SI231b: Matrix Computations

Lecture 4: Basic Concepts (Part 3)

Yue Qiu

qiuyue@shanghaitech.edu.cn

School of Information Science and Technology

ShanghaiTech University

Sept. 26, 2021

Basic Concepts: Part 3

- subspaces, span
- dimension of subspaces, rank
- inner product, orthogonality
- matrix products, computational complexity

If $\mathcal{X}$ and $\mathcal{Y}$ are subspaces of a vector space $\mathcal{V}$, define the sum of two subspaces by

$$\mathcal{X} + \mathcal{Y} = \{\mathrm{x} + \mathrm{y} | \mathrm{x} \in \mathcal{X} \text{ and } \mathrm{y} \in \mathcal{Y}\}$$

then

- the sum $\mathcal{X} + \mathcal{Y}$ is again a subspace of $\mathcal{V}$
- if $\mathcal{S}_X$, $\mathcal{S}_Y$ spans $\mathcal{X}$ and $\mathcal{Y}$, then $\mathcal{S}_X \cup \mathcal{S}_Y$ spans $\mathcal{X} + \mathcal{Y}$

Examples

- If $\mathcal{X} \subset \mathbb{R}^2$ and $\mathcal{Y} \subset \mathbb{R}^2$ are subspaces defined by two different lines through the origin, then $\mathcal{X} + \mathcal{Y} = \mathbb{R}^2$
- If $\mathcal{X}$ is a subspace represents a plane passing through the origin in $\mathbb{R}^3$ and $\mathcal{Y}$ is a subspace defined by the line through the origin that is perpendicular to $\mathcal{X}$, $\mathcal{X} + \mathcal{Y} = \mathbb{R}^3$

Direct Sum of Subspaces

Let $\mathcal{X}$ and $\mathcal{Y}$ be subspaces of a vector space $\mathcal{V}$, then $\mathcal{V}$ is said to be a direct sum of $\mathcal{X}$ and $\mathcal{Y}$, i.e., $\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$, if

$$\mathcal{V} = \mathcal{X} + \mathcal{Y} \quad \text{and} \quad \mathcal{X} \cap \mathcal{Y} = \{0\}$$

Equivalently,

Every vector u from the vector space $\mathcal{V}$ can be uniquely represented by

$$\mathsf{u} = \mathsf{u}_1 + \mathsf{u}_2$$

with $\mathsf{u}_1 \in \mathcal{X}$ and $\mathsf{u}_2 \in \mathcal{Y}$. Then we use

$$\mathcal{V} = \mathcal{X} \oplus \mathcal{Y}$$

to represent the direct sum of $\mathcal{X}$ and $\mathcal{Y}$.

Example:

$$\text{span}\{\mathsf{e}_1,\ \mathsf{e}_2\} \oplus \text{span}\{\mathsf{e}_3\} = \mathbb{R}^3$$

Range Spaces

1. The range of a matrix $A \in \mathbb{R}^{m \times n}$ denoted by $\mathcal{R}(A)$, is defined to be the subspace of $\mathbb{R}^m$ generated by the range of Ax

$$\mathcal{R}(A) = \{y \in \mathbb{R}^m \mid y = Ax,\ x \in \mathbb{R}^n\} \subset \mathbb{R}^m$$

   - also called column space

2. The range of A^T is the subspace of $\mathbb{R}^n$ defined by

$$\mathcal{R}(A^T) = \left\{x \in \mathbb{R}^n \mid x = A^T y,\ y \in \mathbb{R}^m\right\} \subset \mathbb{R}^n$$

   - also called row space

3. $\mathcal{R}(A)$ is the set of all "images" of vectors $x \in \mathbb{R}^n$ under transformation by A, sometimes $\mathcal{R}(A)$ is called the image space of A.

Null Spaces

1. The null space of a matrix $A \in \mathbb{R}^{m \times n}$ denoted by $\mathcal{N}(A)$, is defined to be the subspace of $\mathbb{R}^n$ with

$$\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \subset \mathbb{R}^n$$

- $\mathcal{N}(A)$ is simply the set of all solutions to the homogeneous system $Ax = 0$.

2. Similarly, the nullspace of A^T, i.e., $\mathcal{N}(A^T)$

$$\mathcal{N}(A^T) = \left\{y \in \mathbb{R}^m \mid A^T y = 0\right\} \subset \mathbb{R}^m$$

- also called left-hand nullspace of A since it is the set of all solutions to the left-hand homogeneous system $y^T A = 0^T$

Dimension of Subspaces

The dimension of a nontrivial subspace $\mathcal{S}$ is defined as the number of elements of a basis for $\mathcal{S}$.

- the dimension of of the trivial subspace $\{0\}$ is defined as 0.
- $\dim \mathcal{S}$ will be used as the notation for denoting the dimension of $\mathcal{S}$
- physical meaning: effective degrees of freedom of the subspace
- examples:
  - $\dim \mathbb{R}^m = m$
  - if k is the number of maximal linearly independent vectors of $\{a_1, \ldots, a_n\}$, then $\dim \operatorname{span}\{a_1, \ldots, a_n\} = k$.

Properties:

- let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^m$ be subspaces. If $\mathcal{S}_1 \subseteq \mathcal{S}_2$, then $\dim \mathcal{S}_1 \leq \dim \mathcal{S}_2$.
- let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^m$ be subspaces. If $\mathcal{S}_1 \subseteq \mathcal{S}_2$ and $\dim \mathcal{S}_1 = \dim \mathcal{S}_2$, then $\mathcal{S}_1 = \mathcal{S}_2$.
- let $\mathcal{S} \subseteq \mathbb{R}^m$ be a subspace. Then

$$\dim \mathcal{S} = m \Longleftrightarrow \mathcal{S} = \mathbb{R}^m.$$

- let $\mathcal{S}_1, \mathcal{S}_2 \subseteq \mathbb{R}^m$ be subspaces. We have $\dim(\mathcal{S}_1 + \mathcal{S}_2) \leq \dim \mathcal{S}_1 + \dim \mathcal{S}_2$.
  - as a more advanced result, we also have

$$\dim(\mathcal{S}_1 + \mathcal{S}_2) = \dim \mathcal{S}_1 + \dim \mathcal{S}_2 - \dim(\mathcal{S}_1 \cap \mathcal{S}_2).$$

  - if $\mathcal{S} = \mathcal{S}_1 \oplus \mathcal{S}_2$, then

$$\dim \mathcal{S} = \dim \mathcal{S}_1 + \dim \mathcal{S}_2$$

Rank

The rank of a matrix $A \in \mathbb{R}^{m \times n}$, denoted by $\mathrm{rank}(A)$, is defined as the number of elements of a maximal linearly independent subset of $\{a_1, \ldots, a_n\}$.

- $\mathrm{rank}(A)$ is the maximum number of linearly independent columns of A
- $\dim \mathcal{R}(A) = \mathrm{rank}(A)$ by definition

Facts:

- $\mathrm{rank}(A) = \mathrm{rank}(A^T)$, i.e., the rank of A is also the maximum number of linearly independent rows of A

  Proof?

- $\mathrm{rank}(A + B) \leq \mathrm{rank}(A) + \mathrm{rank}(B)$
- $\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$.
  - Equality holds when A and B are full rank.

- $A \in \mathbb{R}^{m \times n}$ is said to have
  - full column rank if the columns of A are linearly independent (more precisely, the collection of *all* columns of A is linearly independent)
    - $A \in \mathbb{R}^{m \times n}$ being of full column rank $\Longleftrightarrow m \geq n, \text{rank}(A) = n$
  - full row rank if the rows of A are linearly independent
    - $A \in \mathbb{R}^{m \times n}$ being of full row rank $\Longleftrightarrow m \leq n, \text{rank}(A) = m$
  - full rank if $\text{rank}(A) = \min\{m, n\}$; i.e., it has either full column rank or full row rank
  - rank deficient if $\text{rank}(A) < \min\{m, n\}$

A square matrix A is said to be nonsingular or invertible if

- A is full rank
- all the columns of A are linear independent
- $Ax = 0 \iff x = 0$
- alternatively, we say A is singular if $Ax = 0$ for some $x \neq 0$.

The inverse of an invertible A, denoted by A^{-1}, is a square matrix that satisfies

$$A^{-1}A = I.$$

Facts (for a nonsingular A):

- A^{-1} always exists and is unique (or there are no two inverses of A)
- A^{-1} is nonsingular
- $AA^{-1} = I$
- $(A^{-1})^{-1} = A$
- $(AB)^{-1} = B^{-1}A^{-1}$, where A, B are square and nonsingular
- $(A^T)^{-1} = (A^{-1})^T$
    - as a shorthand notation, we will denote $A^{-T} = (A^T)^{-1}$

The inner product of two vectors $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$ is defined as

$$\langle \mathrm{x}, \mathrm{y} \rangle = \sum_{i=1}^{n} y_i x_i = \mathrm{y}^T \mathrm{x}.$$

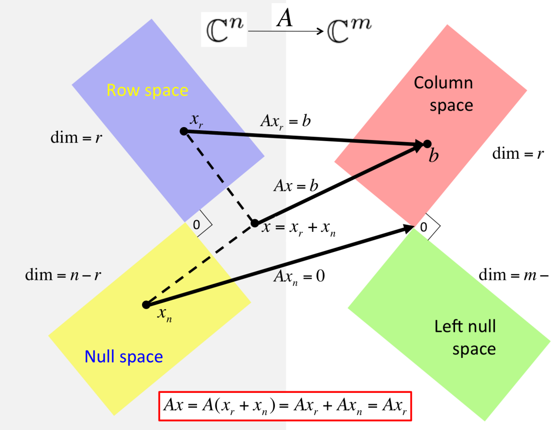
- x, y are said to be orthogonal to each other if $\langle \mathrm{x}, \mathrm{y} \rangle = 0$
- x, y are said to be parallel if $\mathrm{x} = \alpha \mathrm{y}$ for some α

The angle between two vectors $\mathrm{x}, \mathrm{y} \in \mathbb{R}^n$ is defined as

$$\theta = \arccos\left(\frac{\mathrm{y}^T \mathrm{x}}{\|\mathrm{x}\|_2 \|\mathrm{y}\|_2} \right).$$

- x, y are orthogonal if $\theta = \pi/2$
- x, y are parallel if $\theta = 0$ or $\theta = \pi$

Orthogonality of Four Fundamental Subspaces



- $\mathcal{R}(A) \perp \mathcal{N}(A^T)$
- $\mathcal{R}(A^T) \perp \mathcal{N}(A)$
- Details will follow in the later part

Rank Nulllity Theorem

Theorem

For $\mathrm{A} \in \mathbb{R}^{m \times n}$, *we have*

$$rank(\mathrm{A}) + \dim \mathcal{N}(\mathrm{A}) = n$$

Can we prove this?

Cauchy-Schwartz inequality:

$$|x^T y| \leq \|x\|_2 \|y\|_2.$$

Also, the above equality holds if and only if $x = \alpha y$ for some $\alpha \in \mathbb{R}$.

- Proof: suppose $y \neq 0$; the case of $y = 0$ is trivial. For any $\alpha \in \mathbb{R}$,

$$0 \leq \|x - \alpha y\|_2^2 = (x - \alpha y)^T (x - \alpha y) = \|x\|_2^2 - 2\alpha x^T y + \alpha^2 \|y\|_2^2. \qquad (*)$$

Also, the equality above holds if and only if $x = \beta y$ for some β. Let

$$f(\alpha) = \|x\|_2^2 - 2\alpha x^T y + \alpha^2 \|y\|_2^2.$$

The function f is minimized when $\alpha = (x^T y)/\|y\|_2^2$. Plugging this α back to $(*)$ leads to the desired result.

Hölder inequality:

$$|x^T y| \leq \|x\|_p \|y\|_q,$$

for any p, q such that $1/p + 1/q = 1,\ p \geq 1$.

- examples:
  - $(p, q) = (2, 2)$: Cauchy-Schwartz inequality
  - $(p, q) = (1, \infty)$: $|x^T y| \leq \|x\|_1 \|y\|_\infty$.

    This can be easily verified to be true:

    $$|x^T y| \leq \sum_{i=1}^n |x_i y_i| \leq \max_j |y_j| \left(\sum_{i=1}^n |x_i|\right) = \|x\|_1 \|y\|_\infty.$$

Let $\mathrm{A} \in \mathbb{R}^{m\times k}$, $\mathrm{B} \in \mathbb{R}^{k\times n}$, and consider

$$\mathrm{C} = \mathrm{AB}.$$

- column representation:

$$\mathrm{c}_i = \mathrm{A}\mathrm{b}_i, \quad i = 1, \ldots, n$$

- inner-product representation: redefine $\mathrm{a}_i \in \mathbb{R}^k$ as the ith row of A.

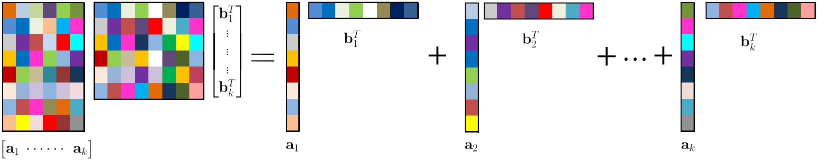
$$\mathrm{AB} = \begin{bmatrix} \mathrm{a}_1^T \\ \vdots \\ \mathrm{a}_m^T \end{bmatrix} \begin{bmatrix} \mathrm{b}_1 & \cdots & \mathrm{b}_n \end{bmatrix} = \begin{bmatrix} \mathrm{a}_1^T\mathrm{b}_1 & \cdots & \mathrm{a}_1^T\mathrm{b}_n \\ \vdots & & \vdots \\ \mathrm{a}_m^T\mathrm{b}_1 & \cdots & \mathrm{b}_m^T\mathrm{b}_n \end{bmatrix}$$

Thus,

$$c_{ij} = \mathrm{a}_i^T\mathrm{b}_j, \quad \text{for any } i, j.$$

- outer-product representation: redefine $b_i \in \mathbb{R}^k$ as the ith row of B. Thus,

$$C = AB = \sum_{i=1}^{k} a_i b_i^T$$



- The matrix of the form $X = ab^T$ for some a, b is called a rank-one outer product.
    - It can be verified that $\text{rank}(X) \le 1$, and $\text{rank}(X) = 1$ if $a \ne 0, b \ne 0$.
- the outer-product representation

$$C = \sum_{i=1}^{k} a_i b_i^T$$

is a sum of k rank-one outer products.

- does it mean that $\text{rank}(C) = k$?
    - $\text{rank}(C) \le \sum_{i=1}^{k} \text{rank}(a_i b_i^T) \le k$ is true [1]
    - but the above equality is generally not attained; e.g., $k = 2$, $a_1 = a_2$, $b_1 = -b_2$ leads to $C = 0$
    - $\text{rank}(C) = k$ only when A and B are full rank (take home exam)

[1] use the rank inequality $\text{rank}(A + B) \le \text{rank}(A) + \text{rank}(B)$.

Sometimes it may be useful to manipulate matrices in a block form.

- let $A \in \mathbb{R}^{m\times n}$, $x \in \mathbb{R}^n$. By partitioning

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

where $A_1 \in \mathbb{R}^{m\times n_1}, A_2 \in \mathbb{R}^{m\times n_2}, x_1 \in \mathbb{R}^{n_1}, x_2 \in \mathbb{R}^{n_2}$, with $n_1 + n_2 = n$, we can write

$$Ax = A_1x_1 + A_2x_2$$

- similarly, by partitioning

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

we can write

$$Ax = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 \\ A_{21}x_1 + A_{22}x_2 \end{bmatrix}$$

Block Matrix Manipulations

- consider AB. By an appropriate partitioning,

$$AB = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = A_1B_1 + A_2B_2$$

- similarly, by an appropriate partitioning,

$$AB = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix} \begin{bmatrix} B_1 & B_2 \end{bmatrix} = \begin{bmatrix} A_1B_1 & A_1B_2 \\ A_2B_1 & A_2B_2 \end{bmatrix}$$

- we showcase two-block partitioning only, but the same manipulations apply to multi-block partitioning like

$$A = \begin{bmatrix} A_{11} & \cdots & A_{1q} \\ \vdots & & \vdots \\ A_{p1} & \cdots & A_{pq} \end{bmatrix}$$

Extension to $\mathbb{C}^n$

- all the concepts described above apply to the complex case
- we only need to replace every "$\mathbb{R}$" with "$\mathbb{C}$", and every "T" with "H"; e.g.,
  - $$\text{span}\{\mathsf{a}_1, \ldots, \mathsf{a}_n\} = \{\mathsf{y} \in \mathbb{C}^m \mid \mathsf{y} = \textstyle\sum_{i=1}^n \alpha_i \mathsf{a}_i, \ \boldsymbol{\alpha} \in \mathbb{C}^n\},$$
  - $\langle \mathsf{x}, \mathsf{y} \rangle = \mathsf{y}^H \mathsf{x}$;
  - $\|\mathsf{x}\|_2 = \sqrt{\mathsf{x}^H \mathsf{x}}$, and so forth.

Extension to $\mathbb{R}^{m\times n}$

- the concepts also apply to the matrix case
  - e.g., we may write

$$\mathrm{span}\{\mathsf{A}_1,\ldots,\mathsf{A}_k\} = \{\mathsf{Y}\in\mathbb{R}^{m\times n} \mid \mathsf{Y} = \textstyle\sum_{i=1}^k \alpha_i \mathsf{A}_i,\ \boldsymbol{\alpha}\in\mathbb{R}^k\}.$$

  - sometimes it is more convenient to *vectorize* X as a vector $\mathsf{x}\in\mathbb{R}^{mn}$, and use the same treatment as in the $\mathbb{R}^n$ case
  - inner product for $\mathbb{R}^{m\times n}$:

$$\langle \mathsf{X},\mathsf{Y}\rangle = \textstyle\sum_{i=1}^m \sum_{j=1}^n x_{ij} y_{ij} = \mathrm{tr}\big(\mathsf{Y}^T\mathsf{X}\big),$$

  - the matrix version of the Euclidean norm is called the Frobenius norm:

$$\|\mathsf{X}\|_F = \sqrt{\textstyle\sum_{i=1}^m \sum_{j=1}^n |x_{ij}|^2} = \sqrt{\mathrm{tr}\big(\mathsf{X}^T\mathsf{X}\big)}$$

- extension to $\mathbb{C}^{m\times n}$ is just as straightforward as in that to $\mathbb{C}^n$

- every vector/matrix operation such as $x + y$, $y^T x$, Ax, $\ldots$ incurs computational costs, and they cost more as the vector and matrix sizes get bigger
- we typically look at floating point arithmetic operations, such as add, subtract, multiply, and divide

- flops: one flop means one floating point arithmetic operation.
- flops count of some standard vector/matrix operations: $x, y \in \mathbb{R}^n$, $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times p}$,
  - $x + y$: n adds, so n flops
  - $y^T x$: n multiplies and $n-1$ adds, so $2n-1$ flops
  - Ax: m inner products, so $m(2n-1)$ flops
  - AB: do "Ax" above p times, so $pm(2n-1)$ flops

- we are often interested in the *order* of the complexity
- big $\mathcal{O}$ notation: given two functions $f(n), g(n)$, the notation

$$f(n) = \mathcal{O}(g(n))$$

  means that there exists a constant $C > 0$ and n_0 such that $|f(n)| \leq C|g(n)|$ for all $n \geq n_0$.
- big $\mathcal{O}$ complexities of standard vector/matrix operations:
  - $\mathrm{x} + \mathrm{y}$: $\mathcal{O}(n)$ flops
  - $\mathrm{y}^T\mathrm{x}$: $\mathcal{O}(n)$ flops
  - Ax: $\mathcal{O}(mn)$ flops
  - AB: $\mathcal{O}(mnp)$ flops

Complexities of Matrix Computations

- Discussion: flop counts do not always translate into the actual efficiency of the execution of an algorithm
- things like pipelining, FPGA, parallel computing (multiple GPUs, multiple servers, cloud computing), etc., can make the story different.
- flop counts also ignore memory usage and other overheads...
- that said, we need at least a crude measure of the computational cost of an algorithm, and counting the flops serves that purpose.

How to Save Computations

- computational complexities depend much on how we design and write an algorithm
- generally, it is about
  - top-down, analysis-guided, designs:
    - seen in class, research papers
    - looks elegant
- facts are
  - usually *not* taught much in class
  - not commonplace in papers
  - subtly depends on your problem at hand
  - a bunch of small differences can make a big difference, say in actual running time
- here we give several, but by no means all, tips for saving computations

How to Save Computations

- apply matrix operations wisely
- Example: try this on Matlab

```
>> A=randn(5000,2);
>> B=randn(2,10000);
>> C=randn(10000,10000);
>>
>> tic; D= A*B*C; toc
Elapsed time is 1.334183 seconds.
>> tic; D= (A*B)*C; toc     % ask Matlab to do AB first
Elapsed time is 1.205725 seconds.
>> tic; D= A*(B*C); toc     % ask Matlab to do BC first
Elapsed time is 0.067979 seconds.
```

- let us analyze the complexities in the last example
  - $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times p}, C \in \mathbb{R}^{p\times p}$, with $n \ll \min\{m, p\}$.
  - We want to compute $D = ABC$.
  - if we compute AB first, and then $D = (AB)C$, the flop count will be

$$\mathcal{O}(mnp) + \mathcal{O}(mp^2) = \mathcal{O}(m(n+p)p) \approx \mathcal{O}(mp^2)$$

  - if we compute BC first, and then $D = A(BC)$, the flop count will be

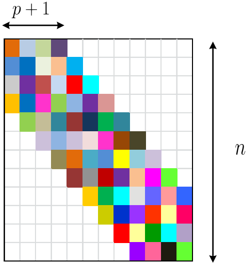
$$\mathcal{O}(np^2) + \mathcal{O}(mnp) = \mathcal{O}((m+p)np).$$

  - the 2nd option is preferable if n is much smaller than m, p

- use structures, if available
- example: let $A \in \mathbb{R}^{n\times n}$ and suppose that

  $$a_{ij} = 0 \ \text{ for all } i, j \text{ such that } |i - j| > p,$$

  for some integer $p > 0$.



  - such a structured A is called banded matrix
  - if we don't use structures, computing Ax requires $\mathcal{O}(n^2)$
  - if we use the banded + sparsity[1] structures, we can compute Ax with $\mathcal{O}(pn)$
  - different problems may have different fancy/advanced structures to be exploited

[1] a vector or matrix is said to be sparse if it contains many zeros

Readings

Readings for lecture 2 and 3

- Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*, SIAM, 2005.

  Chapter 3.1 – 3.7, 4.1 – 4.5, 5.1 – 5.4
- Gene H. Golub and Charles F. Van Loan. *Matrix Computations*, Johns Hopkins University Press, 2013.

  Chapter 1, 2.1 – 2.3