Linear Algebra

Tung Pham

Machine Learning Research Scientist, VinAI

11 Oct 2022

Dear readers,

As part of VinAI's effort to nurture the next generations of AI talents in Vietnam, we decide to release our Linear Algebra course to the public.

At VinAI, we use this course to train our residents and build up strong mathematical fundamentals that are essential for their research careers in AI fields, and later, for their Ph.D. studies at the top Computer Science program.

We do hope that the community (especially universities' students) find this course's material useful for their studies and careers.

Thank you.

VINAI Artificial Intelligence Application and Research JSC

Content

- History of (Linear) Algebra.
- Concepts in Linear Algebra: Linear map, kernel, matrix, range, rank, etc.
- Eigenvalues, eigenvectors.
- Spectral decomposition theorem.
- Other decompositions of a matrix.
- Some special matrices.
- Matrix norms.

Main focuses

- The relationship between linear algebra and geometry.
- Foster the intuition about linear algebra.
- Build/construct your understanding of linear algebra systematically.

What is Linear Algebra?

Wikipedia

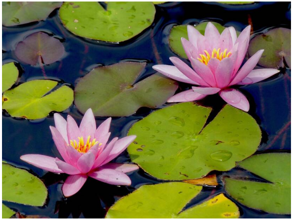
Linear algebra is central to almost all areas of mathematics.
Linear algebra is also used in most sciences and fields of engineering, because it allows modeling many natural phenomena, and computing efficiently with such models.

William Stein

Mathematics is the art of reducing any problem to linear algebra.

India example

One-third of a collection of beautiful water lilies was offered to Shiva, one-fifth to Vishnu, one-sixth to Surya, and one-fourth to the Devi. The six that remained were presented to the guru. How many water lilies were there in all?



Let x be the number of water lilies: $x = \frac{1}{3}x + \frac{1}{5}x + \frac{1}{6}x + \frac{1}{4}x + 6.$

Brief history of Linear Algebra

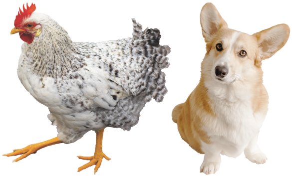
Vietnam example: Dog and chicken

There are 36 dogs and chickens, the total number of legs are 100. How many dogs and chickens are there?

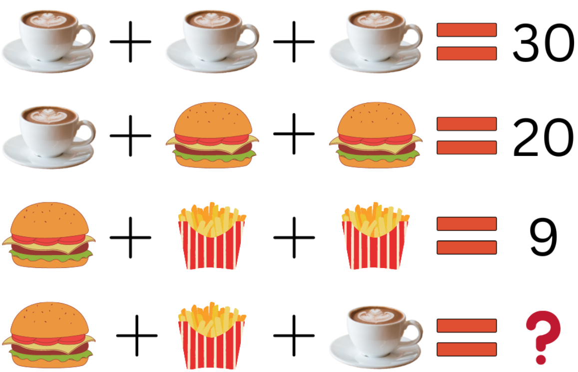
X number of dogs, Y number of chickens

$$X + Y = 36$$
$$4X + 2Y = 100.$$

Solve the equation we get $X = 14, Y = 22$.



Internet Example



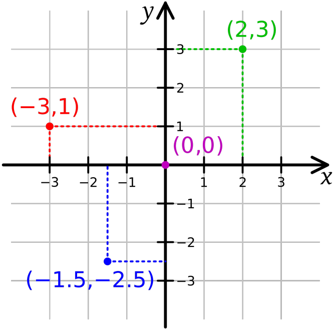
Brief History of Linear Algebra

- Generally, we need to solve a system of linear equations which has some variables.
- The Babylon knew how to solve 2×2 system of linear equations with two unknowns.
- Around 200 BC, the Chinese showed they could solve 3×3 system of linear equations.
- The equation $ax + b = c$ was worked on by people from all walks of life.
- Leibnitz in 17th century introduced the concept of determinant.
- Cramer presented his ideas to solve system of linear equations based on determinants.

Brief History of Linear Algebra

- Euler brought to light that a system of linear equations does not have to have a solution.
- Gauss introduced elimination method to solve system of linear equations.
- In 1848, Sylvester introduced the term “matrix".
- Cayley defined the matrix multiplication.
- With the development of computer, the matrix calculations were speed up.
- Story of Nobel laureate, who needed to compute inverse matrix of 25 by 25 matrix. He had to access to super-computer at that time.

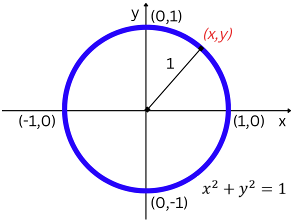
Rene Dercastes in 1637 introduced the coordinate to represent points in plane, which is now called Cartesian coordinates.

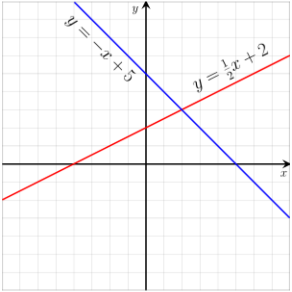


Source: Wikipedia

Why is the Cartesian coordinate important?

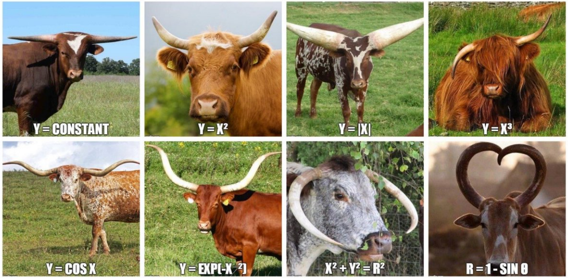
- Points, Lines, Shapes, Objects could be represented as set of numbers.
- Concepts in geometry could be represented as a set of numbers.





Source:Wikipedia

- Dual view of the object:
  - Geometric view helps to image and visualise objects.
  - Algebraic view helps to do calculations.
  - Imagine of high dimensional space.



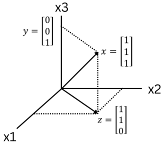
Source: Internet

Problems with Cartesian Coordinates

- The Cartesian coordinates need an origin.
- It also needs a (orthogonal) system of basis vectors/directions for coordinates
- When we change the origin or change the system of vectors, the coordinates also change.
- However, the object is still the same.
- Questions:
  - How do we describe/formularize of those changes, when we change the basis vectors?
  - How we formularize when the object changes?

Vector space

- Given an origin O, for each point A, we view it as a vector $\vec{OA}$.
- The whole space could be considered as a vector space.
- For example



- Addition, subtraction between vectors could be viewed as addition and subtraction on coordinates.
- However, we need a set of basis vectors to span the whole space.

- In vector space, an origin is fixed, every point in the space is determined by a vector from the origin to that point.
- Addition, subtraction between vectors and multiplication a vector with a number will result in a vector.
- There is no coordinate.
- However, we could put coordinate there by viewing the space through the lense of Cartesian coordinates.
- We set up a system of vectors such that the linear combination of them form the whole vector space.
- The system that has minimal number of vectors are called the basis vectors.
- When the whole space is $\mathbb{R}^n$, then the number of basis vectors equals n.

Examples of Vector Space

- Line: $\mathbb{R}$
- Plane: $\mathbb{R}^2$
- The set of polynomials with real coefficients

$$p(x) = a_0 + a_1 x + \ldots + a_n x^n$$

  is also a vector space. Since sum of two polynomials is also a polynomial. Product of polynomial with a real number is also a polynomial.
- Set of continuous functions $f : \mathbb{R} \to \mathbb{R}$ is also a vector space.

Basic concepts

- Linear combination: given vector $v_1, \ldots, v_n$, the linear combination of $v_1, \ldots, v_n$ are the set

$$a_1 v_1 + a_2 v_2 + \ldots + a_n v_n,$$

$a_1, \ldots, a_n$ are the coefficients (scalar) which are often $\mathbb{R}$ or $\mathbb{C}$.

- Linearly dependent vectors $v_1, \ldots, v_k$: there exists number a_i such that some of $a_i \neq 0$ and

$$a_1 v_1 + \ldots + a_k v_k = \vec{\mathbf{0}}.$$

Example: A, B and C are three points, with center of gravity G, then $\vec{GA}$, $\vec{GB}$ and $\vec{GC}$ are dependent vectors.

$$\vec{GA} + \vec{GB} + \vec{GC} = \vec{\mathbf{0}}.$$

Basic concepts

- Linearly independent vectors $v_1, \ldots, v_k$: for all a_i with some $a_i \neq 0$, then

$$a_1 v_1 + \ldots + a_k v_k \neq \vec{\mathbf{0}}.$$

- Dimension of a vector space $\mathcal{V}$: the minimal number of vectors such that their linear combination is the vector space $\mathcal{V}$.
- Given vector space $\mathcal{V}$ and $v_1, \ldots, v_n$ are linearly independent and their linear combination forms $\mathcal{V}$, then the set $v_1, \ldots, v_n$ is called a system of basis vectors.
- There are many tuples of basis vectors.
- Example: the vector space of polynomial, we could choose the basis vectors are

$$1, 2x - 1, 3x^2 - 2x, \ldots$$

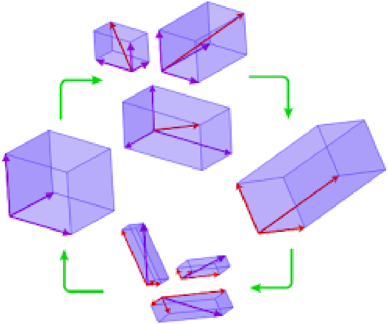
Some basic results of vector space $\mathcal{V}$:

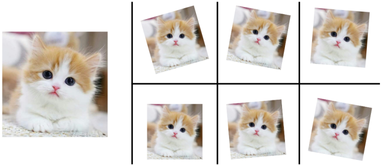
- The number of linearly independent vectors in $\mathcal{V}$ cannot exceed the dimension of $\mathcal{V}$.
- If there are more vectors than the dimension of $\mathcal{V}$ then those vectors are linearly dependent.
- Any vector in $\mathcal{V}$ could be written as the linear combination of basis vectors of $\mathcal{V}$.
- If some vectors are pairwise linearly independent, then it does not mean that they are linearly independent.

Vector Space

- The set of continuous functions on $\mathbb{R}$ is a vector space. The dimension of that space is infinite.
- There are many other cases of vector space in which their space's dimension is infinite.
- $\mathbb{R}^n$ is one of the most popular vector space.
- We often approximate or view other space as $\mathbb{R}^n$.
- In reality, we are only able to see the 3 dimensional space. For higher dimensional space, we could only imagine abstractly.

Example of linear maps





Enlarge your Dataset

Source: Internet

Definition (Linear map)

L is a linear map from vector space $\mathcal{V}$ to $\mathcal{W}$ when

$$L(a\vec{u} + b\vec{v}) = aL(\vec{u}) + bL(\vec{v})$$

for $\vec{u}$ and $\vec{v}$ are two vectors in $\mathcal{V}$, a and b are scalar (real number, complex number etc).

- Example: Linear map $L : \mathbb{R}^2 \to \mathbb{R}^2$ with

$$L([x, y]) = x \times L([1, 0]) + y \times L([0, 1]).$$

- Example: Non-linear map $L : \mathbb{R}^2 \to \mathbb{R}^2$

$$L([x, y]) = [2x + 3y, x - y + 1].$$

Matrix

A rectangle of size $m \times n$ of numbers, which is a linear map from n dimensional space to m dimensional space.

Example:

$$L = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} \quad e_1 = [1, 0], \quad e_2 = [0, 1]$$

$$L(e_1) = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

$$L(e_2) = \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

Matrix

- From the previous example, the map L takes unit vectors e_1 and e_2 map to vectors $(2, 1)$ and $(2, 3)$, which are the columns of matrix L.
- For any vector $v = (x, y)$, v could be written as $x \times (1, 0) + y \times (0, 1)$.
- Then

$$\begin{aligned} L(x, y) &= x \cdot (2, 1) + y \cdot (2, 3) = (2x + 2y, x + 3y) \\ &= \begin{pmatrix} 2 & 2 \\ 1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2x + 2y \\ x + 3y \end{pmatrix}. \end{aligned}$$

When L is a $m \times n$ matrix, $L : \mathbb{R}^n \to \mathbb{R}^m$.

Rank and Kernel of Linear Map

Some definitions/concepts for linear map $L : \mathcal{V} \to \mathcal{W}$

Image of L

Image of L is the set $L(\mathcal{V}) \subset \mathcal{W}$.

Properties: $L(\mathcal{V})$ is a vector space.

Rank of L

The dimension of $L(\mathcal{V})$ is also the rank of L.

Properties: rank of L is the number of independent vectors/columns of L.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}$$

Kernel of L

The kernel of L, denoted by $Ker(L)$, is defined as

$$Ker(L) = \{v \in V : \quad L(v) = 0\}$$

Proposition

$Ker(L)$ is also a vector space.

The proof is easy, assume that $v_1, v_2 \in Ker(L)$,

$$L(v_1) = L(v_2) = \mathbf{0}$$
$$L(a_1 v_1 + a_2 v_2) = a_1 L(v_1) + a_2 L(v_2) = \mathbf{0} \Rightarrow a_1 v_1 + a_2 v_2 \in Ker(L).$$

Kernel of Linear Map (continued)

To solve equation $Lx = \mathbf{0}$, to obtain $Ker(L)$.

$$L = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}; \quad v_1 = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} \quad v_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} \quad v_3 = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$$

$$Lx = x_1 v_1 + x_2 v_2 + x_3 v_3 = \mathbf{0}$$

We know that $v_1 - 2v_2 + v_3 = \mathbf{0}$, and v_1 and v_3 are independent. Hence

$$Ker(L) = \left\{ c(1, -2, 1) \middle| c \in \mathbb{R} \right\}.$$

Kernel of Linear Map (continued)

$$L = \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix}; \quad v_1 = \begin{pmatrix} 1 \\ 4 \\ 7 \end{pmatrix} \quad v_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix} \quad v_3 = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$$

$$v_1 - 2v_2 + v_3 = \mathbf{0}.$$

the column v_2 could be represented as linear combination of v_1 and v_3, v_2 could be eliminated to simplify the linear map L.

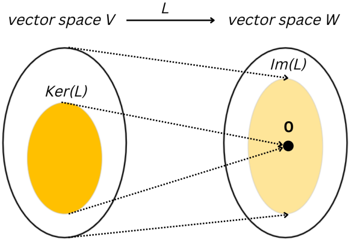
- In reality, some map could be approximated by linear map.
- Dimension reduction is an important area in statistics and machine learning.

Rank and Nullity theorem

Let L be a linear map from $\mathcal{V}$ to $\mathcal{W}$ with the dimension of $\mathcal{V}$ is finite, then

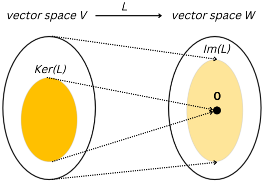
$$Rank(L) + Nullity(L) = Dim(\mathcal{V}).$$

$Rank(L)$ is the dimension of $L(\mathcal{V})$, $Nullity(L)$ is the dimension of $Ker(L)$ and $Dim(\mathcal{V})$ is the dimension of $\mathcal{V}$.



Source: Internet

Rank and Nullity Theorem



Intuition of this theorem: the vector space $\mathcal{V}$ could be decomposed as the sum of two vector spaces,

- Vector space $\mathcal{V}$ is mapped to a subspace of $\mathcal{W}$.
- The other vector space is mapped to zero under map L.
- Those two vector spaces have only one common vector which is zero. Hence, the sum of their dimensions is the dimension of the sum. That proves the theorem.

Proof of Rank and Nullity Theorem

- Assume $dim(Ker(L)) = k$ and $dim(\mathcal{V}) = k + n$.
- Let $v_1, \ldots, v_k$ be a basis of $Ker(L)$.
- We extend the basis $\{v_1, \ldots, v_k\}$ to $\{v_1, \ldots, v_k, v_{k+1}, \ldots, v_{k+n}\}$ such that the linear combination of $v_1, \ldots, v_{k+n}$ is the whole space $\mathcal{V}$.
- We need to show the dimension of $Image(L)$ is equal to n.
- We first show that $L(v_{k+1}), \ldots, L(v_{k+n})$ are linear independent. If the contrary, then there exist some a_i's such that

$$
\begin{aligned}
&a_{k+1}L(v_{k+1}) + \ldots + a_{n+k}L(v_{k+n}) = \mathbf{0} \\
\Rightarrow &L(a_{k+1}v_{k+1} + \ldots + a_{k+n}v_{k+n}) = \mathbf{0} \\
\Rightarrow &u := a_{k+1}v_{k+1} + \ldots + a_{k+n}v_{k+n} \in Ker(L).
\end{aligned}
$$

Proof of Rank and Nullity Theorem

- Since vector $u \in Ker(L)$, u could be written as linear combination of vectors $v_1, \ldots, v_k$. Contradiction to the linear independency assumption between $v_1, v_2, \ldots, v_{k+n}$.
- For any vector $v \in V$, there exist a_i

$$
\begin{aligned}
v &= \underbrace{a_1 v_1 + \ldots + a_k v_k}_{s} + \underbrace{a_{k+1} v_{k+1} + \ldots + a_{k+n} v_{k+n}}_{u} \\
&= s + u.
\end{aligned}
$$

- Let $\mathcal{U}$ be the vector space spanned by $v_{k+1}, \ldots, v_{k+n}$, $L(s+u) = L(u) \Rightarrow L(\mathcal{U}) = Image(L)$.
- $L(v_{k+1}, \ldots, L(v_{k+n})$ are linear independent and $L(\mathcal{U}) = Image(L)$. Dimension of $Image(L)$ is equal to n.

Composition Map

- Let $L_1 : V_0 \to V_1$ and $L_2 : V_1 \to V_2$ be a linear maps between two vector spaces. Then $L_2 \circ L_1 : V_0 \to V_2$.
- Assume that the dimension of V_2, V_1 and V_0 are equal to k, m and n respectively.
- In the language of matrix, let M_1 be the matrix associated with L_1 and M_2 be a matrix of L_2.
- Then size of M_1 is $m \times n$, then size of M_2 must be $k \times m$.
- Multiplying two matrices is equivalent to find the matrix of decomposition map $L_2 \circ L_1$.
- The product of two matrices is not commutative, because the composition map might not be commutative.

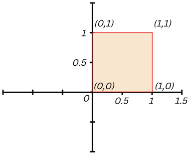
Inverse Map

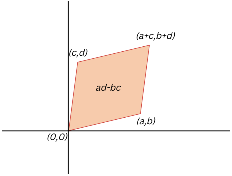
- Given L is a map from $\mathbb{R}^n \to \mathbb{R}^n$.
- Assume that the kernel of L is zero. It means that L is one-to-one map.
  Proof: If $L(v_1) = L(v_2)$ then $L(v_1 - v_2) = \mathbf{0}$. Since the kernel of L is zero, $v_1 - v_2 = \mathbf{0}$.
- We now could define the inverse map $L^{-1} : \mathbb{R}^n \to \mathbb{R}^n$.
- The identity map I is the product of LL^{-1}.
- In the matrix form $I = MM^{-1} = M^{-1}M$, where M is a matrix associated with L.
- It also could explain that $(M_1M_2)^{-1} = M_2^{-1}M_1^{-1}$, because it is an inverse map.

A linear map is a way to rescale and change direction of vector space. It is also a way to transform an object to another object linearly in some directions.

EXAMPLE: A linear map transforms square to a parallelogram.

$$M = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$$





To understand/imagine a linear map, follow its geometric transformation.

EXAMPLE 1 If $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$, then the linear transformation $\mathbf{x} \mapsto A\mathbf{x}$ maps the unit sphere $\{\mathbf{x} : \|\mathbf{x}\| = 1\}$ in $\mathbb{R}^3$ onto an ellipse in $\mathbb{R}^2$, shown in Fig. 1. Find a unit vector $\mathbf{x}$ at which the length $\|A\mathbf{x}\|$ is maximized, and compute this maximum length.

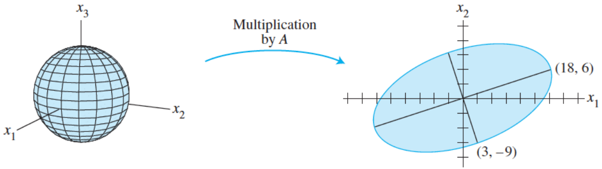


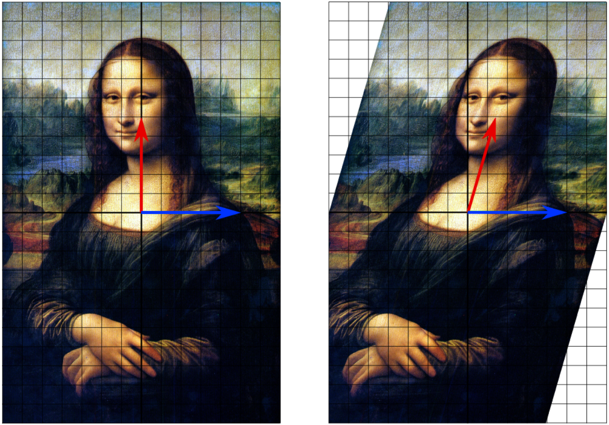
FIGURE 1 A transformation from $\mathbb{R}^3$ to $\mathbb{R}^2$.

Source: Internet

Some Questions about Rank, Nullity and Dimension

- What are the rank of $xy^\top$ and $I + xx^\top$, where x, y are $n \times 1$ vectors?
- If $T : \mathbb{R}^n \to \mathbb{R}^m$, could $rank(T)$ be greater than n or m?
- Prove that $rank(AB) \leq \min\{rank(A), rank(B)\}$.
- If A is linear map from $\mathbb{R}^3$ to $\mathbb{R}^3$. Can both rank and dimension of kernel of A equal 2 at the same time?
- Prove that $rank(A) + rank(B) \geq rank(A + B)$. When does the equality happen?
- Prove that $Ker(T_1) \cap Ker(T_2) \subseteq Ker(T_1 + T_2)$, give example when the equality does not happen?
- Prove or disprove $Ker(T) = Ker(T^2)$ and $Image(T) = Image(T^2)$?

Eigvenvectors and Eigenvalues



Source: Wikipedia.

Eigvenvectors and Eigenvalues

- For a linear map L, find vector v such that

$$Lv \text{ is parallel to } v.$$

- When v is an unit vector with that property, v is called eigenvector.
- It is important to find v, since following v, the operation L is simple.
- The ratio between Lv and v is called the corresponding eigenvalue.
- A pair of eigenvalue and eigenvector (λ, e) of linear map L,

$$Le = \lambda e.$$

Let L be a linear map

- Assume that $e_1, \ldots, e_k$ are eigenvectors of L with the same eigenvalue λ, then every unit vector v of the linear space spanned by $e_1, \ldots, e_k$ is also an eigenvector of L.

  Proof:

$$\begin{aligned} v &= \alpha_1 e_1 + \ldots + \alpha_k e_k \\ \Rightarrow L(v) &= L(\alpha_1 e_1 + \ldots + \alpha_k e_k) \\ &= \lambda\alpha_1 e_1 + \ldots + \lambda\alpha_k e_k \\ &= \lambda v. \end{aligned}$$

  v is an eigenvector of L.

Properties of Eigenvalues and Eigenvectors

- Assume that $(\lambda_1, e_1), (\lambda_2, e_2), \ldots (\lambda_k, e_k)$ are pairs of eigenvalue and eigenvector of L. If the λ_i are non-zero distinct numbers, then $e_1, \ldots, e_k$ are linear independent.

  Proof:

  Assume that $e_1, \ldots, e_{k-1}$ are linearly independent and there exist α_i (some are non-zeros) such that

$$
\begin{aligned}
&\alpha_1 e_1 + \ldots + \alpha_{k-1} e_{k-1} = e_k \\
&L(\alpha_1 e_1 + \ldots + \alpha_{k-1} e_{k-1}) = \lambda_1 \alpha_1 e_1 + \ldots + \lambda_{k-1} \alpha_{k-1} e_{k-1} \\
&L(e_k) = \lambda_k e_k \\
\Rightarrow &\lambda_1 \alpha_1 e_1 + \ldots \lambda_{k-1} \alpha_{k-1} e_{k-1} = \lambda_k e_k = \lambda_k \alpha_1 e_1 + \ldots + \lambda_k \alpha_{k-1} e_{k-1} \\
\Rightarrow &(\lambda_1 \alpha_1 - \lambda_k \alpha_1) e_1 + \ldots + (\lambda_{k-1} \alpha_{k-1} - \lambda_k \alpha_{k-1}) e_{k-1} = \mathbf{0}.
\end{aligned}
$$

  $\Rightarrow e_1, \ldots, e_{k-1}$ are not linear independent. Contradiction.

Properties of Eigenvectors and Eigenvalues

Let $\mathcal{V}$ be a vector space of dimension n. Assume (λ_i) be eigenvalue of a linear map L with multiplicity n_k such that

$$n_1 + \ldots + n_k = n.$$

Then vector space $\mathcal{V}$ could be represented as

$$\mathcal{V} = V_1 \oplus V_2 \oplus \ldots \oplus V_k$$

where V_i is the vector space spanned by eigenvectors which have eigenvalues λ_i, and the V_i are linearly independent.
Note: $\oplus$ means $v \in \mathcal{V}$, then $v = v_1 + \ldots + v_k$, $v_1 \in V_1, \ldots, v_k \in V_k$.

Multiplicity of eigenvalues

A $n \times n$ matrix T has no more than n eigenvalues.

PROOF: By the fundamental theorem of algebra, the determinant of

$$T - \lambda I$$

is a polynomial degree n, thus it has no more than n real solutions. That means there are no more than n real eigenvalues.

- Using the fundamental theorem of algebra.
- Based on topology concept of continuity.
- It is not algebra.

ALGEBRAIC PROOF:

Let V_i be the vector space corresponding to eigenvalues λ_i for $i = 1, 2, \ldots, k$. The dimension of V_i is equal to n_i, n_i is the multiplicity of λ_i.

- We consider $V = V_1 \oplus V_2 \oplus \ldots \oplus V_k$, we need to prove that $\sum_{i=1}^{k} n_i$ is also the dimension of V.
- It is equivalent to show that if $v_1 \in V_1, v_2 \in V_2, \ldots, v_k \in V_k$, then $v_1, v_2, \ldots, v_k$ are independent.
- Assume the contrary, Without Loss of Generality (WLOG) there exist minimal t and $a_i \neq 0; \quad 1 \leq i \leq t$, such that

$$a_1 v_1 + \ldots + a_t v_t = \mathbf{0}$$
$$L(a_1 v_1 + \ldots + a_t v_t) = \lambda_1 a_1 v_1 + \ldots + \lambda_t a_t v_t = \mathbf{0}.$$

ALGEBRAIC PROOF:

- From the previous slide,

$$\lambda_1 a_1 v_1 + \lambda_1 a_2 v_2 + \ldots + \lambda_1 a_t v_t = \mathbf{0}$$
$$\lambda_1 a_1 v_1 + \lambda_2 a_2 v_2 + \ldots + \lambda_t a_t v_t = \mathbf{0}.$$

- Taking the difference,

$$(\lambda_1 - \lambda_2) a_2 v_2 + \ldots + (\lambda_1 - \lambda_t) a_t v_t = \mathbf{0}.$$

Contradiction.

- Thus, $V_1 \oplus V_2 \oplus \ldots \oplus V_k$ has dimension $n_1 + \ldots + n_k$ and it is a subspace of V.
- Hence, $n \geq \sum_{i=1}^{k} n_i$.

- Let A be $n \times n$ invertible matrix with eigenvalues $\{\lambda_i, i \in [1, n]\}$.What are the eigenvalues of A^{-1}?
- Find the eigenvalues and eigenvectors of $I + xx^\top$, where x is $n \times 1$ vectors.
- Assume that A is $m \times n$ matrix and B is a $n \times m$ matrix. Prove that AB and BA have the same non-zero eigenvalues.
- Let A be an upper-triangular matrix. Find all eigenvalues of A.

$$A = \begin{pmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{pmatrix}.$$

- Assume $A \approx B$, could we deduce that $\lambda_i(A) \approx \lambda_i(B)$ and $e_i(A) \approx e_i(B)$, where $\lambda_i(X), e_i(X)$ are the i^{th} eigenvalue and eigenvector of X ?

Matrix and Diagonalisability

- When multiplying two matrices, the more zero-entries, the less computational cost is.
- When two matrices are square matrices with non-zeros are only in the diagonal. Their product is the matrix in which its main diagonal is the product of two diagonals.
- A matrix A is diagonalisable if $A = PDP^{-1}$.
- Here P is invertible and D is diagonal.

EXAMPLE

$$A = \begin{pmatrix} 4 & -1 \\ 6 & -1 \end{pmatrix} \Rightarrow D = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}; \quad P = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Important properties

- $A^n = PD^nP^{-1}$

$$A^n = (PDP^{-1}) \times (PDP^{-1}) \times \ldots \times (PDP^{-1}) = PD^nD^{-1}.$$

- P is a matrix of eigenvectors of A which their eigenvalues are the entries on the diagonal of D.

$$P^{-1}AP = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix} \quad \Rightarrow \quad AP = P \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$

$$\Rightarrow AP_i = \lambda_i P_i$$

P_i is the ith column of P.

Matrix Multiplication

Given matrices A and B, let $D = AB$. What is the formula for D?

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{pmatrix}; \qquad B = \begin{pmatrix} b_{11} & b_{12} & \dots & b_{1p} \\ b_{21} & b_{22} & \dots & b_{2p} \\ \dots & \dots & \ddots & \dots \\ b_{n1} & b_{m2} & \dots & b_{np} \end{pmatrix}$$

$$D = AB; \quad d_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

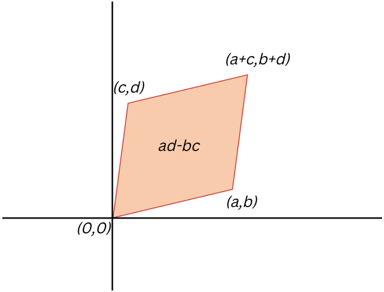
If there are three matrices A,B and C, let $D = ABC$, then

$$d_{ij} = \sum_{k,\ell} a_{ik} b_{k\ell} c_{\ell j}.$$

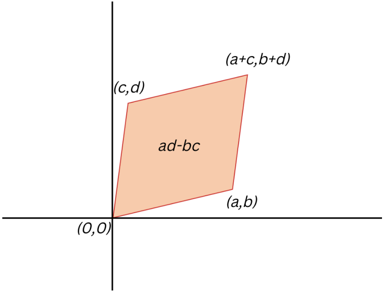
We could use it to verify $(AB)^\top = B^\top A^\top$.

Determinant

- Recall: A linear map is a way to transform a system of basis vectors to a system of vectors.
- The coordinates of new systems are recorded in the columns of the matrix.
- In other way, columns of 2 by 2 matrix form a parallelogram in $\mathbb{R}^2$.
- In high dimensional space, the columns form a parallelotope.



Determinant



- Recall: a linear map transforms a square to a parallelogram, the volume of the parallelogram is the determinant of the matrix.
- The determinant is a real number, it could be negative, since it depends on the direction of the parallelogram / the order of column vectors

$$\det \begin{pmatrix} a & c \\ b & d \end{pmatrix} = ad - bc; \qquad \det \begin{pmatrix} c & a \\ d & b \end{pmatrix} = bc - ad.$$

Compute Determinant

- Leibniz's formulas

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \ddots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}; \qquad \det(A) = \sum_{\sigma} \left[\text{sign}(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \right].$$

- σ is a permutation of $(1, 2, \dots, n)$, that means $\sigma(1), \dots, \sigma(n)$ is a way to rearrange the set $(1, 2, \dots, n)$.
- $\text{sign}(\sigma) = 1$, if we could swap two numbers for an even times to obtain $\sigma(1), \sigma(2), \dots, \sigma(n)$ from $(1, 2, \dots, n)$.
- $\text{sign}(\sigma) = -1$ for otherwise.

EXAMPLE:

$$
\begin{aligned}
n = 4: \quad & \sigma(1) = 2, \sigma(2) = 1, \sigma(3) = 3, \sigma(4) = 4. \\
& (1,2,3,4) \longrightarrow (2,1,3,4) \Rightarrow \text{sign}(\sigma) = -1. \\
n = 3: \quad & \sigma(1) = 2, \sigma(2) = 3, \sigma(3) = 1 \\
& (1,2,3) \longrightarrow (2,3,1) \Rightarrow \text{sign}(\sigma) = 1.
\end{aligned}
$$

Determinant of the 2×2 matrix

$$
\begin{aligned}
\det \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} &= \text{sign}(1,2) a_{11} a_{22} + \text{sign}(2,1) a_{12} a_{21} \\
&= a_{11} a_{22} - a_{12} a_{21}.
\end{aligned}
$$

Some Rules to Compute Determinant

Explain the following rules by geometry

- $\det(I_n) = 1$.
- $\det(A^{-1}) = \det(A)^{-1}$.
- $\det(AB) = \det(A)\det(B)$.
- $\det(cA) = c^n\det(A)$.
- Three matrices A, B and C have all columns the same except the ith columns.The ith column of C equals the sum of ith columns of A and B. Then $\det(C) = \det(A) + \det(B)$.
- If any row or column of A is zero vector, then $\det(A) = 0$.
- If the columns of A are dependent, then $\det(A) = 0$.

Property $\det(A) = \det(A^\top)$ is difficult to explain by geometry.

Questions about Eigenvalues and Determinant

- Eigenvalues of matrix T are the solutions of $\det(\lambda I - T) = 0$.
- Given $\det(A) = \det(A^\top)$, prove that eigenvalues of A are also eigenvalues of $A^\top$. Prove that $\det(\lambda I - A) = \det(\lambda I - A^\top)$.
- Prove that $B^{-1}AB$ has the same eigenvalues as A.
- Let λ_i be eigenvalues of A, calculate the determinant of $I + A$.

Spectral Decomposition

Spectral decomposition

If A is a $n \times n$ symmetric matrix, then $A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$

$$A = U \Lambda U^\top$$

- The u_i are $n \times 1$ eigenvectors of A, Λ is the diagonal matrix of λ_i, U is the unitary/orthogonal matrix of $\left[u_i\right]_{i=1}^{n}$.
- The u_i are orthonormal

$$u_i^\top u_j = \begin{cases} 0 & i \neq j \\ 1 & i = j. \end{cases}$$

- $U^\top U = I = UU^\top$ or $U^\top = U^{-1}$.
- λ_i are real numbers.

Spectral Decomposition

Spectral decomposition is a special case of diagonalisable matrix:

$$U\Lambda U^\top = U\Lambda U^{-1}.$$

The main differences:

- Columns of U are orthogonal vectors with length 1.
- Diagonalisable matrix is not necessary symmetric matrix.
- Any matrix of the form $U\Lambda U^\top$ for Λ is a symmetric matrix.

Spectral Decomposition

We prove the last statement about $U\Lambda U^\top$

$$U = \begin{pmatrix} u_{11} & \dots & u_{1n} \\ \dots & \dots & \dots \\ u_{n1} & \dots & u_{nn} \end{pmatrix}; \quad \Lambda = \begin{pmatrix} \lambda_1 & \dots & 0 \\ \dots & \dots & \dots \\ 0 & \dots & \lambda_n \end{pmatrix}; \quad U^\top = \begin{pmatrix} u_{11} & \dots & u_{n1} \\ \dots & \dots & \dots \\ u_{1n} & \dots & u_{nn} \end{pmatrix}$$

$U = [u_{ij}]; \qquad \Lambda = diag(\lambda_i); \qquad U^\top = [u_{ji}]$

Now let $A = \left[a_{ij}\right] = U\Lambda U^\top$, we do the multiplication

$$a_{ij} = \sum_{k=1}^{n} u_{ik}\lambda_k u_{jk}$$

$$a_{ji} = \sum_{k=1}^{n} u_{jk}\lambda_k u_{ik}$$

Thus $a_{ij} = a_{ji}$.

Sketch of Proof of Spectral Decomposition Theorem

A is a symmetric matrix

- All eigenvalues are real numbers:
  If $Av = \lambda v$, then λ is real number
- Eigenvectors with different eigenvalues are orthogonal.
- If $\mathcal{W}$ is stable under A, then so is $\mathcal{W}^{\perp}$, i.e.

$$A(\mathcal{W}) \subset \mathcal{W} \Rightarrow A(\mathcal{W}^{\perp}) \subset \mathcal{W}^{\perp}.$$

- A on $\mathcal{W}$ and $\mathcal{W}^{\perp}$ are symmetric linear transformation.
  The restriction of A on $\mathcal{W}$ and $\mathcal{W}^{\perp}$ is symmetric.
- A accepts a linear representation on its eigenvectors and eigenvalues.

Proof of Spectral Decomposition Theorem

Step 1

Prove that all eigenvalues are real numbers.

Let v be an eigenvector of A: $v \in \mathbb{C}^n$ and $\overline{v}$ is the conjugate of v. Then

$$\begin{aligned}\overline{\overline{v}^\top A v} &= v^\top \overline{A}\overline{v} = v^\top A \overline{v} \\ &= (v^\top A \overline{v})^\top = \overline{v}^\top A^\top v = \overline{v}^\top A v\end{aligned}$$

since A is symmetric, $A = A^\top$. The LHS = RHS, then they are all real numbers.

Let (λ, v) be a pair of eigenvalue and eigenvector:

$$\overline{v}^\top A v = \overline{v}^\top \lambda v = \lambda \overline{v}^\top v = \lambda \|v\|^2 \in \mathbb{R}.$$

Then $\lambda \in \mathbb{R}$.

Step 2

Prove that eigenvectors with different corresponding eigenvalues are orthogonal.

Let (λ, v) and (μ, u) be pairs of eigenvalue and eigenvector:

$$u^\top Av = u^\top \lambda v = \lambda u^\top v$$
$$v^\top Au = v^\top \mu u = \mu v^\top u.$$

However $u^\top Av = (u^\top Av)^\top = v^\top A^\top (u^\top)^\top = v^\top Au$. It follows that

$$\mu v^\top u = \lambda u^\top v \Rightarrow u^\top v = 0 \Rightarrow u \perp v.$$

Step 3

If $A \in \mathbb{R}^{n\times n}$ is symmetric and W is a subspace of $\mathbb{R}^n$ such that $A(W) \subset W$, then $A(W^\perp) \subset W^\perp$.

Let $x \in W$ and $y \in W^\perp$, then

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

Since $Ax \in W$, $y^\top Ax = 0$, then $x^\top Ay = 0$ for all x. Therefore $Ay \in W^\perp$.

STEP 4

If $\mathcal{W}^{\perp}$ is a subspace and $A(\mathcal{W}^{\perp}) \subset \mathcal{W}^{\perp}$ then $\mathcal{W}^{\perp}$ contains an eigenvector of A.

Choose an orthogonal basis $u_1, \ldots, u_m$ of $\mathcal{W}^{\perp}$, because $Au_i \in \mathcal{W}^{\perp}$. Denote

$$r_{ij} = u_i^{\top} A u_j \Rightarrow r_{ij} = r_{ji}$$

r_{ij} is the coefficient of Au_j on the direction u_i thus

$$Au_j = \sum_{i=1}^{m} r_{ij} u_i$$

The matrix $R = (r_{ij})$ is symmetric. R is a linear map in the orthonormal system of $u_1, \ldots, u_m$

Proof of Spectral Decomposition Theorem

STEP 4 CONTINUED

Since R is symmetric matrix under the orthogonal system $(u_1, \ldots, u_m)$, there exists a pair (λ, w) is the eigenvalues and eigenvector of R in the space $\mathcal{W}$.

Assume the coordinates of w under the orthonormal system $(u_1, \ldots, u_m)$ is $(x_1, \ldots, x_m)$.

- Then
$$w = \sum_{j=1}^{m} x_j u_j.$$
- $Rx = \lambda x$, where $x = (x_1, \ldots, x_m)$
$$\sum_{j=1}^{m} r_{ij} x_j = \lambda x_i.$$

STEP 4 CONTINUED

$$
\begin{aligned}
A(w) = \sum_{j=1}^{m} x_j A(u_j) &= \sum_{j=1}^{m} x_j \sum_{i=1}^{m} r_{ij} u_i \\
&= \sum_{i=1}^{m} \Big(\sum_{j=1}^{m} r_{ij} x_j \Big) u_i \\
&= \sum_{i=1}^{n} \lambda x_i u_i \\
&= \lambda w.
\end{aligned}
$$

It means that w is eigenvector of A.

Step 5

Building the subspace $\mathcal{W}$ by induction.

- We start with the first pair of eigenvalue and eigenvector (λ_1, u_1).
- Set $\mathcal{W}_1 = \mathbb{R}u_1$, then $A(\mathcal{W}_1) \subset \mathcal{W}_1$, then $A(\mathcal{W}_1^\perp) \subset W_1^\perp$.
- Adding eigenvector $u_2 \in \mathcal{W}_1^\perp$ to form $\mathcal{W}_2 = \text{span}(u_1, u_2)$ and so on.
- We obtain a sequence of orthogonal eigenvectors $u_1, \ldots, u_n$ and their corresponding eigenvalues.

The proof is done.

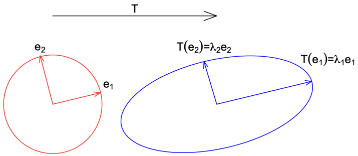
Calculation using Spectral Decomposition

- Let A is a symmetric matrix, then $A = \sum_{i=1}^{n} \lambda_i u_i u_i^\top$, where the (λ_i, u_i) are all pairs of eigenvalues and eigenvectors of A.
- Let v be a vector that has the form $\sum_{i=1}^{n} \alpha_i u_i$ under the orthogonal system (u_i)
- We compute $v^\top A v$

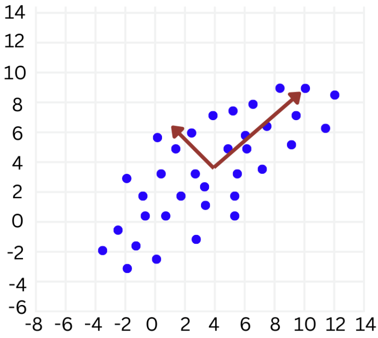
$$
\begin{aligned}
v^\top A v &= (\sum_{i=1}^{n} \alpha_i u_i^\top)(\sum_{i=1}^{n} \lambda_i u_i u_i^\top)(\sum_{i=1}^{n} \alpha_i u_i) \\
&= \sum_{i=1}^{n} \alpha_i^2 \lambda_i .
\end{aligned}
$$

Geometric Representation of Spectral Decomposition

- The spectral decomposition is a way to decompose a symmetric linear map into directions and scales.
- It transforms a circle to ellipse, a ball to ellipsoid.
- The eigenvectors are the directions of the symmetric axes of ellipsoid.
- The eigenvalues are the scales on the ellipsoid.
- Note that the eigenvalues could be negative.



- Principle Components: Given n points $X_1, \ldots, X_n$ in $\mathbb{R}^d$ with $n \geq d$, find the principal components of them



Non-negative, Positive Definite Matrix

- A positive definite matrix is a matrix A that satisfies the condition: $x^\top Ax > 0$ for all $n \times 1$ vector x.
- A semi-positive/non-negative definite matrix is a matrix A that satisfies the condition: $x^\top Ax \geq 0$ for all $n \times 1$ vector x.
- A symmetric matrix with all positive eigenvalues is a positive definite matrix.
- The converse is not true

$$(a, b) \times \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = a^2 + b^2.$$

- A symmetric matrix with a negative eigenvalue is not a positive definite matrix.

Exercises for Spectral Decomposition Theorem

- Are the sum and product of two symmetric matrices symmetric?
- Is $A^\top A$ symmetric?
- Are eigenvalues of $A^\top A$ all non negative?
- Describe the linear map of symmetric matrix A whose eigenvalues are ± 1.
- Find the eigenvalues of

$$A = \begin{pmatrix} a & b & c \\ b & c & a \\ c & a & b \end{pmatrix}.$$

Trace of Matrix

- Trace means "vết"in Vietnamese, something is left after things already happened.
- Recall that matrix is a way to record the linear transformation under a basis. Trace of matrix is something unchanged, under different basis.
- Which means that trace depends only on the linear map, not the coordinates.
- The mathematical definition of trace is

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \langle u_i, Au_i \rangle = \sum_{i=1}^{n} u_i^\top A u_i,$$

  where the u_i is a system of orthonormal vectors.
- When A is square matrix, then trace of A is the sum of main diagonal entries.

Trace of Matrix

Let

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix}; \quad e_i = \begin{pmatrix} 0 \\ \vdots \\ 1_i \\ \vdots \\ 0 \end{pmatrix};$$

$$e_i^\top A e_j = e_i^\top A_j = a_{ij}.$$

Here, A_j is the j^{th} column of A.
Thus a_{ij} is the ith coordinate of vector A_j under the system $\{e_1, \dots, e_n\}$ and

$$A = \left[e_1, \dots, e_n\right]^\top A \left[e_1, \dots, e_n\right].$$

Properties of Trace of Matrix

- $\text{tr}(A) = \text{tr}(A^\top)$.
- $\text{tr}(\alpha A + \beta B) = \alpha\text{tr}(A) + \beta\text{tr}(B)$.
- Let a, b be two $n \times 1$ vectors,
$$\text{tr}(ab^\top) = a^\top b = \langle a, b \rangle.$$
- Let $X = [x_{ij}]$ and $Y = [y_{jk}]$ be two matrices of size $m \times n$ and $n \times m$, then
$$\text{tr}(XY) = \text{tr}(YX).$$
- tr is invariant under orthonormal system.

- Proof of $\mathrm{tr}(XY) = \mathrm{tr}(YX)$:
  Entries at row i and column i of XY is

$$\sum_{j=1}^{n} x_{ij} y_{ji}$$

Then trace of XY is equal to

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} y_{ji}.$$

Similarly, we have trace of YX is equal to

$$\sum_{j=1}^{n} \sum_{i=1}^{m} y_{ji} x_{ij}.$$

- Proof of the invariant property of trace: Let (u_i) be another orthonormal system and B be the matrix under the orthonormal system (u_i) such that

$$A = \left[u_1, \ldots, u_n\right]^\top B \left[u_1, \ldots, u_n\right]$$
$$\Rightarrow A = U^\top BU$$
$$\Rightarrow UAU^\top = B.$$

For any $n \times n$ matrix X and Y, by formula of product of two matrices

$$\operatorname{tr}(XY) = \operatorname{tr}(YX).$$

Then we have

$$\operatorname{tr}(B) = \operatorname{tr}(UAU^\top) = \operatorname{tr}(AU^\top U) = \operatorname{tr}(AI) = \operatorname{tr}(A).$$

Questions about Trace of Matrix

- If the followings are always true
  - $\text{tr}(AB) = \text{tr}(A)\text{tr}(B)$.
  - For A is symmetric, $trace(A)$ is the sum of all eigenvalues.
- Find all matrix of 2 by 2 A such that: $\text{tr}(A^2) = \left[\text{tr}(A)\right]^2$.
- Let A be 2×2 matrix. Prove that

$$A^2 - \text{tr}(A) \cdot A + \det(A) \cdot I_2 = O_2.$$

- For A and B are two $n \times n$ symmetric positive definite matrices, prove that $\text{tr}(AB) > 0$.

Eigenvalues, Trace and Determinant

Eigenvalues, trace and determinant of a matrix

Let A be a $n \times n$ matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$. Then

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i; \qquad \det(A) = \prod_{i=1}^{n} \lambda_i.$$

$$A - \lambda I = \begin{pmatrix} a_{11} - \lambda & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} - \lambda & \ldots & a_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ a_{n1} & a_{n2} & \ldots & a_{nn} - \lambda \end{pmatrix}$$

$\det(A - \lambda I)$ is a polynomial of λ which has roots $\lambda_1, \ldots, \lambda_n$, then

$$\det(A - \lambda I) = \prod_{i=1}^{n} (\lambda_i - \lambda).$$

Eigenvalues, Trace and Determinant

The determinant of $A - \lambda I$ is a polynomial having the form

$$\det(A - \lambda I) = (-1)^n \lambda^n + \Big(\sum_{i=1}^{n} a_{ii}\Big)(-1)^{n-1}\lambda^{n-1} + \dots$$

by the Leibniz's formula. On the other hand,

$$\begin{aligned}\det(A - \lambda I) &= \prod_{i=1}^{n}(\lambda_i - \lambda)\\ &= (-1)^n \lambda^n + \Big(\sum_{i=1}^{n} \lambda_i\Big)(-1)^{n-1}\lambda^{n-1} + \dots\end{aligned}$$

Comparing the coefficients of λ^{n-1} of both polynomials, we obtain

$$\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i.$$

Set $\lambda = 0$,

$$\det(A) = \prod_{i=1}^{n} \lambda_i.$$

It is an algebraic proof. When A is diagonalizable,

$$A = P\Lambda P^{-1},$$

where the columns of P are the eigenvectors of A, diagonal of Λ consists of the eigenvalues of A.

$$\begin{aligned}\mathrm{tr}(A) &= \mathrm{tr}(P\Lambda P^{-1}) = \mathrm{tr}(\Lambda P^{-1}P) = \mathrm{tr}(\Lambda I) \\ &= \mathrm{tr}(\Lambda) = \sum_{i=1}^{n} \lambda_i.\end{aligned}$$

Eigenvalues, trace and determinant

When A is diagonalisable, $A = P\Lambda P^{-1}$. Let v_i be the i^{th} column of P,

$$Av_i = \lambda_i v_i.$$

The parallelepiped formed by $(v_1, \ldots, v_n)$ is transformed by A becoming a parallelepiped formed by $(\lambda_1 v_1, \ldots, \lambda_n v_n)$.

$$\begin{aligned} \det(A) &= \frac{\text{volume}(\lambda_1 v_1, \ldots, \lambda_n v_n)}{\text{volume}(v_1, \ldots, v_n)} \\ &= \prod_{i=1}^{n} \lambda_i. \end{aligned}$$

It is more geometric.

Some Types of Matrix

- Symmetric matrix: $A = A^\top$.
- Diagonal matrix: Every off-diagonal entry is equal to zero. Product of diagonal matrices is again diagonal matrix.
- Orthogonal matrix: Matrix U of $n \times n$ entries where its columns form an orthonormal basis.

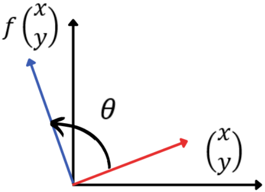
$$\begin{aligned} U(e_i) &= U_i \\ e_i &= (0, \ldots, 1_i, \ldots, 0) \\ U^\top U &= I; \qquad U^\top = U^{-1}. \end{aligned}$$

U_i is the i^{th} column of U.

Rotation and Reflection matrices are two special cases of unitary matrix.

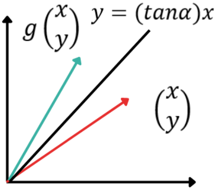
Matrix corresponding to ***rotation***

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$



Matrix corresponding to ***reflection***

$$\begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix}$$



Some Properties of Rotation and Reflection Matrices

- $\text{Rot}(x)\text{Rot}(y) = \text{Rot}(x+y)$.
- $\text{Ref}(x)\text{Ref}(y) = \text{Rot}(2[x-y])$.
- $\text{Rot}(x)\text{Ref}(y) = \text{Ref}(y + \frac{1}{2}x)$.
- $\text{Ref}(x)\text{Rot}(y) = \text{Ref}(x - \frac{1}{2}y)$.
- Rot and Ref preserve the distance between two points.

Some Questions about Orthogonal Matrices

- What is the determinant of an orthogonal matrix?
- What are the possible real eigenvalues and eigenvectors of an orthogonal matrix?
- What is the product of two orthogonal matrices?
- Find all the orthogonal 2×2 matrix.

Singular Value Decomposition

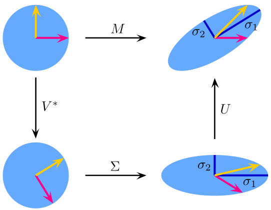
M is a matrix $m \times n$, M could be decomposed as $U\Sigma V^\top$ where U is a unitary matrix of $m \times m$, Σ is a diagonal matrix of $m \times n$ and V is an unitary matrix of $n \times n$. In particular,

$$M = U\Sigma V^\top.$$

- Singular value decomposition: Decompose the linear map by the singular value and unitary matrices.
- For matrix M, we define the singular values of M are the square root of eigenvalues of $MM^\top$ or $M^\top M$.
- Non-diagonal entries of Σ are all equal to zero.

Geometry of Singular Value Decomposition

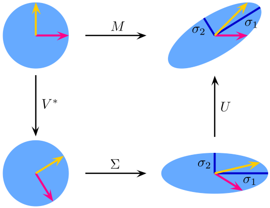
- V transform is like a "rotation/reflection" of the basis.
- Σ is a rescaling transform in the original directions that transforms a circle to ellipse.
- U transform is like another "rotation/reflection" of the basis.
- Two vectors (red and yellow) are two basic vectors.
- A linear map M transforms a circle to an ellipse under a different coordinate system.



$$M = U \cdot \Sigma \cdot V^*$$

Source:Internet

- A linear map M transform a circle to an ellipse under a different coordinate.
- When matrix M is symmetric, the U and V are transpose of each other. It means that we rotate once, we rescale it and then we rotate it back.



$$M = U \cdot \Sigma \cdot V^*$$

Source: Internet

Proof of SVD

- By the spectral decomposition,

$$M^\top M = V\Lambda V^\top = \sum_{i=1}^{\ell} \lambda_i v_i v_i^\top = \sum_{i=1}^{\ell} \sigma_i^2 v_i v_i^\top.$$

- $\sigma_1, \dots, \sigma_\ell$ are all positive, for some $\ell \leq \min\{m, n\}$, which corresponds to $v_1, \dots, v_\ell$, and $\sigma_i^2 = \lambda_i$.
- Denote $V_1 = [v_1, \dots, v_\ell]$.
- We extend the matrix V_1 by adding $n - \ell$ vectors $v_{\ell+1}, \dots, v_n$ such that $v_1, \dots, v_n$ are orthonormal vectors. Denote

$$V_2 = [v_{\ell+1}, \dots, v_n], \qquad V = [V_1, V_2].$$

$$V^\top M^\top M V = \begin{pmatrix} \Lambda & \mathbf{0}_{\ell\times(n-\ell)} \\ \mathbf{0}_{(n-\ell)\times\ell} & \mathbf{0}_{(n-\ell)\times(n-\ell)} \end{pmatrix}.$$

Proof of SVD

From the previous calculations,

$$\begin{bmatrix} V_1 \\ V_2 \end{bmatrix} M^\top M [V_1, V_2] = \begin{bmatrix} V_1^\top M^\top M V_1 & V_1^\top M^\top M V_2 \\ V_2^\top M^\top M V_1 & V_2^\top M^\top M V_2 \end{bmatrix} = \begin{bmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}.$$

- We have $V_2^\top M^\top M V_2 = \|MV_2\|_2^2 = \mathbf{0}$, it means that $MV_2 = \mathbf{0}$.
- We obtain some other equations

$$V_1^\top V_1 = I_\ell$$
$$V_2^\top V_2 = I_{n-\ell}.$$

- For $\sigma_i > 0$, define $u_i = Mv_i/\sigma_i$ for $i = 1, 2, \ldots, \ell$.
- We will prove u_i is an eigenvector of $MM^\top$.

Proof of SVD

- Since u_i is an eigenvector of $MM^\top$,

$$MM^\top u_i = MM^\top M v_i / \sigma_i = M \sigma_i^2 v_i / \sigma_i = \sigma_i^2 u_i.$$

- To check if u_i is an unit vector, we have

$$u_i^\top u_i = \left(\frac{Mv_i}{\sigma_i}\right)^\top \frac{Mv_i}{\sigma_i} = \frac{1}{\sigma_i^2} v_i^\top M^\top M v_i$$
$$= \frac{v_i^\top \sigma_i^2 v_i}{\sigma_i^2} = v_i^\top v_i = 1.$$

- To check u_i is orthogonal to u_j for $i \neq j$

$$u_j^\top u_i = \frac{1}{\sigma_j} v_j M^\top M v_i \frac{1}{\sigma_i} = 0.$$

since v_i, v_j are two different eigenvectors of $M^\top M$.

Proof of SVD

- Let $U_1 = MV_1\Lambda^{-\frac{1}{2}}$, then U_1 is orthogonal matrix of size $m \times \ell$.
- We have

$$I_n = (V_1, V_2)\begin{pmatrix} V_1^\top \\ V_2^\top \end{pmatrix} = V_1 V_1^\top + V_2 V_2^\top$$

$$\begin{aligned} U_1\Lambda^{\frac{1}{2}} V_1^\top &= MV_1\Lambda^{-\frac{1}{2}}\Lambda^{\frac{1}{2}} V_1^\top = MV_1 I_\ell V_1^\top \\ &= M\left(I_n - V_2 V_2^\top\right) = M - MV_2 V_2^\top = M. \end{aligned}$$

  The last equality comes from $MV_2 = \mathbf{0}$.
- We could extend U_1 by adding U_2 of orthonormal vectors to U_1 to form $U = [U_1, U_2]$.

- We build Σ by adding 0 into or removing 0 from the matrix

$$\begin{pmatrix} \Lambda^{\frac{1}{2}} & \mathbf{0}_{\ell\times(n-\ell)} \\ \mathbf{0}_{(n-\ell)\times\ell} & \mathbf{0}_{(n-\ell)\times(n-\ell)} \end{pmatrix}$$

to make it a matrix of $m \times n$.

- For $m \geq n$, M has the following form

$$M = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \Lambda^{\frac{1}{2}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \\ \mathbf{0} \end{bmatrix} \begin{bmatrix} V_1^\top \\ V_2^\top \end{bmatrix} = U_1 \Lambda^{\frac{1}{2}} V_1^\top.$$

- Or we could go in an easier way:

$$M = U_1 \Lambda^{\frac{1}{2}} V_1^\top = \sum_{j=1}^{\ell} \sigma_j u_j v_j^\top.$$

Adding some other $\sigma_{ij} = \sigma_i \mathbf{1}_{i=j}$ to have $\Sigma = \begin{bmatrix} \sigma_{ij} \end{bmatrix}$ we have

$$M = \sum_{i=1}^{m} \sum_{j=1}^{n} \sigma_{ij} u_i v_j^\top.$$

Applications of SVD

Let A be an $m \times n$ matrix which has the SVD: $A = U\Sigma V^\top$ or

$$A = \sum_i \sigma_i u_i v_i^\top$$

For a given rank k, we need to find a matrix of rank k to approximate A.

$$A_k = \arg\min_X \|X - A\|_F^2; \quad \text{subject to } \operatorname{rank}(X) = k.$$

Then

$$A_k = \sum_{i=1}^k \sigma_i u_i v_i^\top.$$

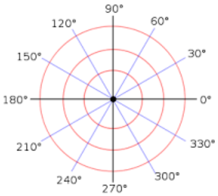
SVD: Exercises

- When do the singular values of a matrix become its eigenvalues?
- Can some singular values of a matrix be equal to zero?
- Given the SVD of M, find the SVD of $M^\top$.
- Prove that the singular values of M and $M^\top$ are the same.
- Prove that the rank of M is equal to the number of non-zero eigenvalues.
- Given the SVD form of M, write a vector x in the convenient form to do the calculation Mx.

Polar decomposition

Given a square matrix A, $A = US$, where U is unitary matrix and S is symmetric non-negative definite matrix.

- The word "polar"means "cực", it is similar to polar coordinate on the sphere,
- or polar representation of complex number $a + b\mathbf{i} = |z|e^{\mathbf{i}\theta}$.



Source: Wikipedia

Polar Decomposition

- $A = US$, U is unitary matrix and S is symmetric positive definite matrix.
- It contains two parts: One is "rotation" or "reflection", other is rescaling along a set of orthogonal basis (eigenvectors).
- Given $A = U\Sigma V^\top$, we could write

$$U\Sigma V^\top = UV^\top(V\Sigma V^\top).$$

  The first part is "rotation", the second part is positive definite matrix.
- The polar decomposition of nonsingular matrix is unique.

VinAI

Lower/Upper Triangular Matrix

Definition

A is upper triangular matrix if $a_{ij} = 0$ for all $i > j$.

Properties

- A is lower triangular matrix if $A^\top$ is upper triangular matrix, it means that $a_{ij} = 0$ for all $i < j$.
- Product of two upper triangular matrices is an upper triangular matrix.
- The same for product of two lower triangular matrices.
- Geometric representation: The first k vectors/columns of matrix A lie in the space spanned by $e_1, \ldots, e_k$, where

$$e_i = [0, \ldots, 1_i, \ldots, 0]^\top.$$

Upper/Lower Triangular Matrix

Solving system of linear equations by using the upper triangular matrix (Gaussian elimination)

$$A = [a_{ij}]; \quad x = [x_1, \dots, x_n]^\top; \qquad b = [b_1, \dots, b_n]^\top$$

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ 0 & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \ldots + a_{2n}x_n &= b_2 \\ &\dots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn}x_n &= b_n. \end{aligned}$$

QR decomposition

Matrix A could be decomposed as product of two matrices: Q and R, Q is an orthogonal matrix and R is an upper triangular matrix, where A is a matrix of $m \times n$, Q is a matrix of $m \times n$ and R is a matrix of $n \times n$.

QR Decomposition

$$\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix} \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}$$

QR Decomposition

$$\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix} \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}$$

- Columns of A are vectors obtained under the transformation A from axis unit vectors.
- Columns of Q $\star$ are the orthonormal vectors.
- Columns of R $\star$ are the coefficients that are used to represent columns of A under the linear combination of columns of Q.
- Important: the entries of the lower part of matrix R are zeros.

QR Decomposition

$$\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} * & * \\ * & * \\ * & * \end{bmatrix} \begin{bmatrix} * & * \\ 0 & * \end{bmatrix}$$

- The first columns of Q and A represent two vectors of the same direction:

$$\frac{1}{\sqrt{2^2+2^2+1^2}}[2,2,1] \to \left[\frac{2}{3}, \frac{2}{3}, \frac{1}{3}\right].$$

- The second vector/column in A is a linear combination of the first and second vectors in Q

$$[3,4,1] = a\left[\frac{2}{3}, \frac{2}{3}, \frac{1}{3}\right] + b\left[x, y, z\right].$$

- The second column of Q has norm 1 and orthogonal to the first column of Q:

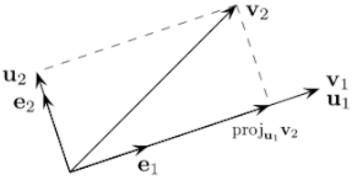
$$a = 3 \times \frac{2}{3} + 4 \times \frac{2}{3} + 1 \times \frac{1}{3} = 5 \Rightarrow b[x, y, z] = \left[\frac{-1}{3}, \frac{2}{3}, \frac{-2}{3}\right]$$
$$\Rightarrow b = 1; \quad [x, y, z] = \left[\frac{-1}{3}, \frac{2}{3}, \frac{-2}{3}\right].$$

- Or the second vector in Q is a linear combination of the first and second vectors in A.
- Similar for the third vector and so on.
- We obtain

$$\begin{bmatrix} 2 & 3 \\ 2 & 4 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} & \frac{-1}{3} \\ \frac{2}{3} & \frac{2}{3} \\ \frac{1}{3} & \frac{-2}{3} \end{bmatrix} \begin{bmatrix} 3 & 5 \\ 0 & 1 \end{bmatrix}.$$

Gram-Schmidt Process

- This process guarantees that the first k vectors of Q forms the same space as the first k vectors of A.
- Hence, the coefficient matrix is upper triangle matrix.
- This process is named Gram-Schmidt process.



$$
\begin{aligned}
\mathbf{u}_1 &= \mathbf{a}_1, & \mathbf{e}_1 &= \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \\
\mathbf{u}_2 &= \mathbf{a}_2 - \mathrm{proj}_{\mathbf{u}_1}\,\mathbf{a}_2, & \mathbf{e}_2 &= \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \\
\mathbf{u}_3 &= \mathbf{a}_3 - \mathrm{proj}_{\mathbf{u}_1}\,\mathbf{a}_3 - \mathrm{proj}_{\mathbf{u}_2}\,\mathbf{a}_3, & \mathbf{e}_3 &= \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} \\
&\vdots & &\vdots \\
\mathbf{u}_k &= \mathbf{a}_k - \sum_{j=1}^{k-1} \mathrm{proj}_{\mathbf{u}_j}\,\mathbf{a}_k, & \mathbf{e}_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}
\end{aligned}
$$

Source: Wikipedia

Matrix Decomposition

Cholesky decomposition

A positive definite matrix A could be decomposed as product of two matrix L and $L^\top$, where L is the lower triangle matrix.

- Since A is symmetric definite, $A = U\Sigma U^\top$.
- Use the QR decomposition for $\Sigma^{1/2}U^\top = QR$, then $A = RQQ^\top R^\top$.
- $A = RIR^\top$, we obtain $A = RR^\top$.
- To solve equation $Ax = b$, rewrite it in the form $LL^\top x = b$. Since L is lower triangle matrix, we could solve the equation easily to find $L^\top x$, then find x.

LU decomposition

A matrix A could be written as the product of L and U, where L is a lower matrix and U is an upper matrix.

LDU decomposition

Matrix $A = LDU$, where D is diagonal matrix, L and U are unitriangle (main diagonal entries are all equal to 1) matrices.

- If $a_{11} = 0$, then l_{11} or u_{11} must equal to zero. If A is not singular, at least one of L or U must be singular. Contradiction.
- They permute A such that the factorisation doable: $PA = LU$.

- Let A be a linear map from $\mathbb{R}^m$ to $\mathbb{R}^n$.
- If $m > n$, then A cannot be injective.
- If $m < n$, then A cannot be surjective,
- In general, if A is not bijective,

How do we define an "inverse"of A from $\mathbb{R}^n$ to $\mathbb{R}^m$?

- Solution: Moore-Penrose inverse.
- The main idea is to decompose the space $\mathbb{R}^m$ and $\mathbb{R}^n$ into sum of linear subspaces that characterises A.

- The first step: Decompose $\mathbb{R}^m$ into sum of subspaces: one is the kernel of A,

$$\mathbb{R}^m = Ker(A) \oplus Ker(A)^\perp.$$

- The second step: Decompose $\mathbb{R}^n$ into sum of subspaces: one is the Image of A,

$$\mathbb{R}^n = Image(A) \oplus Image(A)^\perp.$$

- Note that the map A is bijective between $Ker(A)^\perp$ and $Image(A)$.
- We can define the inverse of A between $Ker(A)^\perp$ and $Image(A)$, then extend the inverse map to $\mathbb{R}^m$ and $\mathbb{R}^n$.
- The map is called Moore-Penrose inverse, denoted by A^+.

Moore-Penrose Inverse

- $A : \mathbb{R}^m \to \mathbb{R}^n$.
- For any $v \in \mathbb{R}^n$, decompose

$$v = v_1 + v_2$$
$$v_1 \in Image(A); \qquad v_2 \in Image(A)^\perp.$$

- Since $v_1 \in Image(A)$, there exists $u \in \mathbb{R}^m$ such that $A(u) = v_1$.
- Decompose $u = u_1 + u_2$ such that $u_1 \in Ker(A)^\perp, u_2 \in Ker(A)$.
- Note: The above decomposition is unique up to u_1.

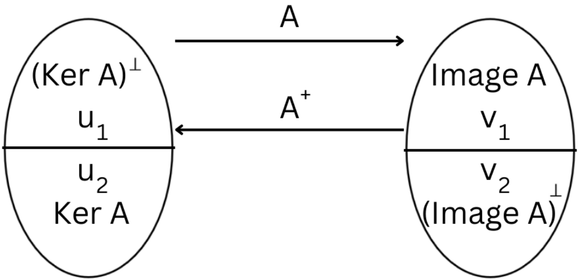
- The proof is followed. Assume that $\exists u, \tilde{u} \in \mathbb{R}^m : A(u) = A(\tilde{u}) = v_1$ and

$$u = u_1 + u_2; \qquad u_1 \in Ker(A)^{\perp}; \quad u_2 \in Ker(A)$$
$$\tilde{u} = \tilde{u}_1 + \tilde{u}_2; \qquad \tilde{u}_1 \in Ker(A)^{\perp}; \quad \tilde{u}_2 \in Ker(A)$$
$$A(u - \tilde{u}) = \mathbf{0} \Rightarrow u - \tilde{u} \in Ker(A) \Rightarrow u_1 - \tilde{u}_1 \in Ker(A) \Rightarrow u_1 = \tilde{u}_1.$$

- The Moore-Penrose inverse is defined as A^+

$$A^+ : \mathbb{R}^n \to \mathbb{R}^m$$
$$A^+(v) = u_1.$$

Moore-Penrose Inverse



Properties of Moore-Penrose Inverse

- $(A^+)^+ = A$ is like $(A^{-1})^{-1} = A$.
  Proof: Given $u = u_1 + u_2$ with $u_1 \in Ker(A)^\perp, u_2 \in Ker(A)$.
  Map between u_1 and v_1 is bijective, thus $(A^+)^+ u_1 = v_1 = Au$
- $AA^+A = A$, it means that $AA^+Au = Au$ for all $u \in \mathbb{R}^m$.
  Proof:

$$Au = v; \qquad A^+v = u_1 \Rightarrow u = u_1 + u_2, \quad u_1 \perp u_2$$
$$Au = v = Au_1; \qquad AA^+Au = AA^+v = Au_1.$$

- $A^+AA^+ = A^+$, it means that $A^+AA^+y = A^+y$.

- $(A^{+}A)^{\top} = A^{+}A$.
  Proof: For $u \in \mathbb{R}^m$,

$$u = u_1 + u_2; \qquad u_1 \in Ker(A)^{\perp}; \quad u_2 \in Ker(A)$$
$$A^{+}Au = u_1.$$

  Then $A^{+}A$ is an orthogonal projection on the subspace $Ker(A)^{\perp}$

$$A^{+}A(\mathbb{R}^m) = Ker(A)^{\perp}.$$

  Thus eigenvalues of $A^{+}A$ is the identity map on $Ker(A)^{\perp}$ and vanishes on $Ker(A)$. Thus $A^{+}A$ is symmetric.

Moore-Penrose Inverse

THE RELATIONSHIP WITH SVD: Given

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

then

$$A^+ = V\Sigma^+ U^\top,$$

where Σ^+ is the transpose of Σ with singular value σ_i is replaced by $\frac{1}{\sigma_i}$.
Proof: For some $\sigma_i > 0$, thus $v_i \in Ker(A)^\perp$

$$A = U\Sigma V^\top = \sum_i \sigma_i u_i v_i^\top \Rightarrow Av_i = \sigma_i u_i$$
$$\Rightarrow A^+ u_i = \frac{1}{\sigma_i} v_i \Rightarrow A^+ = V\Sigma^+ U^\top.$$

PROOF USING SVD:

- $AA^+A = A$.
  Proof: Assume that the dimension of $Ker(A)$ is equal to ℓ

$$\begin{aligned}(U\Sigma V^\top)(V\Sigma^+U^\top)(U\Sigma V^\top) &= U\Sigma I_n\Sigma^+ I_m \Sigma V^\top \\ &= U\Sigma\Sigma^+\Sigma V^\top = U\Sigma V^\top = A.\end{aligned}$$

- $(AA^+)^\top = AA^+$.
  Proof:

$$\begin{aligned}(U\Sigma V^\top V\Sigma^+U^\top)^\top = (U\Sigma\Sigma^+U^\top)^\top &= U(\Sigma^+)^\top\Sigma^\top U \\ &= U\Sigma\Sigma^+U^\top = U\Sigma V^\top V\Sigma^+U^\top = AA^+.\end{aligned}$$

Stochastic Matrix

Left stochastic matrix and right stochastic matrix

$$P = \begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,n} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{bmatrix}$$

The right stochastic matrix is P which has the properties: $\sum_j p_{i,j} = 1$.
The name “right"comes from $P\mathbf{1}_n = \mathbf{1}_n$.

- P is the left stochastic matrix iff $P^\top$ is the right stochastic matrix.
- P is double stochastic matrix iff P is both left and right stochastic matrix.

Stochastic Matrix

- The name stochastic comes from $p_{i,j} = \mathbb{P}(\text{state } j \mid \text{state } i)$.
- Thus, if P_1 and P_2 are right stochastic matrices, then $P_2 P_1$ is also right stochastic matrix.
  Proof: Straight calculation shows that $P_2 P_1$ is right stochastic matrix, or we could define

$$p_{1,i,j} = \mathbb{P}(\text{state } j | \text{ state } i)$$
$$p_{2,j,k} = \mathbb{P}(\text{state } k | \text{ state } j).$$

  Then

$$p_{i,k} = \mathbb{P}(\text{state } k | \text{ state } i) = \sum_j p_{1,i,j} p_{2,j,k}.$$

Some Questions about Matrix Decompositions

- Let $A \in \mathbb{R}^{m \times n}$ and $A = QR$ be its QR factorization. Let A_2 be the first two columns of A, let Q_2 be the first two columns of Q. Find R_2 such that $A_2 = Q_2 R_2$.
- Given the QR decomposition of matrix A with columns of A are linear independent, find the A^+.
- Is this always true $(AB)^+ = B^+ A^+$?
- Prove that the QR decomposition is unique for non-singular matrix.

Vector Norms

- Norms are used to measure how large the quantities are.
- ℓ_0, ℓ_1, $\ell_2, \ldots, \ell_p, \ldots, \ell_\infty$ norms.
- For vector $x = (x_1, \ldots, x_n)$, $\ell_p(x)$ is defined as

$$\ell_p(x) = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

- ℓ_0 norm is a counting norm.
- ℓ_1 norm is absolute norm.
- ℓ_2 norm is Euclidean distance.

Vector Norms

Some basic inequalities for vector norms:

- $\ell_\infty(x) = \max\left\{|x_i|; 1 \leq i \leq n\right\}$.
- For $1 < p < q$, $\ell_p(x) > \ell_q(x)$.
- $\lim_{p\to\infty} \ell_p(x) = \max_{1\leq i\leq n} |x_i|$.
- Cauchy-Schwarz's inequality

$$\ell_1(x) \leq \sqrt{n}\ell_2(x).$$

- Hölder's inequality: For $0 < p < q$

$$\ell_p(x) \leq n^{1/p-1/q}\ell_q(x).$$

- $\ell_p(x)$ is convex with respect to x when $p \geq 1$.
- $\ell_p(x)$ is concave with respect to x when $0 < p < 1$.

Matrix Norms

- Norms are all defined for vectors, they are also defined for matrix.
- For matrix A is a map from $\mathbb{R}^m$ to $\mathbb{R}^n$.

$$\|A\|_{p,q} = \max_{x \neq 0} \frac{\|Ax\|_q}{\|x\|_p}.$$

- When $p = q$, we denote $\|A\|_{p,q} = \|A\|_p$.
- Frobenius norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2(A)}$$

where $\sigma_i(A)$ are the singular values of A.

Frobenius Norm

Assume that $A = [a_{ij}]$ with $i = 1, \ldots, m$ and $j = 1, \ldots, n$.
The SVD of A is

$$A = \sum_{i}^{\min\{m,n\}} \sigma_i u_i v_i^\top = U\Sigma V^\top.$$

Then we have

$$A^\top A = V\Sigma^\top U^\top U\Sigma V^\top = V\Sigma_{n\times n}^2 V^\top$$

$$\|A\|_F^2 = \mathrm{tr}(A^\top A) = \mathrm{tr}(V\Sigma^2 V^\top) = \sum_{i=1}^{\min\{m,n\}} \sigma_i^2.$$

SCHATTEN NORM:

$$\|A\|_{S,p} = \Big(\sum_{i=1}^{\min\{m,n\}} \sigma_i^p(A) \Big)^{\frac{1}{p}}.$$

- $p = 1$ it is called nuclear norm, often used in sparsity since it is related to the rank of A.
- $p = 2$ it is called Frobenius norm.
- $\|A\|_{S,1}$ and $\|A\|_{S,2}$ are often used.
- Sometimes they use notations $\| \cdot \|_p$.

Some inequalities for matrix norms

- $\|A\|_2 \le \|A\|_F \le \sqrt{r}\|A\|_2$.
- $\|A\|_F \le \|A\|_* \le \sqrt{r}\|A\|_F$, where $\|\cdot\|_*$ is nuclear norm.
- $\frac{1}{\sqrt{n}}\|A\|_\infty \le \|A\|_2 \le \sqrt{m}\|A\|_\infty$.
- $\frac{1}{\sqrt{m}}\|A\|_1 \le \|A\|_2 \le \sqrt{n}\|A\|_1$.
- $\|A\|_2 \le \sqrt{\|A\|_1\|A\|_\infty}$.

where matrix $A \in \mathbb{R}^{m\times n}$ of rank r.

Matrix Norms

We prove some inequalities in the previous slides, given the SVD of A

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top .$$

The first inequality: For any unit vector x, we write

$$x = \sum_{i=1}^{n} x_i v_i; \qquad \sum_{i=1}^{n} x_i^2 = 1.$$

Then we have,

$$Ax = \left[\sum_{i=1}^{r} \sigma_i u_i v_i^\top \right] \left[\sum_{j=1}^{n} x_j v_j \right] = \sum_{i=1}^{r} \sigma_i x_i u_i .$$

$$\|Ax\|_2^2 = \sum_{i=1}^{r} \sigma_i^2 x_i^2 \leq \sigma_1^2 \sum_{i=1}^{r} x_i^2 \leq \sigma_1^2.$$

It follows that $\|Ax\|_2 \leq \sigma_1$ for all $\|x\| = 1$, which means that $\|A\|_2 \leq \sigma_1$. Then

$$\|A\|_2 \leq \sqrt{\sum_{i=1}^{r} \sigma_i^2} = \|A\|_F.$$

The second inequality comes from the Cauchy-Schwarz's inequality

$$\Big[\sum_{i=1}^{r} \sigma_i\Big]^2 \leq r \sum_{i=1}^{r} \sigma_i^2.$$

Matrix Norms

Some special cases of $\|A\|_{p,p}$ with $p \in \{1, 2, \infty\}$.
For $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ and $x = (x_1, \ldots, x_n)$

$$Ax = \Big(\sum_{j=1}^{n} a_{ij} x_j\Big); \qquad i = 1, 2, \ldots, m.$$

- $p = 1$:

$$\begin{aligned}\|Ax\|_1 &= \sum_{i=1}^{n} \Big|\sum_{j=1}^{n} a_{ij} x_j\Big| \le \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}||x_j| \\ &= \sum_{j=1}^{n} |x_j| \sum_{i=1} |a_{ij}| \le \max_j \sum_{i=1}^{n} |a_{ij}|.\end{aligned}$$

Equality: $x_{j^*} = 1$ for $j^* = \arg\max_j \sum_{i=1}^{n} |a_{ij}|$.

- $p = 2$: We know that $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and $\|Ax\|_1 \leq \sigma_1$. Let $x = v_1$, then we have

$$Av_1 = \sum_{i=1}^{r} \sigma_i u_i v_i^\top v_1 = \sigma_1 u_1$$
$$\|Av_1\|_2 = \sigma_1.$$

- $p = \infty$: $x = (x_1, \dots, x_n)$ and $\max |x_j| \leq 1$

$$Ax = \Big(\sum_{j=1}^{n} a_{ij} x_j \Big); \qquad i = 1, 2, \dots, m.$$
$$\|Ax\|_\infty = \max_{1 \leq i \leq m} \Big| \sum_{j=1}^{n} a_{ij} x_j \Big| \leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|.$$

Equality: $x_j = \operatorname{sign}(a_{i^* j})$ where $i^* = \arg\max_i \sum_{j=1}^{n} |a_{ij}|$.

Matrix Norms

- The Frobenius norm is like the squared Euclidean distance, which is the sum of the square of column vectors.
- The ℓ_1 norm is often used to replace ℓ_0 norm in the case of sparsity.
- The Schatten norm 1 could be used to "quantify" the rank of matrix when we need to find low-ranked approximation.
- Combination between ℓ_1 and ℓ_2 is used in Elastic Net.
- ℓ_1 and ℓ_∞ are used as linear constraints in optimization problem.

Examples

- Dantzig selector (Candes and Tao, 2004): Solve the convex program

$$\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad X\beta = y.$$

They consider solving the following alternatives

$$\min_{\beta} \|\beta\|_1 \quad \text{subject to} \quad \|X^\top r\|_\infty \le \lambda_p,$$

where $\lambda_p > 0$, where r is the residual vector $r = y - X\beta$.

- Nuclear norm (Candes and Plan, 2009): Let M be a $m \times n$ matrix of interest. However, there are only some entries of M known in the set Ω. Question:

  How do we recover M (if we know M is low-rank)?

  The problem is formularized as

$$\min \operatorname{rank}(X): \quad \text{subject to} \quad X_{ij} = M_{ij}; \quad (i,j) \in \Omega.$$

  Candes and Recht proposed to solve

$$\min \left\|X\right\|_{S,1} \quad \text{subject to} \quad X_{ij} = M_{ij}; \quad (i,j) \in \Omega.$$

It is the end of the course!!!

If you have any question or suggestion to improve the slides, please feel free to drop us an e-mail at contact@vinai.io.

Subscribe & follow us on:

- Youtube: http://youtube.com/VinAIResearch
- Linkedin: http://bit.ly/VinAILinkedIn
- Twitter: https://twitter.com/VinAI_Research