

Linear Alegbra MathNoteBook

Pro. Tom Luo



14 — Week6

14.1 Tuesday

14.1.1 Summary of last two weeks

In the first two weeks, we have learnt how to solve linear system of equations $\mathbf{Ax} = \mathbf{b}$. To understand this equation better, we learn the definition for matrices and vector space. Matrices calculation involve vectors, the columns \mathbf{Ax} are linear combination of n vectors—columns of \mathbf{A} .

Determinants

And then we learn how to describe the **quantity of a matrix**—determinant. The determinant of a square matrix is a single number. That number contains an amazing amount of information about the matrix. There are three main points about determinant:

- *Determinants is related to invertibility, rank, eigenvalue, PSD, ...*
- $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$.
- *The square matrix \mathbf{A} is invertible if and only if $\det(\mathbf{A}) \neq 0$.*

Linear Transformation

Linear transformation is another important topic. The matrix multiplication $T(\mathbf{v}) = \mathbf{Av}$ gives a linear transformation. If we consider a vector as a point in a vector space, then *the linear transformation allows movements in the space*. It “transforms” vector \mathbf{v} to another vector \mathbf{Av} . In view of linear transformation, we can understand $\det(\mathbf{AB}) = \det(\mathbf{A}) \det(\mathbf{B})$ better:

$$\det(\mathbf{A}) = \text{Volume of } \mathbf{Ak}, \text{ where } \mathbf{k} \text{ is a unit cube.}$$

If we transform the \mathbf{k} by \mathbf{A} secondly by \mathbf{B} , actually, it has the same effect of transforming \mathbf{k} by \mathbf{AB} .

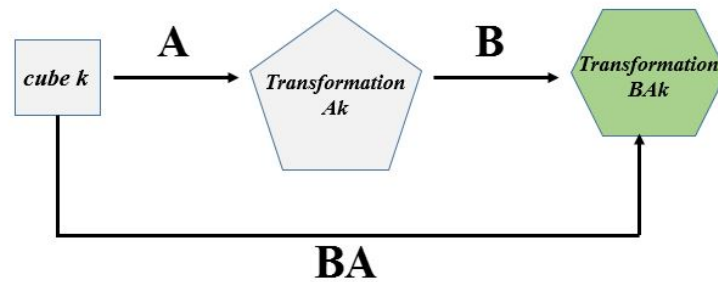


Figure 14.1: Transformation of a vector by A , then by B has the same effect by AB .

Hence if we denote the volume on a graph, we find the volume of $B(Ak)$ is exactly the same as $(BA)k$. Hence we have $\det(B)\det(A) = \det(BA)$.

Moreover, $\det(A) = 0 \iff \text{Volume of } Ak = 0 \iff \dim(Ak) = 0$.

Cramer's Rule also has geometric meaning, which will not be talked in this lecture. (In big data age, people will not use Cramer's rule frequently.)

Linear transformation has a matrix representation under certain basis. How to transform one basis into another basis? We have to use *similar matrices as matrix representation*.

Orthogonality

Why we learn orthogonality? It has two motivations:

1. Linear independence between vectors $\iff \text{Angle} \neq 0^\circ$.
Then we are interested in the special case: orthogonal $\iff \text{Angle} = 90^\circ$.
2. Solving least squares (linear regression).

Input: x = age of propellant, Output: y = shear strength.

Our data contains $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $n = 20$ samples.

We want to find a best line that fit the data:

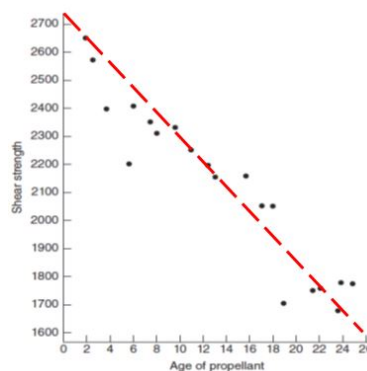


Figure 14.2: The relationship between x and y .

In other words, we want to find x s.t.

$$(\mathbf{A} \mathbf{x} \approx \mathbf{b})$$

age
coefficient
strength

The general least square problem is given by:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|^2$$

where $\mathbf{b} \in \mathbb{R}^m$.

- If $m = n$, this optimization problem is converted into find the solution to equation $\mathbf{Ax} = \mathbf{b}$.
- Otherwise, the least square solution must satisfy $\frac{\partial}{\partial \mathbf{x}} \|\mathbf{Ax} - \mathbf{b}\|^2 = \mathbf{0}$.

$$\implies \mathbf{A}^T \mathbf{Ax} = \mathbf{A}^T \mathbf{b}. \quad (\text{normal equation.})$$

This optimization problem also has geometric meaning:

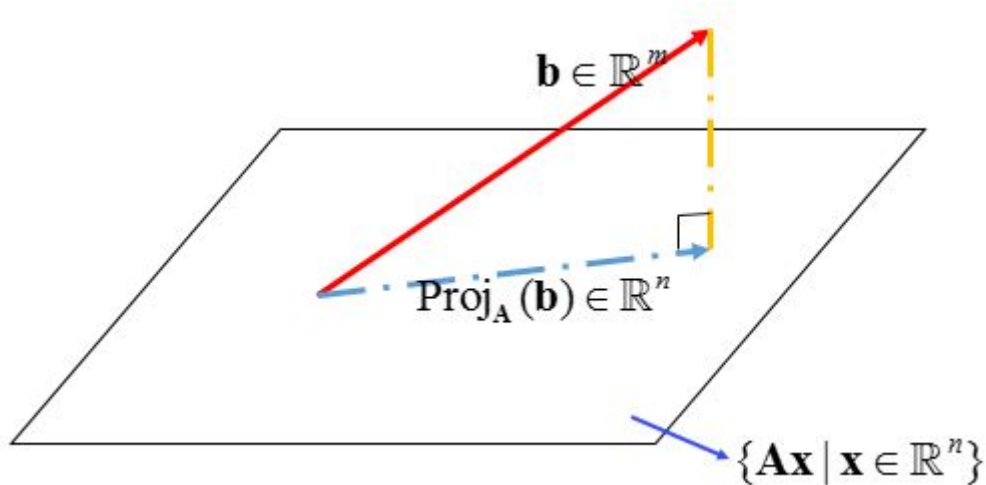


Figure 14.3: Least square problem: find \mathbf{x} such that $\mathbf{Ax} = \text{Proj}_{\mathbf{A}}(\mathbf{b})$.

So you need to memorize we only need to find \mathbf{x} such that $\mathbf{Ax} = \text{Proj}_{\mathbf{A}}(\mathbf{b})$.
But how to find $\text{Proj}_{\mathbf{A}}(\mathbf{b})$? You can write it as inner product:

$$\text{Proj}_{\mathbf{A}}(\mathbf{b}) = \mathbf{A} \frac{1}{\langle \mathbf{A}, \mathbf{A} \rangle} \langle \mathbf{A}, \mathbf{b} \rangle = \mathbf{A} (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}.$$



- The projection of \mathbf{b} onto a vector \mathbf{a} is given by:

$$\text{Proj}_{\mathbf{a}}(\mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle} \mathbf{a}$$

Since the factor $\frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle}$ is a scalar, you can also write the projection as:

$$\text{Proj}_{\mathbf{a}}(\mathbf{b}) = \mathbf{a} \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\langle \mathbf{a}, \mathbf{a} \rangle}$$

- However, the projection of \mathbf{b} onto a subspace $\{\mathbf{Ax} | \mathbf{x} \in \mathbb{R}^n\}$ is given by

$$\text{Proj}_{\mathbf{A}}(\mathbf{b}) = \mathbf{A} \frac{1}{\langle \mathbf{A}, \mathbf{A} \rangle} \langle \mathbf{A}, \mathbf{b} \rangle$$

We **cannot** write this projection as $\frac{1}{\langle \mathbf{A}, \mathbf{A} \rangle} \langle \mathbf{A}, \mathbf{b} \rangle \mathbf{A}$, since the factor $\frac{1}{\langle \mathbf{A}, \mathbf{A} \rangle} \langle \mathbf{A}, \mathbf{b} \rangle$ is a vector instead of a scalar.

The least square solution is given by

$$\mathbf{x} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{b}.$$

If $\mathbf{A} = \mathbf{Q}$, where \mathbf{Q} is a *orthogonal matrix*, then the solution is converted as

$$\mathbf{x} = \mathbf{Q}^T \mathbf{b}.$$

14.1.2 Eigenvalues and eigenvectors

Why do we study eigenvalues and eigenvectors?

- **Motivation 1:** If we consider matrices as the *movements* (linear transformation) for *vectors* in vector space. Then roughly speaking, *eigenvalues* are the *speed* of the movements, *eigenvectors* are the *direction* of the movements
- **Motivation 2:** We know that linear transformation has different matrix representation for different basis. But which representation is **simplest** for one linear transformation? This section gives us answer to this question.

When vectors are multiplied by A , almost all vectors change direction. If \mathbf{x} has the same direction as $A\mathbf{x}$, they are called **eigenvectors**.

The key equation is $A\mathbf{x} = \lambda\mathbf{x}$, The numebr λ is the eigenvalue of A .

Definition 14.1 — Eigenvectors and Eigenvalues. Let A be $n \times n$ matrix. A scalar λ is an **eigenvalue** of A iff \exists a vector $\mathbf{x} \neq \mathbf{0}$ s.t. $A\mathbf{x} = \lambda\mathbf{x}$. The vector \mathbf{x} is called an **eigenvector** (corresponding to λ .)

■ Example 14.1

$$A = \begin{bmatrix} 4 & -2 \\ 1 & 1 \end{bmatrix} \quad \mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$A\mathbf{x} = \begin{bmatrix} 6 \\ 3 \end{bmatrix} = 3 \begin{bmatrix} 2 \\ 1 \end{bmatrix} = 3\mathbf{x}.$$

$\lambda = 3$ is the eigenvalue of A .

$\mathbf{x} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ is the eigenvalue of A associated with $\lambda = 3$.

Proposition 14.1 If \mathbf{x} is an eigenvector of A , so is $\alpha\mathbf{x}$ for all *nonzero* scalar α . (These vectors have the same eigenvalue.)

Calculation

How to find λ and \mathbf{x} ? In other words, how to solve the nonlinear equation $A\mathbf{x} = \lambda\mathbf{x}$, where λ and \mathbf{x} are unknowns? If we can know the eigenvalues λ , then we can solve the system $(\lambda I - A)\mathbf{x} = \mathbf{0}$ to get the corresponding eigenvectors.

But how to find eigenvalues? $A\mathbf{x} = \lambda\mathbf{x}$ has a nonzero solution $\iff (\lambda I - A)\mathbf{x} = \mathbf{0}$ has a nonzero solution $\iff (\lambda I - A)$ is singular $\iff \det(\lambda I - A) = 0$.

This is how to recognize an eigenvalue λ :

Proposition 14.2 The number λ is the eigenvalue of A if and only if $\lambda I - A$ is singular.

$$\text{Equation for the eigenvalues} \quad \det(\lambda I - A) = 0. \quad (14.1)$$

Definition 14.2 — characteristic polynomial. Define $P_A(\lambda) := \det(\lambda I - A)$.

Then $P_A(\lambda) = \det(\lambda I - A)$ is called the **characteristic polynomial** for the matrix A .

And the equation $\det(\lambda I - A) = 0$ is called the **characteristic equation** for the matrix A .

If $P_A(\lambda^*) = 0$, then we say λ^* is the root of $P_A(\lambda)$.

The roots of $P_A(\lambda)$ are the **eigenvalues** of A . $\forall \mathbf{x} \in N(\lambda I - A)$ (*eigenspace*) is an eigenvector associated with λ .

■ **Example 14.2** Find the eigenvalues and eigenvectors of $\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 3 & -2 \end{bmatrix}$.

$$\det(\lambda \mathbf{I} - \mathbf{A}) = \begin{vmatrix} \lambda - 3 & -2 \\ -3 & \lambda + 2 \end{vmatrix} = 0.$$

$$\implies (\lambda + 3)(\lambda - 2) - 6 = 0. \implies \lambda^2 - \lambda - 12 = 0. \implies \lambda_1 = 4 \quad \lambda_2 = -3.$$

Eigenvalues of \mathbf{A} are $\lambda_1 = 4$ and $\lambda_2 = -3$.

In order to get eigenvectors, we solve $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$:

- For λ_1 , $(\mathbf{A} - \lambda_1 \mathbf{I})\mathbf{x} = \begin{bmatrix} -1 & 2 \\ 3 & -6 \end{bmatrix} \mathbf{x} = \mathbf{0}$.

$$\implies \mathbf{x} = \begin{bmatrix} 2x_2 \\ x_2 \end{bmatrix} = x_2 \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Hence any $\alpha \begin{bmatrix} 2 & 1 \end{bmatrix}^T$ ($\alpha \neq 0$) is the eigenvector of \mathbf{A} associated with $\lambda_1 = 4$.

- For λ_2 , similarly, we derive

$$\mathbf{x} = \begin{bmatrix} -x_2 \\ 3x_2 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 3 \end{bmatrix}$$

Hence any $\beta \begin{bmatrix} -1 & 3 \end{bmatrix}^T$ ($\beta \neq 0$) is the eigenvector of \mathbf{A} associated with $\lambda_2 = -3$. ■

Possible difficulty: how to solve $\det(\lambda \mathbf{I} - \mathbf{A}) = 0$?

$P_{\mathbf{A}}(\lambda)$ is a characteristic polynomial with degree n . Actually, we can write $P_{\mathbf{A}}(\lambda)$ as:

$$P_{\mathbf{A}}(\lambda) = \lambda^n - a_1 \lambda^{n-1} + a_2 \lambda^{n-2} - \dots + (-1)^n a_n$$

When n increases, it's hard to find its roots:

- When $n = 2$, solution to $ax^2 + bx + c = 0$ has *closed form*, which means we can express x in terms of a, b, c directly.
- When $n = 3$, solution to $ax^3 + bx^2 + cx + d = 0$ has *closed form*, which has been proved in 15th century.
- When $n = 4$, solution to $ax^4 + bx^3 + cx^2 + dx + e = 0$ also has *closed form*.
- However, when $n \geq 5$, the characteristic equation has *no closed form* solution, which has been proved by Galois and Abel.

Although we cannot find closed form solution for large n , does there exist such solution which is not closed form? Gauss gives us the answer:

Theorem 14.1 — Fundamental theorem of algebra. Every nonzero, single variable, degree n polynomial with *complex coefficients* has *exactly* n complex roots. (Counted with multiplicity.)

What's the meaning of *multiplicity*?


For example, the polynomial $(x - 1)^2$ has one root 1 with multiplicity 2.

Implication:

Hence every polynomial $f(x)$ could be written as

$$\begin{aligned} f(x) &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x_1 + a_0 \\ &= a_n (x - x_1)(x - x_2) \cdots (x - x_n) \end{aligned}$$

where x_i 's are roots for $f(x)$.

 Exact roots are almost impossible to find. But approximate roots (eigenvalues) can be found easily by numerical algorithm. (such as Newton's method.)

14.1.3 Products and Sums of Eigenvalue

Suppose $P_A(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A})$ has n roots $\lambda_1, \dots, \lambda_n$, then we obtain:

$$P_A(\lambda) = \det(\lambda \mathbf{I} - \mathbf{A}) = (\lambda - \lambda_1) \cdots (\lambda - \lambda_n) \quad (14.2)$$

Why the coefficient for λ^n is 1 in equation (14.2)? If we expand $\det(\lambda \mathbf{I} - \mathbf{A})$, we find

$$\det(\lambda \mathbf{I} - \mathbf{A}) = \begin{vmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{n1} & \cdots & \cdots & \lambda - a_{nn} \end{vmatrix} \quad (14.3)$$

So λ only appears in diagonal. If we expand the determinant, the coefficient is obviously 1. Moreover, in (14.2), the coefficient of λ^{n-1} is

$$-(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$$

In (14.3), λ^{n-1} only appears among $(\lambda - a_{11})(\lambda - a_{22}) \cdots (\lambda - a_{nn})$. Hence the coefficient of λ^{n-1} is

$$-(a_{11} + a_{22} + \cdots + a_{nn})$$

Consequently, we derive

$$\sum \lambda_i = \text{trace} = \sum a_{ii}$$

The sum of the entries on the main diagonal is called the **trace** of \mathbf{A} .

If we let $\lambda = 0$ in (14.2), then we obtain $\det(-\mathbf{A}) = (-1)^n \lambda_1 \lambda_2 \cdots \lambda_n$.

And obviously, $\det(-\mathbf{A}) = (-1)^n \det(\mathbf{A})$.

Hence $(-1)^n \det(\mathbf{A}) = (-1)^n \lambda_1 \lambda_2 \cdots \lambda_n \implies \det(\mathbf{A}) = \lambda_1 \lambda_2 \cdots \lambda_n$. So we have two useful conclusions:

Theorem 14.2 *The product of the n eigenvalues equals the determinant.
The sum of the n eigenvalues equals the sum of the n diagonal entries.*

14.1.4 Application: Page Rank and Web Search

If we do keyword search on google, every keyword will return 20k pages. But how to generate more useful pages for us? Our goal is to *compute a vector* $\mathbf{x} = [x_1 \ x_2 \ \cdots \ x_n]$ in \mathbb{R}^n , each x_i represents the importance of the page. Then we only need to generate the most important pages for us.

The information we use: links.

We use the number of links to judge whether a page is important. For example,

- A page *linked to by* 10^5 pages is more important than a page *linked by* 10 pages.
- If two pages *both linked to by* 100 pages, are they the same important?

The answer is no. For example, in your own blogs, if you create 100 pages that all link to your home page, then obviously your home page is not so important. These 100 pages are created by yourself, which are called **fake pages**.

We do the following assumptions:

- Assume we have $100n$ people are visiting pages. We have n pages.
- We assume every page have 100 visitors. They follow the links on that page.
- And we assume *multiple of links are equally split*. (For example, if one page have 5 links, then there will be $100/5 = 20$ people follow each link.)

To start with, the distribution of people for n pages is given by:

$$\mathbf{x}^0 = \begin{pmatrix} x_1^0 \\ x_2^0 \\ \vdots \\ x_n^0 \end{pmatrix}$$

We assume the probability people will go from page j to page i is a_{ij} .

Question: If page j links to by 5 pages, what's a_{ij} ?

Answer: due to assumption 3, roughly speaking, $a_{ij} = \frac{1}{5}$.

However, this answer is not absolutely right. Since assumption is not always true. If we consider *stochastic process*, then a_{ij} is given by:

$$a_{ij} = 0.85 \times \frac{1}{5} + 0.15 \times \frac{1}{n}.$$

If we write matrix $\mathbf{A} = [a_{ij}]_{1 \leq i, j \leq n}$, then the (i, j) entry of \mathbf{A} represents the probability that a random Web surfer will link from page j to page i .

Next step distribution:

Hence when each surfer follow one link of the pages, the distribution of people among pages is given by

$$\mathbf{x}^1 = \mathbf{A}\mathbf{x}^0 = \begin{pmatrix} x_1^1 \\ x_2^1 \\ \vdots \\ x_n^1 \end{pmatrix}$$

In general, we obtain $\mathbf{x}^{k+1} = \mathbf{A}\mathbf{x}^k$.

If the sequence $\{\mathbf{x}^k\}$ converges, then take the limit $k \rightarrow \infty$ suppose $\lim_{k \rightarrow \infty} \mathbf{x}^k = \mathbf{x}^*$. Then we obtain:

$$\mathbf{x}^{k+1} = \mathbf{A}\mathbf{x}^k \implies \mathbf{x}^* = \mathbf{A}\mathbf{x}^*$$

Hence we only need to get \mathbf{x}^* , which is the *eigenvector* of \mathbf{A} .

Once we get the distribution of people for pages, we get the importance of each page.

