linear Algebra (3Blue1Brown) by Grant Sanderson
2. A Matrix Algebra Companion for Statistical Learning by Gaston Sanchez
3. UvA Machine Learning I by Erik Bekkers
4. Stanford CS109 Probability for Computer Scientists by Alex Tsun
5. Cornell CS3220 Computational Mathematics for CS by Anil Damle
6. Stanford CS229 Machine Learning by Andrew Ng
7. UofT CSC 411 Introduction to ML by M. Ren and M. Mackay
8. UIUC CS357 Numerical Methods by Andreas Kloeckner
9. UW Math 407 Linear Optimization by James V Burke
10. MIT 18.700 Linear Algebra by James McKernan

RESOURCES

A. MIT Linear Algebra by Gilbert Strang
B. Introduction to Probability by Bertsekas and Tsitsiklis
C. Harvard Statistics 110 by Joe Blitzstein