# Linear algebra for EE


Jitkomut Songsiri

Department of Electrical Engineering
Faculty of Engineering
Chulalongkorn University

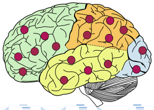
CUEE


September 3, 2023

# Outline

# How to read this handout

1. readers are assumed to have a background on elementary linear algebra in undergrad level (see chapter 'Background and notations (not taught)')
2. the note is used with lecture in EE500 (you cannot master this topic just by reading this note) – class lectures include
   - graphical concepts, math derivation of details/steps in between
   - computer codes to illustrate examples
3. pay attention to the symbol ✎; you should be able to prove such ✎ result
4. each chapter has a list of references; find more formal details/proofs from in-text citations
5. almost all results in this note can be Googled; readers are encouraged to 'stimulate neurons' in your brain by proving results without seeking help from the Internet first
6. typos and mistakes can be reported to `jitkomut@gmail.com`



# Background and notations (not taught)

# Sufficient and necessary conditions

consider a (true) conditional statement: $P \Rightarrow Q$, we say

- $P$ is **sufficient** for $Q$
- $Q$ is **necessary** for $P$
- $P$ **only if** $Q$

example: if $x = -3$ then $|x| = 3$ (a true conditional statement)

- '$P$ is sufficient for $Q$' means

  the truth of $x = -3$ is sufficient for concluding the truth of $|x| = 3$

- '$P$ only if $Q$' and '$Q$ is necessary for $P$' have the same meaning:

  $x = -3$ is *true only* under the condition that $|x| = 3$ (because if $|x| \neq 3$ then $x = -3$ can't be true) however, $|x| = 3$ is *not a sufficient condition* for $x = -3$

(because if $|x| = 3$ then $x$ can be either $3$ or $-3$)

*i.e.*, the converse of the statement: 'if $x = -3$ then $|x| = 3$' is false

consider a (true) biconditional statement: $P \Leftrightarrow Q$, we say

$P$ is **sufficient** and **necessary** for $Q$

when $P \Rightarrow Q$ **and** $Q \Rightarrow P$

example: $|x| = 2$ if and only if $x^2 = 4$ (a true biconditional statement)

- saying $|x| = 2$ is equivalent to saying $x^2 = 4$

# Vector notation

$n$-vector $x$:

$$x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

- also written as $x = (x_1, x_2, \ldots, x_n)$
- set of $n$-vectors is denoted $\mathbf{R}^n$ (Euclidean space)
- $x_i$: $i$th **element** or **component** or **entry** of $x$
- it is common to denote $x$ as a column vector
- $x^T = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}$ is then a row vector

# Matrix notation

an $m \times n$ matrix $A$ is defined as

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix}, \quad \text{or} \quad A = [a_{ij}]_{m \times n}$$

- $a_{ij}$ are the **elements**, or **coefficients**, or **entries** of $A$
- set of $m \times n$-matrices is denoted $\mathbf{R}^{m \times n}$
- $A$ has $m$ rows and $n$ columns ($m, n$ are the **dimensions**)
- the $(i, j)$ entry of $A$ is also commonly denoted by $A_{ij}$
- $A$ is called a **square** matrix if $m = n$

# Special matrices

**zero matrix:** $A = 0$

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$a_{ij} = 0$, for $i = 1, \ldots, m, j = 1, \ldots, n$

**identity matrix:** $A = I$

$$A = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

a square matrix with $a_{ii} = 1, a_{ij} = 0$ for $i \neq j$

**diagonal matrix:** a square matrix with $a_{ij} = 0$ for $i \neq j$

$$A = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_n \end{bmatrix}$$

**triangular matrix:** a square matrix with zero entries in a triangular part

**upper triangular**

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}$$

$a_{ij} = 0$ for $i \geq j$

**lower triangular**

$$A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$a_{ij} = 0$ for $i \leq j$

# Block matrix notation

**example:** $2 \times 2$-block matrix $A$

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}$$

for example, if $B, C, D, E$ are defined as

$$B = \begin{bmatrix} 2 & 1 \\ 3 & 8 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 1 & 7 \\ 1 & 9 & 1 \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 \end{bmatrix}, \quad E = \begin{bmatrix} -4 & 1 & -1 \end{bmatrix}$$

then $A$ is the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 & 1 & 7 \\ 3 & 8 & 1 & 9 & 1 \\ 0 & 1 & -4 & 1 & -1 \end{bmatrix}$$

note: dimensions of the blocks must be compatible

# Column and Row partitions

write an $m \times n$-matrix $A$ in terms of its columns or its rows

$$A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} = \begin{bmatrix} b_1^T \\ b_2^T \\ \vdots \\ b_m^T \end{bmatrix}$$

- $a_j$ for $j = 1, 2, \ldots, n$ are the columns of $A$
- $b_i^T$ for $i = 1, 2, \ldots, m$ are the rows of $A$

**example:** $A = \begin{bmatrix} 1 & 2 & 1 \\ 4 & 9 & 0 \end{bmatrix}$

$$a_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 2 \\ 9 \end{bmatrix}, \quad a_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad b_1^T = \begin{bmatrix} 1 & 2 & 1 \end{bmatrix}, \quad b_2^T = \begin{bmatrix} 4 & 9 & 0 \end{bmatrix}$$

# Matrix-vector product

product of $m \times n$-matrix $A$ with $n$-vector $x$

$$Ax = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n \\ \vdots \\ a_{m1}x_1 + a_{m2}x_2 + \ldots + a_{mn}x_n \end{bmatrix}$$

- dimensions must be compatible: $\#$ columns in $A = \#$ elements in $x$

if $A$ is partitioned as $A = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}$, then

$$Ax = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

- $Ax$ is a linear combination of the column vectors of $A$
- the coefficients are the entries of $x$

# Product with standard unit vectors

**post-multiply with a column vector**

$$Ae_k = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{bmatrix} = \text{the } k\text{th column of } A$$

**pre-multiply with a row vector**

$$\begin{aligned} e_k^T A &= \begin{bmatrix} 0 & 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix} \\ &= \begin{bmatrix} a_{k1} & a_{k2} & \cdots & a_{kn} \end{bmatrix} = \text{the } k\text{th row of } A \end{aligned}$$

# Trace

**definition:** trace of a square matrix $A$ is the sum of the diagonal entries in $A$

$$\mathbf{tr}(A) = a_{11} + a_{22} + \cdots + a_{nn}$$

**example:**

$$A = \begin{bmatrix} 2 & 1 & 4 \\ 0 & -1 & 5 \\ 3 & 4 & 6 \end{bmatrix}$$

trace of $A$ is $2 - 1 + 6 = 7$

**properties** ✎

- $\mathbf{tr}(A^T) = \mathbf{tr}(A)$
- $\mathbf{tr}(\alpha A + B) = \alpha\,\mathbf{tr}(A) + \mathbf{tr}(B)$
- $\mathbf{tr}(AB) = \mathbf{tr}(BA)$

# System of linear equations

a linear system of $m$ equations in $n$ variables

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

in matrix form: $Ax = b$

problem statement: given $A, b$, find a solution $x$ (if exists)

# Three types of linear equations

- **square** if $m = n$ ($A$ is square)

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

- **underdetermined** if $m < n$ ($A$ is fat)

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$$

- **overdetermined** if $m > n$ ($A$ is skinny)

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

# Existence and uniqueness of solutions

**range space** of $A \in \mathbf{R}^{m\times n}$ is

$$\begin{aligned} \mathcal{R}(A) &= \{\, y \in \mathbf{R}^m \mid y = Ax, \ \text{ for } x \in \mathbf{R}^n \,\} \\ \mathbf{rank}(A) &\triangleq \dim(\mathcal{R}(A)) \end{aligned}$$

**nullspace** of $A$ is

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

**important properties:** ✎

- a linear system $y = Ax$ has a solution if and only if $y \in \mathcal{R}(A)$
- equivalently, $y = Ax$ has a solution if and only if $\mathbf{rank}(A) = \mathbf{rank}([A \mid y])$
- if the linear system has a solution, the solution is unique if and only if $\mathcal{N}(A) = \{0\}$

# Inverse of matrices

**definition:** a *square* matrix $A$ is called **invertible** or **nonsingular** if there exists $B$ s.t.

$$AB = BA = I$$

- $B$ is called an **inverse** of $A$
- it is also true that $B$ is invertible and $A$ is an inverse of $B$
- if no such $B$ can be found $A$ is said to be **singular**

assume $A$ is invertible

- an inverse of $A$ is unique
- the inverse of $A$ is denoted by $A^{-1}$

# Facts about invertible matrices

assume $A, B$ are invertible

**facts** ✎

- $(\alpha A)^{-1} = \alpha^{-1} A^{-1}$ for nonzero $\alpha$
- $A^T$ is also invertible and $(A^T)^{-1} = (A^{-1})^T$
- $AB$ is invertible and $(AB)^{-1} = B^{-1} A^{-1}$
- $(A + B)^{-1} \neq A^{-1} + B^{-1}$

✌ **Theorem:** for a square matrix $A$, the following statements are equivalent

1. $A$ is invertible
2. $Ax = 0$ has only the trivial solution $(x = 0)$
3. the reduced echelon form of $A$ is $I$
4. $A$ is invertible if and only if $\det(A) \neq 0$

# Inverse of diagonal matrix

$$A = \begin{bmatrix} a_1 & 0 & \cdots & 0 \\ 0 & a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_n \end{bmatrix}$$

a diagonal matrix is invertible iff the diagonal entries are all nonzero

$$a_{ii} \neq 0, \quad i = 1, 2, \ldots, n$$

the inverse of $A$ is given by

$$A^{-1} = \begin{bmatrix} 1/a_1 & 0 & \cdots & 0 \\ 0 & 1/a_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 1/a_n \end{bmatrix}$$

the diagonal entries in $A^{-1}$ are the inverse of the diagonal entries in $A$

# Inverse of triangular matrix

**upper triangular** **lower triangular**

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix} \qquad A = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$$a_{ij} = 0 \text{ for } i \geq j \qquad a_{ij} = 0 \text{ for } i \leq j$$

a triangular matrix is invertible iff the diagonal entries are all nonzero

$$a_{ii} \neq 0, \quad \forall i = 1, 2, \ldots, n$$

- product of lower (upper) triangular matrices is lower (upper) triangular
- the inverse of a lower (upper) triangular matrix is lower (upper) triangular

# Eigenvalues

$\lambda \in \mathbf{C}$ is called an **eigenvalue** of $A \in \mathbf{C}^{n \times n}$ if

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

equivalent to:

- there exists nonzero $x \in \mathbf{C}^n$ s.t. $(\lambda I - A)x = 0$, *i.e.*,

$$Ax = \lambda x$$

  any such $x$ is called an **eigenvector** of $A$ (associated with eigenvalue $\lambda$)
- there exists nonzero $w \in \mathbf{C}^n$ such that

$$w^T A = \lambda w^T$$

  any such $w$ is called a **left eigenvector** of $A$

# Computing eigenvalues

- $\mathcal{X}(\lambda) = \det(\lambda I - A)$ is called the **characteristic polynomial** of $A$
- $\mathcal{X}(\lambda) = 0$ is called the **characteristic equation** of $A$
- eigenvalues of $A$ are the root of characteristic polynomial

# Properties

- if $A$ is $n \times n$ then $\mathcal{X}(\lambda)$ is a polynomial of order $n$
- if $A$ is $n \times n$ then there are $n$ eigenvalues of $A$
- even when $A$ is real, eigenvalues and eigenvectors can be complex, *e.g.*,

$$A = \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix}, \quad A = \begin{bmatrix} -2 & 0 & 1 \\ -6 & -2 & 0 \\ 19 & 5 & -4 \end{bmatrix}$$

- if $A$ and $\lambda$ are real, we can choose the associated eigenvector to be real
- if $A$ is real then eigenvalues must occur in complex conjugate pairs
- if $x$ is an eigenvector of $A$, so is $\alpha x$ for any $\alpha \in \mathbf{C}$, $\alpha \neq 0$
- an eigenvector of $A$ associated with $\lambda$ lies in $\mathcal{N}(\lambda I - A)$

# Important facts

denote $\lambda(A)$ an eigenvalue of $A$

- $\lambda(\alpha A) = \alpha\lambda(A)$ for any $\alpha \in \mathbf{C}$
- $\mathbf{tr}(A)$ is the sum of eigenvalues of $A$
- $\det(A)$ is the product of eigenvalues of $A$
- $A$ and $A^T$ share the same eigenvalues ✎
- $\lambda(\overline{A^T}) = \overline{\lambda(A)}$ ✎
- $\lambda(A^m) = (\lambda(A))^m$ for any integer $m$
- $A$ is invertible if and only if $\lambda = 0$ is not an eigenvalue of $A$ ✎

# Eigenvalue decomposition

if $A$ is diagonalizable then $A$ admits the decomposition

$$A = TDT^{-1}$$

- $D$ is diagonal containing the eigenvalues of $A$
- columns of $T$ are the corresponding eigenvectors of $A$
- note that such decomposition is not unique (up to scaling in $T$)

**recall:** $A$ is diagonalizable if and only if all eigenvectors of $A$ are independent

# References


1. W.K. Nicholson, *Linear Algebra with Applications*, McGraw-Hill, 2006
2. H.Anton and C. Rorres, *Elementary Linear Algebra*, John Wiley, 2011

# Block matrix and quadratic form

# Leading blocks and determinants

let's illustrate by an example of square matrices

$$A = \begin{bmatrix} 0 & -2 & -2 & 1 \\ 0 & 2 & 1 & 2 \\ -3 & -1 & -2 & 0 \\ -1 & 0 & 1 & -3 \end{bmatrix}$$

$A$ has four **leading blocks:**

$$A_1 = 0, \quad A_2 = \begin{bmatrix} 0 & -2 \\ 0 & 2 \end{bmatrix}, \quad A_3 = \begin{bmatrix} 0 & -2 & -2 \\ 0 & 2 & 1 \\ -3 & -1 & -2 \end{bmatrix}, \quad A_4 = A$$

that correspond to four **leading determinants:** (also called **principal minors**)

$$\det(A_1) = 0, \quad \det(A_2) = 0, \quad \det(A_3) = -6, \quad \det(A_4) = \det(A) = -7$$

# Linear function

given $w \in \mathbf{R}^n$ and let $x \in \mathbf{R}^n$ be a vector variable

a **linear function** $f : \mathbf{R}^n \to \mathbf{R}$ is given by

$$f(x) = w^T x = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

(✎ review its linear properties, *i.e.*, superposition)

an **affine function** is a linear function plus a constant: $f(x) = w^T x + b$

- $\frac{\partial f}{\partial x_i} = w_i$ gives the rate of change of $f$ in $x_i$ direction
- the set $\{x \mid w^T x + b = \text{ constant }\}$ is a hyperplane in $\mathbf{R}^n$ with the normal vector $w$
- linear functions are used in linear regression model and linear classifier

# Energy form

given a (real) square matrix $A$, an energy form is a quadratic function of vector $x$:

$$f : \mathbf{R}^n \to \mathbf{R}, \quad f(x) = x^T A x = \sum_i \sum_j a_{ij} x_i x_j$$

- $x^T A x$ is the same as the energy form using $(A + A^T)/2$ as the coefficient because

$$x^T A x = (x^T A x)^T = \frac{x^T (A + A^T) x}{2}$$

- using $A = \frac{A+A^T}{2} + \frac{A-A^T}{2}$, we can later on assume that an energy form requires only the symmetric part of $A$
- reverse question: given an energy form, can you determine what $A$ is ?

$$x_1^2 + 2x_2^2 + 3x_3^2 - x_1 x_2 + 2x_2 x_3 \quad \triangleq \quad x^T A x$$

# Energy form and completing the square

recall how to complete the square:

$$x_1^2 + 3x_2^2 + 14x_1x_2 = (x_1 + 7x_2)^2 - 46x_2^2$$

given these matrices, expand the energy form and complete the square

$$A = \begin{bmatrix} 4 & 6 \\ 6 & 13 \end{bmatrix}, \quad B = \begin{bmatrix} 4 & 6 \\ 6 & 9 \end{bmatrix}, \quad C = \begin{bmatrix} 4 & 6 \\ 6 & -4 \end{bmatrix}$$

- $x^T A x =$
- $x^T B x =$
- $x^T C x =$

# Quadratic function

given $P \in \mathbf{R}^{n\times n}, q \in \mathbf{R}^n, r \in \mathbf{R}$, a **quadratic** function $f : \mathbf{R}^n \to \mathbf{R}$ is of the form

$$f(x) = (1/2)x^T Px + q^T x + r$$

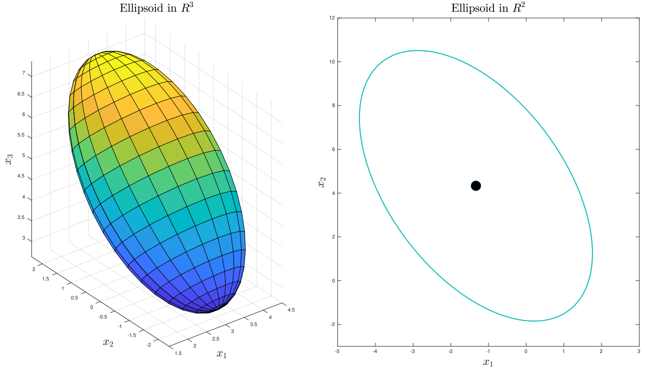
- $x^T Px$ is aka an energy form (due to the quadratic form that appears in the energy/power of some physical variables)

  $$\text{electrical power} = i^2 R, \quad \text{kinetic energy} = \frac{1}{2}mv^2, \quad \text{energy stored in spring} = \frac{1}{2}kx^2$$

- the contour shape of $f$ depends on the property of $P$ (positive definite, indefinite, magnitude of eigenvalues, direction of eigenvectors) – as we will learn shortly

# Surface plot of quadratic function

let $f(x) = (1/2)x^T Px + q^T x$ where $\lambda(P) \succ 0$



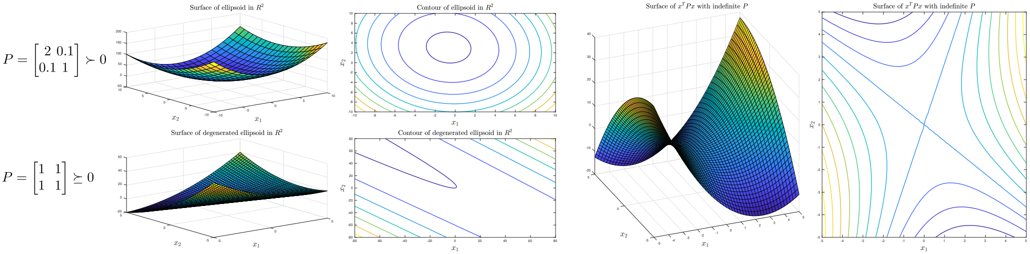

the contour plot of $f$

**ellipsoid** in $\mathbf{R}^n$

- when all eigenvalues of $P$ are positive, $P$ is **positive definite**
- direction and width of principal axes are related to eigenvalues/eigenvectors of $P$ (more on this later)

# Surface plot of quadratic function

let $f(x_1, x_2) = (1/2)(x^T P x) + q^T x$ and three cases of $P$




- case 1: all eigenvalues of $P$ are positive
- case 2: all eigenvalues of $P$ are non-negative (one is zero)
- case 3: $P = \begin{bmatrix} 2 & 1 \\ 1 & -1 \end{bmatrix}$ eigenvalues of $P$ are positive and negative

# Symmetric matrix

**definition:** a (real) square matrix $A$ is said to be **symmetric** if $A = A^T$

**notation:** $A \in \mathbf{S}^n$

**examples:**

$$\begin{bmatrix} X & Y \\ Y^T & Z \end{bmatrix} \text{ with symmetric } X, Z, \quad A = \mathbf{E}[XX^T] \text{ (correlation matrix)}$$

✎ **basic facts:**

- for any (rectangular) matrix $A$, $AA^T$ and $A^TA$ are always symmetric
- if $A$ is symmetric and invertible, then $A^{-1}$ is symmetric
- if $A$ is invertible, then $AA^T$ and $A^TA$ are also invertible

# Properties of symmetric matrix

**spectral theorem:** if $A$ is a real symmetric matrix then the following statements hold

1. all eigenvalues of $A$ are real
2. all eigenvectors of $A$ are orthogonal
3. $A$ admits a decomposition
$$A = UDU^T$$
where $U^TU = UU^T = I$ ($U$ is unitary) and a diagonal $D$ contains $\lambda(A)$
4. for any $x$, we have
$$\lambda_{\min}(A)\|x\|_2^2 \;\leq\; x^TAx \;\leq\; \lambda_{\max}(A)\|x\|_2^2$$
the first (and second) inequalities are tight when $x$ is the eigenvector corresponding to $\lambda_{\min}$ (and $\lambda_{\max}$ respectively)

# Proofs

1. assume $Ax = \lambda x$ and $\lambda, x$ could be complex, denote $x^* = \bar{x}^T$

$$\begin{aligned} (x^*Ax)^* &= x^*A^*x = x^*Ax = x^*\lambda x = \lambda x^*x \\ &= (x^*\lambda x)^* = \bar{\lambda} x^*x \end{aligned}$$

   since $x^*x \neq 0$, we must have $\lambda = \bar{\lambda}$

2. assume $Ax_1 = \lambda_1 x_1$ and $Ax_2 = \lambda_2 x_2$ (now all $(\lambda_i, x_i)$ are real)

$$\begin{aligned} x_2^T A x_1 &= x_2^T \lambda_1 x_1 = \lambda_1 x_2^T x_1 \\ &= x_1^T A x_2 = x_1^T \lambda_2 x_2 = \lambda_2 x_1^T x_2 \end{aligned}$$

   equating two terms give $(\lambda_1 - \lambda_2) x_2^T x_1 = 0$

   for simple case, we can assume that $\lambda_i$'s are distinct, so $x_2^T x_1 = 0$ $(x_2 \perp x_1)$

# Positive definite matrix

**definition:** a symmetric matrix $A$ is **positive semidefinite**, written as $A \succeq 0$ if

$$x^T A x \geq 0, \quad \forall x \in \mathbf{R}^n$$

and is said to be **positive definite**, written as $A \succ 0$ if

$$x^T A x > 0, \quad \text{for all } \textit{nonzero } x \in \mathbf{R}^n$$

✱ the curly $\succeq$ symbol is used with matrices (to differentiate it from $\geq$ for scalars )

example: $A = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \succeq 0$ because

$$x^T A x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2x_2^2 - 2x_1 x_2 = (x_1 - x_2)^2 + x_2^2 \geq 0$$

exercise: ✎ check positive semidefiniteness of matrices on page 33

# How to test if $A \succeq 0$?

**Theorem:** $A \succeq 0$ if and only if all eigenvalues of $A$ are non-negative

($A \succ 0$ if and only if $\lambda(A) > 0$)

**Sylvester's criterion:** if every principal minor of $A$ (including $\det A$) is non-negative then $A \succeq 0$

proof in Horn Theorem 7.2.5

example 1: $A = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix} \succ 0$ because

- eigenvalues of $A$ are $0.38$ and $2.61$ (real and positive)
- the principle minors are $1$ and $\begin{vmatrix} 1 & -1 \\ -1 & 2 \end{vmatrix} = 1$ (all positive)

example 2: $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \succeq 0$ because eigenvalues of $A$ are 0 and 3

# Properties of positive definite matrix

1. if $A \succeq 0$ then all the diagonal terms of $A$ are nonnegative
2. if $A \succeq 0$ then all the leading blocks of $A$ are positive semidefinite
3. if $A \succeq 0$ then $BAB^T \succeq 0$ for any $B$ ✍ (exercise)
4. if $A \succeq 0$ and $B \succeq 0$, then so is $A + B$
5. a diagonal psdf $D = \mathbf{diag}(d_1, d_2, \ldots, d_n)$ admits a square root denoted by $D^{1/2}$

$$D^{1/2}D^{1/2} = D \quad \text{where} \quad D^{1/2} := \mathbf{diag}(\sqrt{d_1}, \sqrt{d_2}, \ldots, \sqrt{d_n})$$

   (this choice of $D^{1/2}$ is also positive semidefinite)

6. if $A \succeq 0$ then $A$ has a square root, denoted as a symmetric $A^{1/2}$ such that

$$A^{1/2}A^{1/2} = A$$

# Square root of positive semidefinite matrix

**definition:** a square root of $A \succeq 0$ is a symmetric matrix denoted by $A^{1/2}$ such that

$$A^{1/2}A^{1/2} = A$$

example:

$$D = \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}, \quad D^{1/2} = \begin{bmatrix} \sqrt{2} & 0 \\ 0 & -\sqrt{6} \end{bmatrix}, \quad A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad A^{1/2} = \frac{1}{2}\begin{bmatrix} 1+\sqrt{3} & 1-\sqrt{3} \\ 1-\sqrt{3} & 1+\sqrt{3} \end{bmatrix}$$

how to find a square root?: one way is from the eigenvalue decomposition

$$A = UDU^T = UD^{1/2}D^{1/2}U^T = UD^{1/2}U^TUD^{1/2}U^T \Rightarrow A^{1/2} := UD^{1/2}U^T$$

- $A^{1/2}$ is not unique but we can choose $A^{1/2}$ that is positive semidefinite
- ❄ $A^{1/2}$ is NOT the matrix with entries $\sqrt{a_{ij}}$
- different definition exists: if $A = B^TB$ then $B$ is called a square root of $A$

# Positive definite matrices in applications

1. covariance matrix: $C = \mathbf{E}[(X-\mu)(X-\mu)^T]$
2. Hessian of convex functions: *e.g.*, $f(x) = \sum_{i=1}^n x_i \log(x_i)$
3. given $Q \succ 0$ there exists a unique $P \succ 0$ satisfying the **Lyapunov equations**

   $$\text{(continuous) } A^T P + PA + Q = 0, \qquad \text{(discrete) } A^T PA - P + Q = 0$$

   if and only if the autonomous linear system is asymptotically stable
4. a matrix in a form of $A^T A$ is called a **Gram matrix**, *e.g.*, appear in quadratic term of dual SVM (Gram is pdf when $A$ is full rank)
5. another name of Gram is Gramian matrix (as in control theory)

   $$W_c = \int_0^\infty e^{A\tau} BB^T e^{A^T\tau} d\tau, \quad \text{can be solved via } AW_c + W_c A^T = -BB^T$$

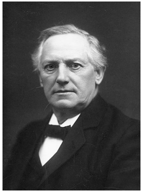
   controllability: $(A, B)$ is controllable iff $W_c \succ 0$

# Gram matrix

for an $m \times n$ matrix $A$ with columns $a_1, \dots, a_n$, the product $G = A^T A$ is called the **Gram matrix**

Gram matrix is positive semidefinite

Jørgen Pedersen Gram



$$G = A^T A = \begin{bmatrix} a_1^T a_1 & a_1^T a_2 & \cdots & a_1^T a_n \\ a_2^T a_1 & a_2^T a_2 & \cdots & a_2^T a_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n^T a_1 & a_n^T a_2 & \cdots & a_n^T a_n \end{bmatrix}$$

$$x^T G x = x^T A^T A x = \|Ax\|^2 \geq 0, \quad \forall x$$

- if $A$ has zero nullspace then $Ax = 0 \leftrightarrow x = 0$; this implies that $A^T A \succ 0$
- let $X$ be a data matrix, partitioned in $N$ rows as $x_k^T$'s; we typically encounter $G = XX^T = \sum_{k=1}^N x_k x_k^T$ as the **sample covariance matrix**

# Negative definite and indefinite

more definitions

- $A$ is called a **negative semidefinite** matrix if $-A$ is positive semidefinite

$$A = \begin{bmatrix} -2 & 1 \\ 1 & -3 \end{bmatrix} \preceq 0 \quad \text{(all eigenvalues of } A \text{ are non-positive)}$$

  (recall the Lyapunov theory in control: $A^T P + PA \preceq 0$)

- if $A$ is neither positive semidefinite matrix nor negative semidefinite matrix, $A$ is said to be **indefinite**

$$A = \begin{bmatrix} 2 & -3 \\ -3 & 1 \end{bmatrix} \not\succeq 0, \quad \text{(eigenvalues of } A \text{ have mixed signs)}$$

  (its energy form $x^T Ax$ is not monotone – can be increasing or decreasing, depending on $x$)

# Exercises on positive definite matrix

1. for which $a$ and $c$ is this matrix pdf ?

$$A = \begin{bmatrix} a & a & a \\ a & a+c & a-c \\ a & a-c & a+c \end{bmatrix}$$

2. let $x \in \mathbf{R}^n$, is $xx^T \succeq 0$? is $xx^T \succ 0$?
3. if $A \succeq 0$, and let $\alpha > 0$, is $A + \alpha I \succ 0$?
4. prove that if $A \succeq 0$ then $BAB^T \succeq 0$ for any $B$
5. let $A \succ 0$, under what condition on $B$ is $BAB^T \succ 0$?
6. let $A = \begin{bmatrix} 2 & 4 \\ 4 & 9 \end{bmatrix}$ i) check if $A \succ 0$, ii) find the smallest $\alpha \in \mathbf{R}$ such that $A + \alpha I \succeq 0$

# Common misunderstanding about pdf matrices

1. $A \succeq 0$ does NOT mean all entries of $A$ are positive!
2. if $x^T A x \geq 0$ for *some* $x$, it does NOT imply that $A \succeq 0$
3. the converse of some statements on page 42 is NOT true
   - ✘ if all diagonal terms of $A$ are nonnegative then $A \succeq 0$
   - ✘ if all the leading blocks of $A$ are positive semidefinite then $A \succeq 0$
   - ✘ if $A + B \succeq 0$ then $A$ and $B$ are positive semidefinite

# Can we compare two psdf matrices?

let $A, B$ be positive semidenite matrices

**definition:** we say $A \succeq B$ ($A$ is greater than $B$ in matrix sense) if

$$A - B \succeq 0$$

example: $A = \begin{bmatrix} 5 & 1 \\ 1 & 3 \end{bmatrix} \succeq 0, \quad B = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix} \succeq 0, \quad A - B = \begin{bmatrix} 3 & 2 \\ 2 & 2 \end{bmatrix} \succeq 0$

however, $A$ and $B$ are not comparable if $A - B \not\succeq 0$ (and denoted by $A \not\succeq B$)

$$A = \begin{bmatrix} 4 & -1 \\ -1 & 2 \end{bmatrix} \succeq 0, \quad B = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix} \succeq 0, \quad A - B = \begin{bmatrix} 1 & -2 \\ -2 & 1 \end{bmatrix} \not\succeq 0$$

(such relation is called **partial ordering**)

a necessary condition for $A \succeq B$ is that $\mathbf{diag}(A) \succeq \mathbf{diag}(B)$

# Congruent transformation

let $A$ be a symmetric matrix and $B$ be any invertible matrix

**definition:** a transformation $f : \mathbf{S}^n \rightarrow \mathbf{S}^n$ given by

$$f(A) = B^T AB$$

is said to be **congruent** to $A$ and has the following properties: **law of inertia**

1. $B^T AB$ has the same number of (positive)(negative)(zero) eigenvalues as $A$ (proof in Strang page 177)
2. for a special case when $A \succ 0$, the result is clear, *i.e.*,

$$B^T AB \succ 0 \quad \Longleftrightarrow \quad A \succ 0, \quad \text{provided that } B \text{ is invertible}$$

example: let $X$ be a random vector and $Y = BX$; then $\mathbf{cov}(Y) = B\,\mathbf{cov}(X)B^T$

# Positive semidefinite ordering

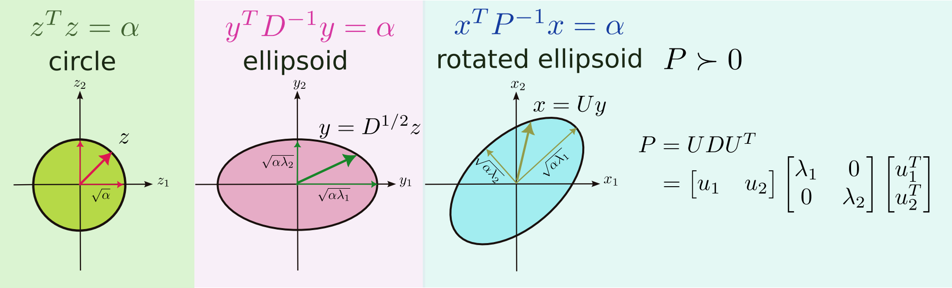
1. if $A \succeq B$ then $A^{-1} \preceq B^{-1}$ (provided that $A, B$ are invertible)
2. $\lambda_{\max}(A)I \succeq A \succeq \lambda_{\min}(A)I$
3. if $A \succeq B$ then $S^T AS \succeq S^T BS$ for any $S$

- proof of [1] involves spectral radius and singular value of matrices (see detail in Horn, Corollary 7.7.4 page 495)
- proof of [2] and [4] are straightforward; just use the definition

# Ellipsoid in $\mathbf{R}^n$

given $P \succ 0, x_c \in \mathbf{R}^n, \alpha > 0$, an ellipsoid in $\mathbf{R}^n$ is parametrized by

$$\mathcal{E} = \{ \, x \in \mathbf{R}^n \mid (x - x_c)^T P^{-1} (x - x_c) \leq \alpha \, \}$$



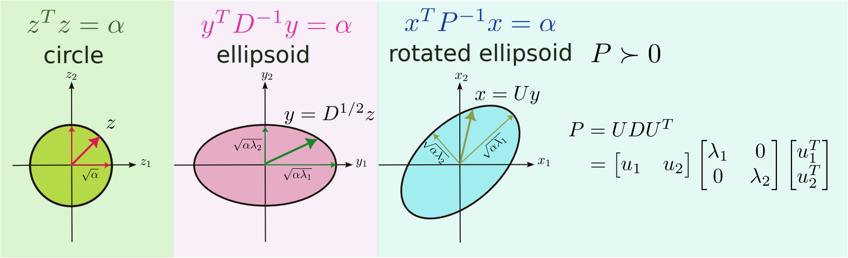

$P \succ 0$ has an eigenvalue decomposition: $P = UDU^T$

1. principal axes of ellipsoids are eigenvectors of $P$: $u_1, u_2, \ldots, u_n$
2. the widths of principal axes are $\sqrt{\alpha \lambda_i}$ where $\lambda_i$'s are eigenvalues of $P$

# How to sketch an ellipsoid

ingredients:

- $P = UDU^T \Rightarrow P^{-1} = UD^{-1}U^T$ where $D = \mathbf{diag}(\lambda_1, \ldots, \lambda_n)$
- $U$ is unitary, *i.e.*, $U^TU = I$ and if $x = Uy$ then $\|x\| = \|y\|$

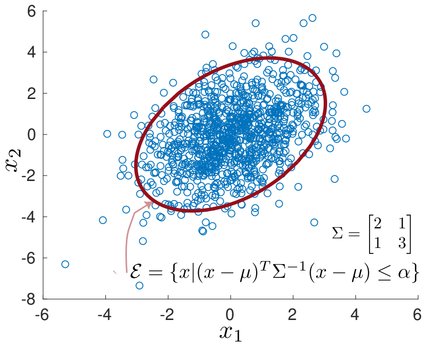



- R to L: $x^TP^{-1}x = x^TUD^{-1}U^Tx = x^TD^{-1/2}D^{-1/2}U^Tx$ and make transformations $y = U^Tx$ and $z = D^{-1/2}y$
- L to R: plot shape in $z$ (easy), scale/dilate $z$ to get shape in $y$, and rotate $y$ to get the shape in $x$

# Ellipsoid as in Gaussian confidence region

**basic facts:** suppose $X$ is Gaussian with covariance $\Sigma_x$

- if $Z = AX + b$ (affine) then $Z$ is also Gaussian with covariance $\Sigma_z = A\Sigma_x A^T$
- for $X \sim \mathcal{N}(0, \Sigma)$ and if $\Sigma = UDU^T$ then $Z = D^{-1/2}U^T X$ is a standard Gaussian
- sum square of $n$ standard Gaussians is a Chi-square of $n$ degree of freedom




- $x \sim \mathcal{N}(0, \Sigma)$ and transform $x$ to $z$
- decompose $\Sigma = UDU^T$ and transform $z = D^{-1/2}U^T x$ to make $\mathbf{cov}(z) = I$

$$P(x^T \Sigma^{-1} x \leq \alpha) = P(z^T z \leq \alpha) = P(\mathcal{X}_n^2 \leq \alpha)$$

- size of ellipsoid ($\alpha$) is computed to guarantee that $P(x \in \mathcal{E}) \geq$ a desired value

$$\alpha = F_{\chi^2}^{-1}(0.9)$$

# Schur complement

a consider a block matrix $X$ partitioned as

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

- Schur complement of $D$ in $X$ is defined as

$$S = A - BD^{-1}C, \quad \text{if} \;\; \det D \neq 0$$

we can show that $\det X = \det D \det S$

| | | | |
|---|---|---|---|
| 7 | 1 | 0 | 3 |
| 1 | 4 | 1 | 5 |
| 0 | 1 | 2 | -2 |
| 3 | 5 | -2 | 9 |

- Schur complement of $A$ in $X$ is defined as

$$S = D - CA^{-1}B, \quad \text{if} \;\; \det A \neq 0$$

we can show that $\det X = \det A \det S$

| | | | |
|---|---|---|---|
| 7 | 1 | 0 | 3 |
| 1 | 4 | 1 | 5 |
| 0 | 1 | 2 | -2 |
| 3 | 5 | -2 | 9 |

# How Schur complement arises in Gaussian elimination

consider a system of linear equations in two-block variables and get rid of $x_2$ first

$$Ax_1 + Bx_2 = y_1, \quad Cx_1 + Dx_2 = y_2$$

if $D^{-1}$ exists, we can eliminate $x_2$ first; $x_2 = D^{-1}y_2 - D^{-1}Cx_1$

plug $x_2$ in the first equation and solve for $x_1$

$$Ax_1 + B(D^{-1}y_2 - D^{-1}Cx_1) = y_1 \quad \Rightarrow \; (A - BD^{-1}C)x_1 = y_1 - BD^{-1}y_2$$

denote $S = A - BD^{-1}C$ and if it is invertible, ✎ the solution is given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} S^{-1}y_1 - S^{-1}BD^{-1}y_2 \\ -D^{-1}CS^{-1}y_2 + (D^{-1} + D^{-1}CS^{-1}BD^{-1})y_2 \end{bmatrix}$$

# Inverse of block matrix

express the solution $(x_1, x_2)$ as a formula for the inverse of a block matrix

$$X^{-1} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} S^{-1} & -S^{-1}BD^{-1} \\ -D^{-1}CS^{-1} & D^{-1} + D^{-1}CS^{-1}BD^{-1} \end{bmatrix}$$

❋ note that the Schur complemnt is the inverse of the $(1,1)$ block of $X^{-1}$!

in fact, an LDU decomposition of $X$ is

$$X = \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}$$

this proves that the determinant of $X$ is $\det(A - BD^{-1}C) \det D$

# Schur complement of positive semidefinite matrix

$$X = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}, \quad S_D = A - BD^{-1}B^T, \quad S_A = D - B^TA^{-1}B,$$

**facts:**

- $X \succ 0$ if and only if $D \succ 0$ and $S_D \succ 0$
- if $D \succ 0$ then $X \succeq 0$ if and only if $S_D \succeq 0$
- $\det X = \det D \det S_D = \det A \det S_A$

$$X = \underbrace{\begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix}}_{\text{full rank}} \begin{bmatrix} S_D & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I & 0 \\ D^{-1}B^T & I \end{bmatrix}$$

a form of congrugent transformation

interesting meaning when $X \succ 0$: we have $S_D \succ 0$ and $D \succ 0$

$$A - S_D = BD^{-1}B^T \succeq 0 \quad \Longleftrightarrow \quad A \text{ is bigger than } S_D \text{ !}$$

analogous results for $S_A$

- $X \succ 0$ if and only if $A \succ 0$ and $S_A \succ 0$
- if $A \succ 0$ then $X \succeq 0$ if and only if $S_A \succeq 0$

# Applications of Schur complement

| | | | |
|---|---|---|---|
| 7 | 1 | 0 | 3 |
| 1 | 4 | 1 | -2 |
| 0 | 1 | 2 | -2 |
| 3 | -2 | -2 | 9 |

- conditional covariance matrix of $X|Y$ (Gaussian case)

$$\Sigma = \begin{bmatrix} \Sigma_x & \Sigma_{xy} \\ \Sigma_{xy}^T & \Sigma_y \end{bmatrix}, \quad \Sigma_{x|y} = \begin{bmatrix} 7 & 1 \\ 1 & 4 \end{bmatrix} - \begin{bmatrix} 0 & 3 \\ 1 & -2 \end{bmatrix} \begin{bmatrix} 2 & -2 \\ -2 & 9 \end{bmatrix}^{-1} \begin{bmatrix} 0 & 3 \\ 1 & -2 \end{bmatrix}^T$$

  (clearly, $\Sigma_{x|y} \preceq \Sigma_x$ – if $\Sigma_{xy} \neq 0$, knowing $Y$ helps reduce covariance in $X$)
- elimination of variable in solving a linear system
- inverse of block matrix

# Matrix inversion lemmas

**Woodbury formula:** let $A$ be invertible and let $C, U, V$ be rectangular matrices

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

(useful when $k < n$ or that $U$ is tall and $V$ is fat giving $C^{-1} + VA^{-1}U$ in smaller size than $n$)

**Sherman-Morrison formula:** when $U, V$ reduce to outer product of vectors

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

(useful when $A^{-1}$ is simple – the denominator in RHS turns to be scalar)

the inverse of perturbation of $A$ corrected by a low-rank update is obtained by a cheap perturbation of $A^{-1}$

# Example of matrix inversion lemma

recall that the inverse of a diagonal matrix $D = \mathbf{diag}(d)$ is $D^{-1} = \mathbf{diag}(1/d)$ (simple)

$$\left( \begin{bmatrix} 2 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 4 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \\ 2 \end{bmatrix} \begin{bmatrix} 0 & -3 & 1 \end{bmatrix} \right)^{-1} =$$

compare the matrix inversion result with the direct calculation

when the dimension of $u, v$ is large, and if $A$ is diagonal

- $A^{-1}$ is obtained as cheaply as $\mathcal{O}(n)$
- calculations of $v^T A^{-1} u$ and $A^{-1} u v^T A^{-1}$ are also in $\mathcal{O}(n)$

# Push-through identity

let $A \in \mathbf{R}^{m\times n}, B \in \mathbf{R}^{n\times m}$ and assume that $I + AB$ is invertible

**facts:** ✎

- $I + BA$ is invertible
- **push-through identity**

$$B(I + AB)^{-1} = (I + BA)^{-1}B$$

($B$ is pushed from the left to right)

hint: start with $B(I + AB) = (I + BA)B$

# Pseudo-inverse

consider a full rank matrix $A \in \mathbf{R}^{m \times n}$ in three cases

- **tall matrix:** $A$ is full rank $\Leftrightarrow$ columns of $A$ are LI $\Leftrightarrow$ $A^T A$ is invertible

$$((A^T A)^{-1} A^T) A = (A^T A)^{-1} (A^T A) = I$$

  the **pseudo-inverse** of $A$ (or left-inverse) is $A^\dagger = (A^T A)^{-1} A^T$

- **wide matrix:** $A$ is full rank $\Leftrightarrow$ row of $A$ are LI $\Leftrightarrow$ $AA^T$ is invertible

$$A(A^T (AA^T)^{-1}) = (AA^T)(AA^T)^{-1} = I$$

  the **pseudo-inverse** of $A$ (or right-inverse) is $A^\dagger = A^T (AA^T)^{-1}$

- **square matrix:** $A$ is full rank $\Leftrightarrow$ $A$ is invertible and both formula of pseudo-inverses reduce to the ordinary inverse $A^{-1}$

✎ the pseudo inverses of the three cases have the same dimension ?

# Symmetry in the complex world$^{x+iy}$

let $A \in \mathbf{C}^{n\times n}$ and denote the operator $A^*$ as

$$A^* = \bar{A}^T \quad \text{(complex conjugate transpose)}$$

**definition:** $A$ is said to be **Hermittian** or **self-adjoint** if $A^* = A$

example: $\begin{bmatrix} 2 & 3-2i \\ 3+2i & 1 \end{bmatrix}$ clearly see that $A^* = A \iff a_{ij} = \bar{a}_{ji}$

**facts:** if $A$ is self-adjoint

- eigenvalues of self-adjoint matrix are real
- eigenvectors are mutually orthogonal
- $A$ admits a decomposition: $A = UDU^*$ where $U$ is unitary, *e.g.*, $U^*U = UU^* = I$

# References


1. S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least Squares*, Cambridge, 2018
2. G. Strang, *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press, 2019
3. R.A. Horn and C.R. Johnson, *Matrix Analaysis*, 2nd Edition, Cambridge, 2012
4. C. C. Aggrawal, *Linear algebra and optimization for machine learning:A textbook*, Springer, 2020

# Normed vector space and Inner product space

# Vector space

a vector space or linear space (over **R**) consists of

- a set $\mathcal{V}$
- a vector sum $+ : \mathcal{V} \times \mathcal{V} \to \mathcal{V}$
- a scalar multiplication : $\mathbf{R} \times \mathcal{V} \to \mathcal{V}$
- a distinguished element $0 \in \mathcal{V}$

$\mathcal{V}$ is called a vector space over **R**, denoted by $(\mathcal{V}, \mathbf{R})$ if elements, called *vectors* of $\mathcal{V}$ satisfy the following main operations:

1. **vector addition:**

$$x, y \in \mathcal{V} \quad \Rightarrow \quad x + y \in \mathcal{V}$$

2. **scalar multiplication:**

$$\text{for any } \alpha \in \mathbf{R}, x \in \mathcal{V} \quad \Rightarrow \quad \alpha x \in \mathcal{V}$$

# Example of vector spaces

- $\mathbf{R}^n$, $\mathbf{R}^{m \times n}$
- set of polynomials of degree less than or equal to $n$
- set of continuous functions on $(a, b)$

$\mathcal{M}$ is called a **subspace** of vector space $\mathcal{V}$ if $\mathcal{M}$ is a subset of $\mathcal{V}$, and $\mathcal{M}$ is a vector space itself

**examples:**

- $\{x \in \mathbf{R}^n \mid x_1 = 0\}$
- set of diagonal matrices of size $n \times n$
- range space and nullspace of a matrix $A$

# Normed vector space

a **normed linear space** is a vector space $\mathcal{V}$ over a $\mathbf{R}$ with a map

$$\|\cdot\| : \mathcal{V} \to \mathbf{R}$$

called a **norm** that satisfies

- homogenity

$$\|\alpha x\| = |\alpha|\|x\|, \qquad \forall x \in \mathcal{V}, \forall \alpha \in \mathbf{R}$$

- triangle inequality

$$\|x + y\| \leq \|x\| + \|y\|, \qquad \forall x, y \in \mathcal{V}$$

- positive definiteness

$$\|x\| \geq 0, \quad \|x\| = 0 \iff x = 0, \qquad \forall x \in \mathcal{V}$$

# Example of vector and matrix norms

$x \in \mathbf{R}^n$ and $A \in \mathbf{R}^{m\times n}$

- 2-norm (Euclidean norm)

$$\|x\|_2 = \sqrt{x^Tx} = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

$$\|A\|_F = \sqrt{\mathbf{tr}(A^TA)} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2}$$

- 1-norm

$$\|x\|_1 = |x_1| + |x_2| + \cdots + |x_n|, \quad \|A\|_1 = \textstyle\sum_{ij} |a_{ij}|$$

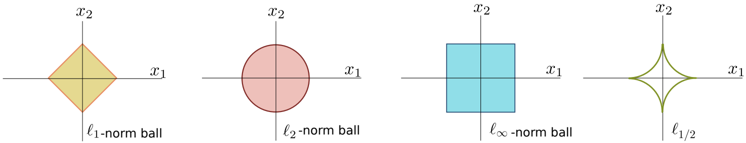
- $\infty$-norm

$$\|x\|_\infty = \max_k\{|x_1|, |x_2|, \ldots, |x_n|\}, \quad \|A\|_\infty = \max_{ij} |a_{ij}|$$

clearly, $\|x\|$ measures the vector size; $\|x - y\|$ measures the distance between $y$ and $x$

# $\ell_p$-norm

$$\|x\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$$




- a unit-norm ball is the set $\{x \in \mathbf{R}^n \mid \|x\| \leq 1\}$
- $\ell_0$ is defined as $\|x\|_0 = \mathbf{card}(x)$ (the number of nonzero elements in $x$)
- $\ell_{1/2}$ is NOT a norm due to violation of triangle inequality

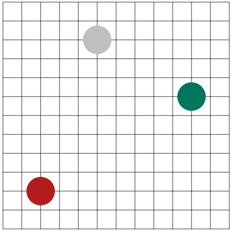
$$x = (1,0), y = (0,1), \quad \|x\|_{1/2} = \|y\|_{1/2} = 1, \quad \text{but} \quad \|x+y\|_{1/2} = \|(1,1)\|_{1/2} = 2^2$$

- $\ell_0, \ell_{1/2}$ are **not truly a norm**; in fact, $\ell_p$ is a norm when $1 \leq p < \infty$

# Norm as a distance function

for $\mathbf{R}^n$, we can use different norms to measure the distance between $x$ and $y$

✎ mark the distance between **red** and **green** dots using



distance function induced by different norms

- $\ell_1$-norm: **Manhattan/taxicab distance**
- $\ell_2$-norm: **Euclidean distance**
- $\ell_p$-norm: **Minkowski distance** for $p \geq 1$
- $\ell_\infty$-norm: **Chebyshev distance**

- a distance value should be non-negative
- the distance from $x$ to $y$ should be the same as measuring from $y$ to $x$

a distance function can be formulated mathematically as the idea of a **metric**

# Metric space

a **metric** is a function $d : \mathcal{X} \times \mathcal{X} \to \mathbf{R}_+$ that gives a distance meaning of two points

a metric (or distance function) must satisfy the three properties for all $x, y \in \mathcal{X}$

1. $d(x, y) = 0$ if and only if $x = y$ (definiteness)
2. $d(x, y) = d(y, x)$ (symmetry)
3. $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

**definition:** any set $\mathcal{X}$ that is equipped with a matric is called a **metric space** $(\mathcal{X}, d)$

- any normed linear space $(\mathcal{V}, \|\cdot\|)$ is then a metric space with the distance function $d(x, y) := \|x - y\|$
- the triangle inequality is satisfied by following

$$d(x, z) := \|x - z\| = \|x - y + y - z\| \leq \|x - y\| + \|y - z\| = d(x, y) + d(y, z)$$

# Further reading about distance

1. let $\mathcal{X}$ be a metric space and $\mathcal{M} \subset \mathcal{X}$ and $x \in \mathcal{X}$

$$\mathbf{dist}(\mathcal{M}, x) = \inf_{z \in \mathcal{M}} d(z, x)$$

   (the distance between **a set and a point** – taking the minimum distance)

2. let $\mathcal{C}$ and $\mathcal{D}$ be two subsets of a metric space $\mathcal{X}$ – the distance between **two sets** is

$$\mathbf{dist}(\mathcal{C}, \mathcal{D}) = \inf_{x \in \mathcal{C}, y \in \mathcal{D}} d(x, y)$$

$$\mathbf{dist}(\mathcal{C}, \mathcal{D}) = \inf_{x \in \mathcal{C}, y \in \mathcal{D}} \|x - y\| \quad \text{if the distance is induced from a norm}$$

3. measure error between two inputs: given any two vectors $x, y$ or matrices $A, B$, to compare if $x = y$ or $A = B$ (mathematically) we should check numerically that

$$\|x - y\| \leq \epsilon, \quad \|A - B\| \leq \epsilon \qquad \text{(choice of norm may affect the computation)}$$

# Applications of vector norms

questions involving norms

- find a vector $x$ having the smallest norm (measured by any norm choice) while $x$ stays in a set (hyperplane, convex sets)

$$\underset{x}{\text{minimize}} \quad \|x\| \quad \text{subject to} \quad Ax = y$$

- we can choose several choices of distance functions in kNN to measure the $k$-nearest neighbors
- $\ell_2$-norm (as MSE) and $\ell_1$-norm (as MAE) are typical loss functions $\rho$ in regression problems

$$\underset{\theta}{\text{minimize}} \quad \sum_{i=1}^{N} \rho(y_i - f(x_i;\theta))$$

where $\rho(r)$ can be $|r|, r^2$

# Separable property

$x =$

| | | | | | |
|---|---|---|---|---|---|
| 1 | -2 | 0 | 3 | -5 | 4 |

| | | | | | |
|---|---|---|---|---|---|
| 1 | -2 | 0 | 3 | -5 | 4 |

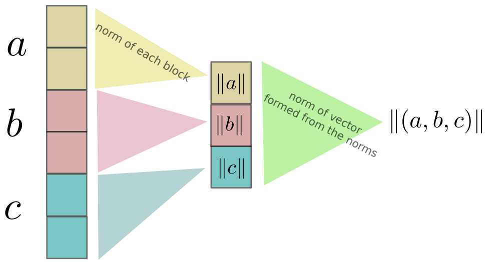
$\triangleq x = (x_1, x_2, x_3), \quad x_k \in \mathbf{R}^2$

let's verify that

- $\|x\|_2^2 = \|x_1\|_2^2 + \|x_2\|_2^2 + \|x_3\|_2^2$
- $\|x\|_1 = \|x_1\|_1 + \|x_2\|_1 + \|x_3\|_1$
- $\|x\|_\infty = \max_{i=1,2,3} \{\|x_1\|_\infty, \|x_2\|_\infty, \|x_3\|_\infty\}$

in fact, $\ell_p$-norm of a stacked vector is ✎

$$\|(a,b,c)\|_p = \|(\|a\|_p, \|b\|_p, \|c\|_p)\|_p$$



# Operator norm

**matrix operator norm** of $A \in \mathbf{R}^{m \times n}$ is defined as

$$\|A\| = \max_{\|x\| \neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

aka as the **induced norm**

**properties:**

1. for any $x$, $\|Ax\| \leq \|A\|\|x\|$ (by the definition)
2. $\|aA\| = |a|\|A\|$ (scaling)
3. $\|A + B\| \leq \|A\| + \|B\|$ (triangle inequality)
4. $\|A\| = 0$ if and only if $A = 0$ (positiveness)
5. $\|AB\| \leq \|A\|\|B\|$ (submultiplicative)

# Examples of operator norms

- **2-norm** (aka as **spectral norm**)

$$\|A\|_2 \triangleq \max_{\|x\|_2=1} \|Ax\|_2 = \sqrt{\lambda_{\max}(A^T A)} = \sigma_{\max}(A) \quad \text{(max singular value)}$$

- **1-norm**

$$\|A\|_1 \triangleq \max_{\|x\|_1=1} \|Ax\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|$$

| | | | |
|---|---|---|---|
| 1 | -2 | 0 | -3 |
| 0 | 3 | 1 | 2 |
| 5 | 0 | 2 | -2 |
| 0 | 7 | 8 | 0 |

- **$\infty$-norm**

$$\|A\|_\infty \triangleq \max_{\|x\|_\infty=1} \|Ax\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

✎ verify that the above operator norms have the given expressions

# More on metric norms

- **nuclear norm:** sum of singular values (no. of nonzero $\sigma_i$ determines $\mathbf{rank}(X)$)

$$\|X\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(X)$$

  (recall a singular value is $\sigma_i(X) = \sqrt{\lambda_i(X^T X)}$)

- **spectral radius $\rho(X)$:** let $\lambda_1, \ldots, \lambda_n$ be $n$ eigenvalues of $X$

$$\rho(X) = \max_k \{ \, |\lambda_1|, |\lambda_2|, \ldots, |\lambda_n| \, \}$$

  ✎ spectral radius is NOT a norm ☞ check which norm condition is violated

- useful relations ✎: $\rho(A) \leq \|A\|_2 \leq \|A\|_F \leq \|A\|_*$

  proof hint: definition of operator norm ; max eigenvalue $<$ sum of eigenvalue ;
  $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$

# Applications of matrix norms

1. analog of least-squares for matrix parameter: minimize$_X$ $\|Y - HX\|_F^2$
2. deriving norm of output from a matrix-vector multiplication

$$\begin{aligned} x(t+1) = Ax(t) \Rightarrow x(t) &= A^t x(0) \\ &\Rightarrow \|x(t)\| \leq \|A\| \|A^{t-1}x(1)\| \leq \cdots \leq \|A\|^t \|x(0)\| \end{aligned}$$

   the inquality is obtained by the matrix operator norm
3. let $S = A^T A$, the maximum of $R(x) = \frac{x^T S x}{x^T x}$ is called the **Rayleigh quotient** which turns out to be the squared spectral norm of $A$, $\sigma_{\max}^2(A)$
4. low-rank approximation: minimize $\|A - X\|_F^2$ subject to $\mathbf{rank}(X) \leq r$ (find a low-rank $X$ that best approximates $A$ in Frobenius norm sense)
5. problem: minimize $f(X) + \lambda \|X\|_*$ (a regularized regression with parameter $X$ that has a low-rank prior)

# Equivalence of norms

two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ on a vector space $\mathcal{V}$ are said to be **equivalent**

if there exists constants $\alpha, \beta$ such that

$$\alpha\|x\|_A \leq \|x\|_B \leq \beta\|x\|_A, \quad \forall x \in \mathcal{V}$$

examples: $\ell_1, \ell_2, \ell_\infty$-norms for $x \in \mathbf{R}^n$ are all equivalent ✎

$$\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \leq n\|x\|_\infty$$

(non-trivial: prove $\|x\|_\infty \leq \|x\|_2$ using Cauchy-Swarz inequality with $y = e_j$ making $y^T x = \|x\|_\infty$)

**applications:** for an error $e \in \mathbf{R}^N$, $\text{MSE} = \frac{1}{N}\|e\|_2^2$, $\text{RMSE} = \frac{1}{\sqrt{N}}\|e\|_2$, $\text{MAE} = \frac{1}{N}\|e\|_1$

$$\text{MAE} \leq \text{RMSE} \leq \sqrt{N}\text{MAE}$$

which bound is useful ? – meaning that it provides a *tight* upper/lower bound

# Inner product space

an inner product space is a vector space $\mathcal{V}$ over $\mathbf{R}$ with a map

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathbf{R}$$

for all $x, y, z \in V$ and all scalars $a \in \mathbf{R}$, an inner product satisfies

1. symmetry: $\langle x, y \rangle = \langle y, x \rangle$
2. linearity in the first argument:

$$\langle ax, y \rangle = a\langle x, y \rangle, \quad \langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$$

3. positive definiteness

$$\langle x, x \rangle \geq 0, \quad \text{and} \quad \langle x, x \rangle = 0 \Leftrightarrow x = 0$$

# Examples of inner product spaces

- $\mathbf{R}^n : \langle x, y \rangle = y^T x = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$ (canonical/vanilla inner product)
- $\mathbf{R}^n$ : for $W \succ 0$, $\langle x, y \rangle_W = y^T W x$ (weighted inner product)

  ($W \succ 0$ is a positive definite matrix, *i.e.*, $x^T W x > 0$ for all $x \neq 0$)
- $\mathbf{R}^{m \times n} : \langle X, Y \rangle = \mathbf{tr}(Y^T X)$
- $C[a, b]$: set of all real-valued continuous functions on $[a, b]$ whose inner product is defined as

$$\langle f, g \rangle = \int_a^b f(t) g(t) dt$$

# Applications of inner product in $\mathbf{R}^n$

the inner product $x^T y$ has a meaningful interpretation in applications

- co-occurrence: let $a, b$ are $n$-vectors that describe occurrence, *i.e.*, each elements is either $0$ or $1$; then $a^T b$ gives the total number of indices for which $a_i$ and $b_i$ are both one
- score/weight/feature: $s = w^T f$ where $f$ is a feature vector, $w$ is the weight vector, and $s$ is the total score
- probability/expected value: expected value $= f^T p$ where $p$ is a probabability vector, and $f_i$ is the value if outcome $i$ occur
- polynomial evaluation: $p(x) = c_0 + c_1 x + \cdots + c_n x^n$ then we can present $p(t) = c^T z$ where $c = (c_0, \ldots, c_n)$ and $z = (1, t, \ldots, t^n)$

# Induced norm

every inner product space induces a norm that is defined by

$$\|x\| \triangleq \sqrt{\langle x, x\rangle} \qquad \text{(satisfy all properties of norm)}$$

Cauchy-Schwarz inequality: $|\langle x, y\rangle| \leq \|x\|\|y\|$

✎ show that the induced norm satisfies the triangle inequality

$$\begin{aligned}
\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x, x\rangle + \langle y, y\rangle + \langle x, y\rangle + \langle y, x\rangle \\
&= \|x\|^2 + \|y\|^2 + 2\Re\langle x, y\rangle \leq \|x\|^2 + \|y\|^2 + 2|\langle x, y\rangle| \\
&\leq \|x\|^2 + \|y\|^2 + 2\|x\|\|y\| = (\|x\| + \|y\|)^2
\end{aligned}$$

(the last inequality follows from Cauchy-Schwarz inequality)

✎ if $\langle x, y\rangle = y^T W x$ is used for the inner product, what is the induced norm ?

# Cauchy-Schwarz inequality (CS)

for any $x, y$ in an inner product space $(\mathcal{V}, \mathbf{R})$

$$|\langle x, y\rangle| \leq \|x\|\|y\|$$

moreover, for $y \neq 0$,

$$\langle x, y\rangle = \|x\|\|y\| \quad \Longleftrightarrow \quad x = cy, \quad \exists c \in \mathbf{R}$$

**proof of non-trivial case** $(y \neq 0)$: for any scalar $\alpha$

$$0 \leq \|x + \alpha y\|^2 = \|x\|^2 + \alpha^2\|y\|^2 + 2\alpha\langle x, y\rangle$$

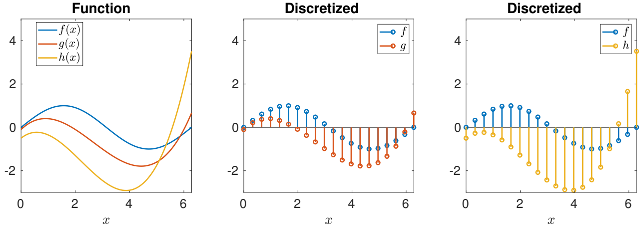
if $y \neq 0$, then we can choose $\alpha = -\dfrac{\langle x, y\rangle}{\|y\|^2}$ and the CS inequality follows

interpretation as **cosine similarity:** $-1 \leq \cos\theta \triangleq \frac{\langle x,y\rangle}{\|x\|\|y\|} \leq 1$

# Cosine similarity function

let's find the similarity between $f(x) = \sin(x) \in C[0, 2\pi]$ and each of two polynomials:

$$g(x) = 0.1x^3 - 0.8x^2 + 1.2x - 0.1, \quad h(x) = 0.15x^3 - x^2 + x - 0.5$$




- similarity between $f(x)$ and $g(x)$: $\dfrac{\int_0^{2\pi} \sin(x) g(x) dx}{\sqrt{\int_0^{2\pi} \sin(x) dx \cdot \int_0^{2\pi} g(x) dx}}$
- after discretizing $f(x)$ to a vector $f \in \mathbf{R}^n$, the similarity index is computed using inner product in $\mathbf{R}^n$: similarity $= \dfrac{f^T g}{\|f\|_2 \|g\|_2}$

# Orthogonality

let $(\mathcal{V}, \mathbf{R})$ be an inner product space

- $x$ and $y$ are **orthogonal**:

$$x \perp y \quad \Longleftrightarrow \quad \langle x, y \rangle = 0$$

- **orthogonal complement** in $\mathcal{V}$ of $S \subseteq \mathcal{V}$, denoted by $S^\perp$, is defined by

$$S^\perp = \{x \in \mathcal{V} \mid \langle x, s \rangle = 0, \;\; \forall s \in S\}$$

**fact:** ✎ $S^\perp$ is a vector space

- for $\mathcal{M} \subseteq \mathbf{R}^n$, $\mathbf{R}^n$ admits the **orthogonal decomposition**:

$$\mathbf{R}^n = \mathcal{M} \oplus \mathcal{M}^\perp, \quad \text{and} \;\; \dim(\mathbf{R}^n) = \dim(\mathcal{M}) + \dim(\mathcal{M}^\perp)$$

any $y \in \mathbf{R}^n$ is uniquely decomposed as $y = m + \tilde{m}$ where $m \in \mathcal{M}$ and $\tilde{m} \in \mathcal{M}^\perp$

# Examples of orthogonality

these are orthogonal pairs

$$(1,0,-1) \perp (1,1,1), \quad \begin{bmatrix} 1 & 0 \\ -2 & 3 \end{bmatrix} \perp \begin{bmatrix} 1 & 1 \\ 0 & -1/3 \end{bmatrix}, \quad C[0,1] : x \perp (4x^2 - 2)$$

✎ please verify

- $S = \{ x \in \mathbf{R}^n \mid a^T x = 0 \}$ and $S^\perp = \mathrm{span}\{a\}$
- $S = \left\{ A \in \mathbf{R}^{2\times 2} \mid A = \begin{bmatrix} 0 & a_{12} \\ a_{21} & 0 \end{bmatrix} \right\}$ and $S^\perp = \left\{ B \in \mathbf{R}^{2\times 2} \mid B = \begin{bmatrix} b_{11} & 0 \\ 0 & b_{22} \end{bmatrix} \right\}$
- $S = \mathrm{span}\{(1,0,0)\}$ and $S^\perp = \mathrm{span}\{(0,1,0),(0,0,1)\}$

$$\mathbf{R}^3 = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right\} \oplus \mathrm{span}\left\{ \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

# Parallogram law

we start with $x, y$ in an inner product space and $\|\cdot\|$ is the induced norm

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle = \langle x,x\rangle + \langle y,y\rangle + \langle x,y\rangle + \langle y,x\rangle \\ \|x-y\|^2 &= \langle x-y, x-y\rangle = \langle x,x\rangle + \langle y,y\rangle - \langle x,y\rangle - \langle y,x\rangle\end{aligned}$$

- **Pythagoras' theorem:** when $x \perp y$, squared norm of the sum reduces to

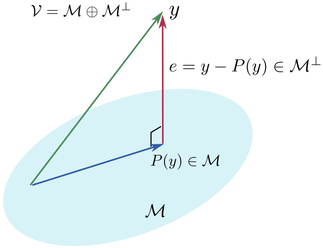
$$\|x+y\|^2 = \|x\|^2 + \|y\|^2$$

- the **parallelogram law:** by adding the above two identities

$$2\|x\|^2 + 2\|y\|^2 = \|x+y\|^2 + \|x-y\|^2$$

# Orthogonal projection

let $x, y$ be vectors in an inner product space $\mathcal{V}$ equipped with $\langle \cdot, \rangle$ and let $\mathcal{M} \subseteq \mathcal{V}$




**orthogonal projection of $y$ onto $\mathcal{M}$**

**definition:** find a mapping $P : \mathcal{V} \to \mathcal{M}$ such that

$$e = y - P(y)$$

is orthogonal to any vector in $\mathcal{M}$

✍ concept of orthogonality depends on the inner product associated with $\mathcal{V}$

**orthogonality condition:** $y - P(y) \perp \mathcal{M}$

# Procedure of finding the orthogonal projection of $y$ onto $\mathcal{M}$

- let $\{\phi_1, \phi_2, \ldots, \phi_m\}$ be a basis for $\mathcal{M}$
- $P(y)$ must be a linear combination of $\phi_k$'s (since $\mathcal{R}(P) \subseteq \mathcal{M}$)

$$P(y) = a_1\phi_1 + \cdots + a_2\phi_m$$

- $y - P(y) \perp \mathcal{M} \iff \langle y - P(y), \phi_k \rangle = 0$ for all $k$ and it gives

**orthogonality condition:** $$\begin{aligned} \langle y, \phi_k \rangle &= \langle P(y), \phi_k \rangle, \quad k = 1, 2, \ldots, m \\ &= \langle a_1\phi_1 + a_2\phi_2 + \cdots + a_m\phi_m, \phi_k \rangle \end{aligned}$$

this forms a system of $m$ linear equations in $a_k$'s

example: if $\mathcal{M}$ has only one basis vector $\phi$, we have $\langle y, \phi \rangle = a_1 \langle \phi, \phi \rangle$

# Procedure of finding the orthogonal projection of $y$ onto $\mathcal{M}$

solve $m$ linear equations to find coefficients $a_k$

$$\begin{bmatrix} \langle \phi_1, \phi_1 \rangle & \langle \phi_2, \phi_1 \rangle & \cdots & \langle \phi_m, \phi_1 \rangle \\ \langle \phi_1, \phi_2 \rangle & \langle \phi_2, \phi_2 \rangle & \cdots & \langle \phi_2, \phi_m \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \phi_m, \phi_1 \rangle & \langle \phi_m, \phi_2 \rangle & \cdots & \langle \phi_m, \phi_m \rangle \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} \langle y, \phi_1 \rangle \\ \langle y, \phi_2 \rangle \\ \vdots \\ \langle y, \phi_m \rangle \end{bmatrix}, \quad \triangleq \quad Ga = b$$

- $G$ with $g_{ij} = \langle \phi_i, \phi_j \rangle$ is called a **Gram matrix** (clearly symmetric and can be shown to be positive definite)
- for this reason, $G$ is invertible and $a = G^{-1}b$
- $b$ is linear in $y$, it is clear that $P(y) = a_1\phi_1 + \cdots + a_m\phi_m$ is then linear in $y$

# Projection onto a vector

if a basis for $\mathcal{M}$ is $\{\phi\}$ (only one basis vector), then $P(y) = a\phi$

$$\langle y, \phi \rangle = a\langle \phi, \phi \rangle \quad \Rightarrow \quad P(y) = \frac{\langle y, \phi \rangle}{\langle \phi, \phi \rangle}\phi$$

1. project $y$ onto $x$ in $\mathbf{R}^n$:

$$P(y) = \alpha x, \qquad P(y) = \frac{\langle x, y \rangle}{\langle x, x \rangle} \cdot x = \frac{(y^T x)x}{\|x\|^2} = \|y\| \cos\theta \cdot \frac{x}{\|x\|}$$

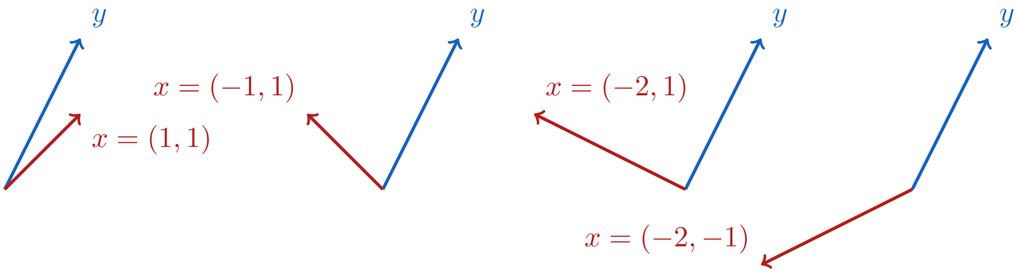
2. project $Y$ onto $X$ in $\mathbf{R}^{m \times n}$:

$$Y = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 3 & -1 \end{bmatrix}, X = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}, \ \langle X, Y \rangle = \mathbf{tr}(Y^T X) = 3, \langle X, X \rangle = \mathbf{tr}(X^T X) = 4$$

$$P(Y) = \frac{\langle X, Y \rangle}{\langle X, X \rangle} \cdot X = \frac{3}{4} X = \frac{3}{4} \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

# Try out the formula

find the projection of $y = (1,2)$ onto the subspace spanned by $x$

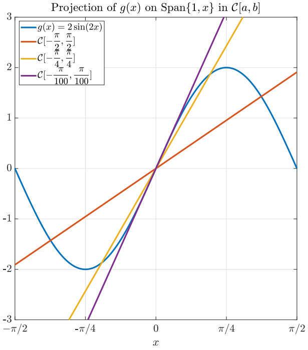


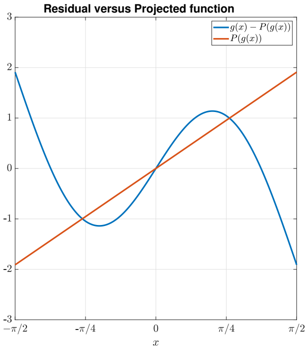

✎ which pair of $(y, x)$ has the highest cosine similarity index?

(review acute/obtuse angles between vectors)

# Projection of a function

example: project $g(x) = 2\sin(2x) \in C[-\frac{\pi}{a}, \frac{\pi}{a}]$ onto a subspace spanned by $\{1, x\}$




inner product:

$$\langle f, g \rangle = \int_{-\pi/a}^{\pi/a} f(t)g(t)dt$$

on $C[-\frac{\pi}{2}, \frac{\pi}{2}], C[-\frac{\pi}{4}, \frac{\pi}{4}]$, and $C[-\frac{\pi}{100}, \frac{\pi}{100}]$

- three projections: $P(g(x))$ are different by the support of function (but all of them are linear in $x$)
- as the support becomes smaller, $P(g(x))$ tends to be the tangent line of $g(x)$ at 0

# Calculations

the orthogonality condition forms a system of 2 equations

$$\begin{bmatrix} \langle 1,1\rangle & \langle 1,x\rangle \\ \langle 1,x\rangle & \langle x,x\rangle \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \langle g(x),1\rangle \\ \langle g(x),x\rangle \end{bmatrix} \quad \Rightarrow \begin{bmatrix} \frac{2\pi}{a} & 0 \\ 0 & \frac{2\pi^3}{3a^2} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \sin(\frac{2\pi}{a}) - \frac{2\pi}{a}\cos(\frac{2\pi}{a}) \end{bmatrix}$$

(as we use the inner product for $C[-\pi/a, \pi/a]$)

$$P(g(x)) = a_1 + a_2 x = \frac{3a^3}{2\pi^3}\left[\sin(2\pi/a) - \frac{2\pi}{a}\cos(2\pi/a)\right]x \quad \triangleq \quad \frac{12}{c^3}[\sin(c) - c\cos(c)]x$$

- $C[-\frac{\pi}{2}, \frac{\pi}{2}]$: the projection is $P(g(x)) = \frac{12}{\pi^2}x$
- $C[-\frac{\pi}{4}, \frac{\pi}{4}]$: the projection is $P(g(x)) = \frac{96}{\pi^3}x$
- ✎ as $a$ is sufficiently large, $P(g(x)) \to 4x$

# Orthogonal complements of range space and nullspace

let $A \in \mathbf{R}^{m \times n}$

✎ verify that

$$\mathcal{R}(A)^{\perp} = \mathcal{N}(A^T), \quad \mathcal{N}(A)^{\perp} = \mathcal{R}(A^T)$$

therefore, we have orthogonal decompositions

$$\mathbf{R}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^T), \quad \mathbf{R}^n = \mathcal{N}(A) \oplus \mathcal{R}(A^T)$$

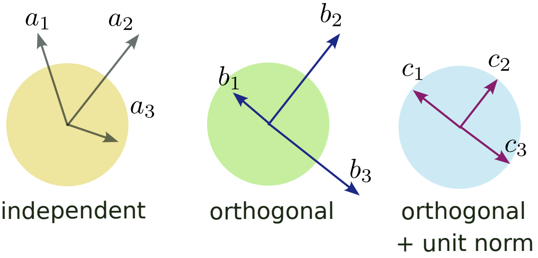
example: $A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \\ -1 & 1 \end{bmatrix}$

$$\mathbf{R}^3 = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \\ 1 \end{bmatrix} \right\} \oplus \mathrm{span}\left\{ \begin{bmatrix} 2 \\ -1 \\ 2 \end{bmatrix} \right\}, \quad \mathbf{R}^2 = \mathrm{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix} \right\} \oplus \{0\}$$

# Linear independence vs Orthogonality

definition: a set $\{\phi_i\}_{i=1}^n \subseteq \mathcal{V}$ can be a basis for $n-$ dimensional vector space $\mathcal{V}$ if

(1) $\mathrm{span}\{\phi_1, \ldots, \phi_n\} = \mathcal{V}$, (2) $\{\phi_1, \ldots, \phi_n\}$ is linearly independent




- $(1, 2, -1), (1, 0, -1), (1, -3, 4)$ are independent but not orthogonal
- $(0, 0, -1), (1, 1, 0), (1, -1, 0)$ are orthogonal and independent

fact: ✎ orthogonal vectors are also independent

# Orthonormal basis

$\{\phi_k\}_{k=1}^n \subset \mathcal{V}$ is said to be an **orthonormal** set if

$$\langle \phi_i, \phi_j \rangle = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and is called an **orthonormal basis** for an $n$-dimensional $\mathcal{V}$ if

1. $\{\phi_k\}_k$ is an orthornomal set
2. $\mathrm{span}\{\phi_1, \phi_2, \ldots, \phi_n\} = \mathcal{V}$

example for $\mathbf{R}^n$:

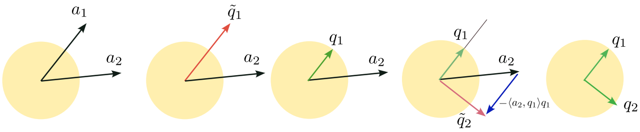
$$\phi_1 = (0, 0, -1), \quad \phi_2 = \frac{1}{\sqrt{2}}(1, 1, 0), \quad \phi_3 = \frac{1}{\sqrt{2}}(1, -1, 0)$$

we can construct an orthonormal basis from the **Gram-Schmidt** orthogonalization

# Gram-Schmidt algorithm (GS)

given vectors $a_1, a_2, \ldots, a_p$, GS algorithm finds orthogonal vectors $q_1, \ldots, q_m$ that

- for $i = 1, \ldots, m$, $a_i$ is a linear combination of $q_1, \ldots, q_m$, and $q_i$ is a linear combination of $a_1, a_2, \ldots, a_i$
- if $a_1, \ldots, a_{j-1}$ are LI but $a_1, \ldots, a_j$ are dependent, GS detects the first vector $a_j$ that is a linear combination of previous $a_1, \ldots, a_{j-1}$




**algorithm:**

1. project vector $a_k$ onto the previous $k-1$ orthonormal vectors
2. $\tilde{q}_k$ is the residual after the projection (hence, must to orthogonal to the previous $a_1, \ldots, a_{k-1}$ vectors
3. normalize $\tilde{q}_k$ to have a unit norm: $q_k := \tilde{q}_k / \|\tilde{q}_k\|$

# Orthogonal expansion

let $\{\phi_i\}_{i=1}^n$ be an orthonormal basis for a vector $\mathcal{V}$ of dimension $n$

for any $x \in \mathcal{V}$, we have the orthogonal expansion:

$$x = \sum_{i=1}^n \langle x, \phi_i \rangle \phi_i$$

meaning: we can project $x$ into orthogonal subspaces spanned by each $\phi_i$

the norm of $x$ is given by

$$\|x\|^2 = \sum_{i=1}^n |\langle x, \phi_i \rangle|^2$$

can be easily calculated by the sum square of projection coefficients

# Kernel 

a kernel $K : [a,b] \times [a,b] \to \mathbf{R}$ is a continuous function with the symmetric property

$$K(x,y) = K(y,x), \qquad \forall x, y \in [a,b]$$

**Mercer's condition:** a real-valued $K(x,y)$ is said to satisfy Mercer's condition if

$$\int\int g(x)K(x,y)g(y)dxdy \geq 0$$

**positive-definite:** $K$ is said to be positive-definite if

$$\sum_{i=1}\sum_{j=1}^{n} K(x_i,x_j)c_ic_j \geq 0, \quad \forall x_i \in [a,b], \ \forall c_i \in \mathbf{R}$$

# Further reading

- open/closed sets, supremum, infimum
- Hölder's inequality (Strang page 96)
- dual norm (see page 637 of Boyd and Vandenberghe 2014)
- composite norms: $x = (x_1, x_2, \ldots, x_K)$ where each $x_i \in \mathbf{R}^p$

$$\|x\|_{p,1} = \sum_{i=1}^{K} \|x_i\|_p$$

- similarity measure:
  - cosine similarity
  - Mahalanobis distance (between a point $x$ and a distribution $\mathcal{D}$)

# Finite/Countable/Bounded sets

- a finite set is a set that has a **finite** number of elements

$$\{(3,4),(1,1),(0,0)\}, \left\{ \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}, \begin{bmatrix} 3 & -1 \\ 10 & 9 \end{bmatrix} \right\}, \quad \text{but } \mathbf{R}^{m\times n} \text{ is not finite}$$

- a set is **countable** if each element in the set is uniquely associated to a unique natural number (or can be counted at a time)

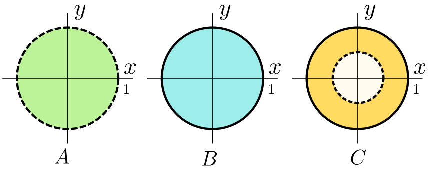
  $\{1,2,3,\ldots\}$ is countable (but not finite), set of diagonal matrices is not countable

- a subset $C$ of a normed vector space is **bounded** if there exists $M > 0$ such that $\|x - v\| < M$ for all $x, v \in C$
  - $\mathrm{span}\{(1,1)\}$ is not bounded
  - $\{x \in \mathbf{R}^2 \mid x = (1,1) + t(2,3) \mid t \in [0,1]\}$ is bounded (but not finite)

# Open and closed sets

concepts about open and closed sets are generalized to normed vector space[1]

let $C$ be a subset of a normed space $\mathcal{V}$




$x \in C$ is called an **interior point** of $C$ if there exists $\epsilon > 0$ for which

$$\{y \mid \|y - x\| \leq \epsilon\} \subseteq C$$

(if all points of $\epsilon$-neighborhood of $x$ are also stay in $C$)

- the set of all interior points of $C$ is denoted by $\mathbf{int}\, C$
- a set $C$ is said to be **open** if $\mathbf{int}\, C = C$ (every point in $C$ is an interior point)
- ✎ what is interior of $A$? is $A$ open ?
- a set $C$ is called **closed** if its complement $\mathcal{V} \setminus C$ is open

✎ is $B$ closed ?

[1] more general definitions for metric/topological space

# Supremum and infimum

let $C \subseteq \mathbf{R}$

- the **supremum** of the set $C$, denoted by $\sup C$ is the **least upper bound** of $C$

$$\sup(0,2) = 2, \;\; \sup(0,2] = 2, \;\; \sup\{(2,-1)^T x \mid \|x\|_2 < 1\} = \sqrt{5}$$

- $\max C$ denotes the maximum element in $C$ (that can be explicitly specified)
- $\sup C$ may or may not be in the set $C$; when $\sup C = C$, we say the supremum of $C$ is *attained* or *achieved*
- we take $\sup = -\infty$ and $\sup C = \infty$ when $C$ is unbounded above
- the **infimum** of $C$, denoted by $\inf C$, is the **greatest lower bound** of $C$

$$\inf(0,2) = 0, \;\; \sup[0,2] = 0, \;\; \sup\{(2,-1)^T x \mid \|x\|_2 < 1\} = -\sqrt{5}$$

- we take $\inf = \infty$ and $\inf C = -\infty$ when $C$ is unbounded below

# Hölder's inequality

the $\ell_p$ and $\ell_q$ norms are dual[2] in the sense that $\frac{1}{p} + \frac{1}{q} = 1$

$$\ell_1 \Leftrightarrow \ell_\infty, \quad \ell_2 \text{ is self-dual}$$

Hölder's inequality is an extension of Cauchy-Schwarz to all dual pairs:

$$|\langle x, y \rangle| \leq \|x\|_p \|y\|_q, \qquad p, q \in [1, \infty) \quad \text{with} \quad \frac{1}{p} + \frac{1}{q} = 1$$

(proofs can depend on the inner product space in question)

[2] there is more formal definition of dual norm/dual space

# Dual norm in $\mathbf{R}^n$

let $\|\cdot\|$ be a norm on $\mathbf{R}^n$; the **dual norm**, denoted $\|\cdot\|_*$ is defined as

$$\|z\|_* = \sup \{z^T x \mid \|x\| \leq 1 \}$$

(✎ verify that it is a norm )

- consider the operator norm of $z^T$ with the norm $\|\cdot\|$ on $\mathbf{R}^n$

$$\sup_{\|x\|\leq 1} \frac{\|z^T x\|}{\|x\|} = \sup_{\|x\|\leq 1} \frac{|z^T x|}{\|x\|} \implies \text{can be regarded as the dual norm}$$

- ✎ it can be shown that the dual norm of $\ell_2$ is itself and the dual norm of $\ell_\infty$ is $\ell_1$
- the dual of the dual norm is the original norm ($\|x\|_{**} = \|x\|$)
- from the definition of dual norm, we always have the inequality

$$z^T x \leq \|x\| \|z\|_* \quad \text{(a special case of Hölder's inequality for } \mathbf{R}^n\text{)}$$

# Dual norm in $\mathbf{R}^{m \times n}$

let $\|\cdot\|$ be a norm in $\mathbf{R}^{m \times n}$

the associated dual norm for this space is defined by generalizing the idea of inner product for matrices: $\langle X, Z \rangle = \mathbf{tr}(Z^T X)$

$$\|Z\|_* = \sup \{\mathbf{tr}(Z^T X) \mid \|X\| \leq 1 \}$$

for example, consider the spectral norm $\|X\|_2$

$$\begin{aligned} \|Z\|_{2*} &= \sup \{\mathbf{tr}(Z^T X) \mid \|X\|_2 \leq 1 \} \\ &= \sigma_1(Z) + \sigma_2(Z) + \cdots + \sigma_r(Z) = \mathbf{tr}(Z^T Z)^{1/2} \end{aligned}$$

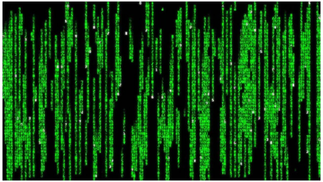
where $r = \mathbf{rank}(Z)$ – the dual norm of spectral norm turns out to be the nuclear norm

# References


1. C. C. Aggrawal, *Linear algebra and optimization for machine learning:A textbook*, Springer, 2020
2. G. Strang, *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press, 2019
3. A.N. Kolmogorov and S.V. Fomin, *Introductory real analysis*, Dover, 1970
4. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, 2014

# Special matrices

# Common matrices used in applications



| | | |
|---|---|---|
| symmetric | Hermittian | skew-symmetric |
| positive definite | Gram | nilpotent |
| unitary | orthogonal | permutation |
| idempotent | nilpotent | companion |
| Toeplitz | Hankel | Vandermonde |
| banded | doubly stochastic | adjacency |

# Unitary matrix

a *complex* matrix $U \in \mathbf{C}^{n\times n}$ is called **unitary** if

$$U^*U = UU^* = I, \qquad (U^* \triangleq \bar{U}^T)$$

example: let $z = e^{-i2\pi/3}$

$$U = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & z & z^2 \\ 1 & z^2 & z^4 \end{bmatrix} = \frac{1}{\sqrt{3}}\begin{bmatrix} 1 & 1 & 1 \\ 1 & e^{-i2\pi/3} & e^{-i4\pi/3} \\ 1 & e^{-i4\pi/3} & e^{-i8\pi/3} \end{bmatrix}$$

**facts:** ✎

- a unitary matrix is always invertible and $U^{-1} = U^*$
- columns vectors of $U$ are mutually orthogonal
- 2-norm is preserved under a unitary transformation: $\|Ux\|_2^2 = (Ux)^*(Ux) = \|x\|_2^2$

# Example: Discrete Fourier transform (DFT)

DFT of the length-$N$ time-domain sequence $x[n]$ is defined by

$$X[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-\mathrm{i}2\pi kn/N}, \quad 0 \leq k \leq N-1$$

define $z = e^{-\mathrm{i}2\pi/N}$, we can write the DFT in a matrix form as

$$\begin{bmatrix} X[0] \\ X[1] \\ X[2] \\ \vdots \\ X[N-1] \end{bmatrix} = \frac{1}{\sqrt{N}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & z^1 & z^2 & \cdots & z^{N-1} \\ 1 & z^2 & z^4 & \cdots & z^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z^{N-1} & z^{2(N-1)} & \cdots & z^{(N-1)(N-1)} \end{bmatrix} \begin{bmatrix} x[0] \\ x[1] \\ x[2] \\ \vdots \\ x[N-1] \end{bmatrix}$$

or $\mathbf{X} = \mathbf{D}\mathbf{x}$ where $\mathbf{D}$ is called the **DFT matrix** and is unitary ($\therefore \mathbf{x} = \mathbf{D}^*\mathbf{X}$)

# Unitary property of DFT

the columns of DFT matrix are of the form:

$$\phi_k = (1/\sqrt{N}) \begin{bmatrix} 1 & e^{-i2\pi k/N} & e^{-i2\pi k\cdot 2/N} & \cdots & e^{-i2\pi k(N-1)/N} \end{bmatrix}^T$$

use $\langle \phi_l, \phi_k \rangle = \phi_k^* \phi_l$ and apply the sum of geometric series:

$$\langle \phi_l, \phi_k \rangle = \frac{1}{N} \sum_{n=0}^{N-1} e^{i2\pi(k-l)n/N} = \frac{1}{N} \cdot \frac{1 - e^{i2\pi(k-l)}}{1 - e^{i2\pi(k-l)/N}}$$

the columns of DFT matrix are therefore *orthogonal*

$$\langle \phi_l, \phi_k \rangle = \begin{cases} 1, & \text{for } k = l + rN, \quad r = 0, 1, 2, \ldots \\ 0, & \text{for } k \neq l \end{cases}$$

# Orthogonal matrix

a *real* matrix $U \in \mathbf{R}^{n \times n}$ is called **orthogonal** if

$$UU^T = U^T U = I$$

**properties:** ✎

- an orthogonal matrix is special case of unitary for real matrices
- an orthogonal matrix is always invertible and $U^{-1} = U^T$
- columns vectors of $U$ are mutually orthogonal
- norm is preserved under an orthogonal transformation: $\|Ux\|_2^2 = \|x\|_2^2$

**example:**

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

# Projection matrix

$P \in \mathbf{R}^{n \times n}$ is said to be a **projection** matrix if $P^2 = P$ (aka **idempotent**)

- $P$ is a linear transformation from $\mathbf{R}^n$ to a subspace of $\mathbf{R}^n$, denoted as $S$
- columns of $P$ are the projections of standard basis vectors and $S$ is the range of $P$
- if $P$ is applied twice on a vector in $S$, it gives the same vector

examples: identity and

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}, \quad \begin{bmatrix} 3 & -6 \\ 1 & -2 \end{bmatrix}, \quad I - X(X^TX)^{-1}X^T \;\; \text{(in regression)}$$

**properties:** ✎

- eigenvalues of $P$ are all equal to $0$ or $1$
- $I - P$ is also idempotent
- if $P \neq I$, then $P$ is singular

# Orthogonal projection matrix

a matrix $P \in \mathbf{R}^{n\times n}$ is called an **orthogonal projection** matrix if

$$P^2 = P = P^T$$

**properties:**

- $P$ is bounded, *i.e.*, $\|Px\| \leq \|x\|$

$$\|Px\|_2^2 = x^T P^T P x = x^T P^2 x = x^T P x \leq \|Px\| \|x\|$$

- if $P$ is an orthogonal projection onto a line spanned by a unit vector $u$,

$$P = uu^T$$

(we see that $\mathbf{rank}(P) = 1$ as the dimension of a line is $1$)

- another example: $P = X(X^T X)^{-1} X^T$ for any matrix $X$ – (in regression)

# Permutation

a **permutation** matrix $P$ is a square matrix that has exactly one entry of 1 in each row and each column and has zero elsewhere

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

**facts:** ✎

- $P$ is obtained by interchanging any two rows (or columns) of an identity matrix
- $PA$ results in permuting rows in $A$, and $AP$ gives permuting columns in $A$
- $P^TP = I$, so $P^{-1} = P^T$ (simple)
- the modulus of all eigenvalues of $P$ is one, *i.e.*, $|\lambda_i(P)| = 1$
- a permutation matrix is an example of **doubly stochastic** matrix

# Stochastic matrix

a (real) square matrix $A$ with non-negative entries is called

1. a **row/right stochastic** if each row sums to 1: $\sum_j a_{ij} = 1$ or $\mathbf{1}^T A = \mathbf{1}^T$
2. a **column/left stochastic** if each column sums to 1: $\sum_i a_{ij} = 1$ or $A\mathbf{1} = \mathbf{1}$
3. a **doubly stochastic** if each row and column sums to 1

row/left stochastic: $\begin{bmatrix} 0.2 & 0.1 & 0 \\ 0.3 & 0.9 & 1 \\ 0.5 & 0 & 0 \end{bmatrix}$, doubly: $\begin{bmatrix} 0.1 & 0.5 & 0.4 \\ 0.2 & 0.2 & 0.6 \\ 0.7 & 0.3 & 0 \end{bmatrix}$, $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

- a stochastic matrix clearly has 1 as an eigenvalue
- ✎ the spectral radius of any stochastic matrix is one
- a left stochastic matrix appears in Markov chain as the **transition probability** matrix: $p(t+1) = Ap(t)$ where $A_{ij}$ is the conditional probability that state $j$ from time $t$ jumps to state $i$ at time $t+1$

# Vandermonde

appears in polynomial evaluation at multiple points

we are not related !

$$p(t) = c_1 + c_2 t + \cdots + c_{n-1} t^{n-2} + c_n t^{n-1}$$

$$V = \begin{bmatrix} 1 & t_1 & \cdots & t_1^{n-1} \\ 1 & t_2 & \cdots & t_2^{n-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & t_n & \cdots & t_n^{n-1} \end{bmatrix}$$

(with a geometric progression in each row)

✎ one can show that the determinant of $V$ can be expressed as

$$\det(V) = \prod_{1 \leq i < j \leq n} (t_j - t_i)$$

hence, $V$ is invertible as long as $t_i$'s are **distinct**

# Companion matrix

$$A = \begin{bmatrix} -a_1 & a_2 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad a_1, \ldots, a_n \in \mathbf{R}$$

1. appears as the state-space dynamic matrix of autoregressive (AR) process

$$y(t) = a_1 y(t-1) + a_2 y(t-2) + \cdots + a_n y(t-n) + u(t)$$

2. ✎ the characteristic polynomial of $A$ is given by

$$\lambda^n + a_1\lambda^{n-1} + a_2\lambda^{n-2} + a_{n-1}\lambda + a_n = 0$$

3. stationarity of AR process is obtained via the root test depending on $a_1, \ldots, a_n$

# Companion matrices in state-space system

## controllable canonical form

$$A = \begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ 0 & 1 & & 0 & -a_{n-2} \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & -a_1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_n \end{bmatrix}$$

$\mathcal{C} = I_n$ and $(A, B)$ is controllable

## observable cannonical form

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & & 0 \\ \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \cdots & -a_1 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

$\mathcal{O} = I_n$ and $(A, C)$ is observable

## controller canonical form

$$A = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$C = \begin{bmatrix} c_1 & c_2 & c_3 & \cdots & c_n \end{bmatrix}$$

$\mathcal{C}$ is an upper triangular matrix with 1's on the diagonal and $(A, B)$ is controllable

## observer canonical form

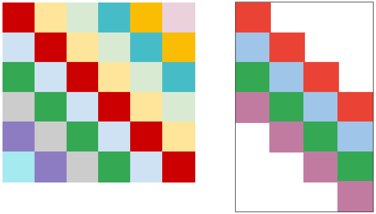
$$A = \begin{bmatrix} -a_1 & 1 & 0 & \cdots & 0 & 0 \\ -a_2 & 0 & 1 & & 0 & 0 \\ \vdots & \vdots & & \ddots & & \vdots \\ -a_{n-1} & 0 & 0 & & 1 & 0 \\ -a_n & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{bmatrix}$$

$$C = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \end{bmatrix}$$

$\mathcal{O}$ is a lower triangular with 1's on the diagonal and $(A, C)$ is observable

# Toeplitz

Toeplitz matrix has constant entries along each descending diagonal from left to right



- $T_{ij} =$ constant when $i - j$ is fixed
- $T$ needs not be square

- the set of $n \times n$ Toeplitz matrices forms a subspace for $\mathbf{R}^{n \times n}$
- an $n \times n$ Toeplitz $T$ has at most $2n - 1$ unique values
- two Toeplitz matrices can be added in $\mathcal{O}(n)$ time
- the linear system $y = Tx$ can be solved by the **Levinson algorithm** in $\mathcal{O}(n^2)$
- can be found in convolution system, covariance matrix, polynomial multiplication
  - See more in Boyd and Vandenberghe page 137 and `https://ee.stanford.edu/~gray/toeplitz.pdf`

# Convolution: impulse response

consider an input-output relationship in a convolution form

$$y(t) = \sum_{k=0}^{\infty} h_k u(t-k) = h_0 u(t) + h_1 u(t-1) + \cdots + h_t u(0)$$

the input-output response in vector format has a **Toeplitz** system

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{N-1} \\ y_N \end{bmatrix} = \begin{bmatrix} h_0 & & & & \\ h_1 & h_0 & & & \\ \vdots & \ddots & \ddots & & \\ h_{N-1} & h_{N-2} & \ddots & h_0 & \\ h_N & h_{N-1} & \cdots & h_1 & h_0 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{N-1} \\ u_N \end{bmatrix} \triangleq y = T(h)u$$

when considering $M$-order FIR (finite impulse response) where $h_t = 0$ for $t = M+1, M+2, \ldots$, $T(h)$ becomes a banded matrix

# Autocorrelation matrix

for a wide-sense stationary process (WSS), define auto-correlation function:

$$R(\tau) = \mathbf{E}[x(t+\tau)x(t)^T], \quad R(-\tau) = R(\tau)^T$$

which has non-negative property: for any $a_j, a_j \in \mathbf{R}^n$ and for $1 \leq i, j \leq n$

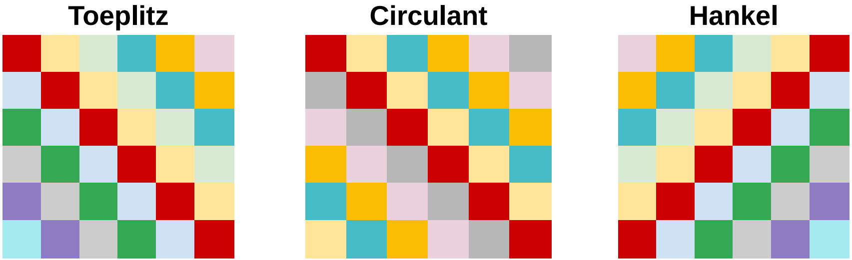
$$\sum_i \sum_j a_i^T R(i-j) a_j \geq 0$$

which is equivalent to positivity of a quadratic form with a **Toeplitz** coefficient matrix:

$$\begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix}^T \begin{bmatrix} R_0 & R_{-1} & \cdots & R_{-(n-2)} & R_{-(n-1)} \\ R_1 & R_0 & R_{-1} & \cdots & R_{-(n-2)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ R_{n-2} & \cdots & R_1 & R_0 & R_{-1} \\ R_{n-1} & R_{n-2} & \cdots & R_1 & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{bmatrix} \geq 0$$

# Hankel and Circulant matrices

Toeplitz matrix's siblings




- **circulant matrix:** each row is a cyclic shift of the row above (e.g., covariance matrix of WSS process)
- **Hankel matrix:** ascending skew-diagonal from left to right is constant (e.g., input-output relationship from state-space model)

# Nilpotent matrix

$A \in \mathbf{R}^{n \times n}$ is *nilpotent* if

$$A^k = 0, \quad \text{for some positive integer } k$$

**Example:** any triangular matrices with $0$'s along the main diagonal

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{(shift matrix)}$$

also related to deadbeat control for linear discrete-time systems

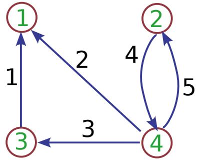
**facts:** ✎

- the characteristic equation for $A$ is $\lambda^n = 0$
- all eigenvalues are $0$

# Graphs

a **graph**: consists of

1. **nodes (or vertices):** labeled by $\{1, 2, \ldots, n\}$
2. **edges:** set $\mathcal{E}$ of $(i, j)$ describing connections between node $i$ and $j$ where 'connection' can be defined in many ways
   - **directed graph:** the connections are bi-directional
   - **undirected graph:** the connections are undirectional (or symmetric)




Directed graph

edge

| node | | | | | |
|---|---|---|---|---|---|
| | 1 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 0 | -1 | 1 |
| | -1 | 0 | 1 | 0 | 0 |
| | 0 | -1 | -1 | 1 | -1 |

incidence matrix

- directed edge from node $j$ to $i$ can be described by a **relation** set

$$\mathcal{R} = \{(1,3), (1,4), (2,3), (2,4), (3,4), (4,2)\}$$

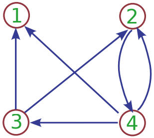
- undirected edge between node $i$ and $j$ can be described by a set of pair $(i, j)$:

$$\{(1,3), (1,4), (2,3), (2,4), (3,4)\}$$

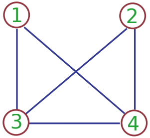
# Graph matrix: Adjacency

a relation $\mathcal{R}$ on $\{1, 2, \ldots, n\}$ is represented by the $n \times n$ matrix $A$ with

$$A_{ij} = \begin{cases} 1, & (i,j) \in \mathcal{R} \\ 0, & (i,j) \notin \mathcal{R} \end{cases}$$



Directed graph



Undirected graph

example of how a relation is defined:

- directed edge: variable $j$ causes variable $i$
- undirected edge: covariance, partial covariance

directed

$$\mathcal{R} = \{(1,3), (1,4), (2,3), (2,4), (3,4), (4,2)\}$$

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

undirected

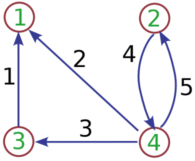
$$\mathcal{R} = \{(1,3), (1,4), (2,3), (2,4), (3,4)\}$$

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

# Graph matrix: Incidence

a **directed graph** can be described by its $n \times m$ **incidence** matrix, defind as

$$A_{ij} = \begin{cases} 1, & \text{edge } j \text{ points to node } i \\ -1, & \text{edge } j \text{ points from node } i \\ 0, & \text{otherwise} \end{cases}$$




Directed graph

edge

| node | | | | | |
|---|---|---|---|---|---|
| | 1 | 1 | 0 | 0 | 0 |
| | 0 | 0 | 0 | -1 | 1 |
| | -1 | 0 | 1 | 0 | 0 |
| | 0 | -1 | -1 | 1 | -1 |

incidence matrix

- dimension of incidence matrix: no. of edges x no. of nodes
- each column has only two nonzero entries (-1 and 1)

- the $i$th row sum gives a total net flow of node $i$
- unlike adjacency matrix, incidence matrix explicitly labels the edges $1, 2, \ldots, m$

# References


1. S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least squares*, Cambridge, 2018
2. C. C. Aggrawal, *Linear algebra and optimization for machine learning:A textbook*, Springer, 2020
3. G. Strang, *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press, 2019
4. S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge, 2014

# Matrix decompositions

# Decompositions

1. SVD (singular value decomposition)
2. QR
3. LU
4. Cholesky

# SVD decomposition

let $A \in \mathbf{R}^{m \times n}$ be a rectangular matrix; there exists the SVD form of $A$

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


- $U \in \mathbf{R}^{m \times m}, V \in \mathbf{R}^{n \times n}$ are orthogonal matrices
- $\Sigma \in \mathbf{R}^{m \times n}$ with $\Sigma_{ii} = \sigma_i \geq 0$ and $\Sigma_{ij} = 0$ for $i \neq j$
- for a rectangular $A$, $\Sigma$ has a diagonal submatrix $\Sigma_1$ with dimension of $\min(m, n)$

$$A_{\text{tall}} = \begin{bmatrix} U_1 & | & U_2 \end{bmatrix} \left[ \frac{\Sigma_1}{0} \right] V^T = U_1 \Sigma_1 V^T, \quad A_{\text{fat}} = U \begin{bmatrix} \Sigma_1 & | & 0 \end{bmatrix} \left[ \frac{V_1^T}{V_2^T} \right] = U \Sigma_1 V_1^T$$

# Singular vectors and singular value

suppose $\mathbf{rank}(A) = r$, $A$ has $r$ positive singular values in descending order

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

and there exist **left singular vector** $u_1, \ldots, u_m$ that are *orthogonal* in $\mathbf{R}^m$ and **right singular vector** $v_1, \ldots, v_n$ that are *orthogonal* in $\mathbf{R}^n$ such that

$$Av_1 = \sigma_1 u_1, \quad Av_2 = \sigma_2 u_2, \ldots, \quad Av_r = \sigma_r u_r, \quad Av_{r+1} = \cdots = Av_n = 0$$

or in matrix form: $AV = U\Sigma$ (where $U$ and $V$ are orthogonal matrices)

$$A \left[\begin{array}{ccc|ccc} v_1 & \cdots & v_r & v_{r+1} & \cdots & v_n \end{array}\right] = \left[\begin{array}{ccc|ccc} u_1 & \cdots & u_r & u_{r+1} & \cdots & u_m \end{array}\right] \left[\begin{array}{ccc|c} \sigma_1 & & & 0 \\ & \ddots & & 0 \\ & & \sigma_r & 0 \\ \hline 0 & 0 & 0 & \mathbf{0} \end{array}\right]$$

unlike eigenvalue decomposition: $AX = X\Lambda$, SVD needs two sets of singular vectors

# How to find $U, \Sigma, V$

for $A = U\Sigma V^T$, we can write

$$A^T A = V\Sigma^T \Sigma V^T \triangleq Q\Lambda Q^T, \qquad AA^T = U\Sigma\Sigma^T U^T \triangleq Q\Lambda Q^T$$

- $V$ contains orthonormal eigenvectors of $A^T A$
- $U$ contains orthonormal eigenvectors of $AA^T$
- $\sigma_1^2, \ldots, \sigma_r^2$ are the nonzero eigenvalues of both $A^T A$ and $AA^T$

steps of finding $U, \Sigma, V$:

1. choose orthonormal eigenvectors $v_1, \ldots, v_r$ of $A^T A$
2. choose $\sigma_k = \sqrt{\lambda_k(A^T A)}$ for $k = 1, \ldots, r$
3. from $Av = \sigma u$, compute $u_k = \frac{Av_k}{\sigma_k}$ for $k = 1, \ldots, r$
4. the last $v_{r+1}, \ldots, v_n$ are in $\mathcal{N}(A)$ and the last $u_{r+1}, \ldots, u_m$ are in $\mathcal{N}(A^T)$ (just pick any orthonormal bases for those subspaces)

# Example: Computing SVD

$$A = \begin{bmatrix} 1 & 0 & 2 \\ -2 & 1 & 0 \end{bmatrix}, \quad A^T A = \begin{bmatrix} 5 & -2 & 2 \\ -2 & 1 & 0 \\ 2 & 0 & 4 \end{bmatrix}$$

- find the right singular vector (eigenvectors of $A^T A$)

$$A^T A = Q \Lambda Q^T, Q = \begin{bmatrix} \frac{3}{\sqrt{14}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{21}} \\ -\frac{1}{\sqrt{14}} & \frac{1}{\sqrt{6}} & \frac{4}{\sqrt{21}} \\ \frac{2}{\sqrt{14}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{21}} \end{bmatrix}, \quad D = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \sigma_1^2 = 7, \sigma_2^2 = 3$$

then $V = Q$ and $\Sigma = \begin{bmatrix} \sqrt{7} & 0 & 0 \\ 0 & \sqrt{3} & 0 \end{bmatrix}$

# Example: Computing SVD

- find the left singular vector $U$ as the normalized image of right singular vector

$$u_1 = \frac{Av_1}{\sigma_1} = \begin{bmatrix} 1 & 0 & 2 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} \frac{1}{\sqrt{14 \cdot 7}} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$u_2 = \frac{Av_2}{\sigma_2} = \begin{bmatrix} 1 & 0 & 2 \\ -2 & 1 & 0 \end{bmatrix} \begin{bmatrix} -1 \\ 1 \\ 2 \end{bmatrix} \frac{1}{\sqrt{6 \cdot 3}} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$U = \begin{bmatrix} u_1 & u_2 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix}$$

- the SVD form of $A$ is

$$\begin{bmatrix} 1 & 0 & 2 \\ -2 & 1 & 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{7} & 0 & 0 \\ 0 & \sqrt{3} & 0 \end{bmatrix} \begin{bmatrix} \frac{3}{\sqrt{14}} & -\frac{1}{\sqrt{6}} & \frac{2}{\sqrt{21}} \\ -\frac{1}{\sqrt{14}} & \frac{1}{\sqrt{6}} & \frac{4}{\sqrt{21}} \\ \frac{2}{\sqrt{14}} & \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{21}} \end{bmatrix}^T$$

# Reduced vs Truncated SVD form

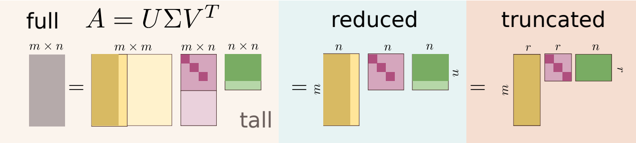
consider $A \in \mathbf{R}^{m \times n}$ and $A^T A$ has size $n \times n$

- the number of nonzero $\lambda(A^T A)$ is less than or equal to $n$
- suppose the number of nonzero $\sigma(A) = \sqrt{\lambda(A^T A)}$ is $r < n$
- the **reduced SVD form** is to use the diagonal $\Sigma_1 \in \mathbf{R}^{n \times n}$ as in the red terms

$$A_{\text{tall}} = \left[\begin{array}{c|c} U_1 & U_2 \end{array}\right] \left[\begin{array}{c} \Sigma_1 \\ \hline 0 \end{array}\right] V^T = U_1 \Sigma_1 V^T, \quad A_{\text{fat}} = U \left[\begin{array}{c|c} \Sigma_1 & 0 \end{array}\right] \left[\begin{array}{c} V_1^T \\ \hline V_2^T \end{array}\right] = U \Sigma_1 V_1^T$$

  and if $r < n$ then $\Sigma_1$ contains $r$ nonzero diagonal entries

- the **truncated SVD** is to further extract only the non-zero diagonal block of $\Sigma_1$



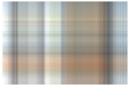
# SVD application: Low rank approximation

when $A$ has nonzero $r$ singular values: $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$

$$\text{truncated form: } A = U_r \Sigma_r V_r^T = \sum_{k=1}^{r} \sigma_k u_k v_k^T \quad (r\text{-sum of rank-1 matrices})$$

**original**

$r = 1$



$r = 10$

$r = 40$

$r = 80$

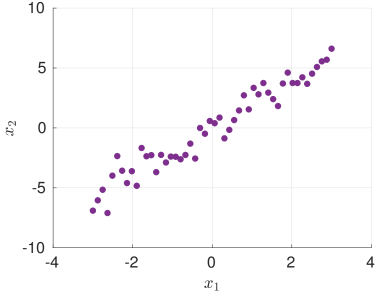
**Eckart-Young theorem:** consider $A \in \mathbf{R}^{m \times n}$ of rank $r$ and $X \in \mathbf{R}^{m \times n}$ of rank $k$; for any $k \leq r$ with $A_k = \sum_{j=1}^{k} \sigma_j u_j v_j^T$ it holds that

$$A_k = \underset{X:\mathbf{rank}(X)=k}{\operatorname{argmin}} \|A - X\|_2, \quad \text{with error } \|A - A_k\|_2 = \sigma_{k+1}$$

the best rank$-k$ approximation of $A$ is the first $k$ pieces in SVD decomposition

# SVD application: PCA

data points are clustered along a subspace (here, line) in $\mathbf{R}^p$




- question: reduce the variable dimension but keep most information in the data
- setting: find the directions that contain $k$-largest variance in data covariance
- data matrix $X \in \mathbf{R}^{p\times N}$ and its covariance is $C = XX^T/(N-1)$

- **total variance** in the data: $T = \mathbf{tr}(C) = \frac{\mathbf{tr}(XX^T)}{N-1} = \frac{\mathbf{tr}(X^TX)}{N-1} = \frac{\|X\|_F^2}{N-1}$
- SVD of $X$ is $U\Sigma V^T$, so covariance is $C = \frac{U\Sigma^2 U^T}{N-1}$
- total variance is also expressed as the sum of $r$ non-zero singular values:

$$T = (\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_r^2)/(N-1)$$

# SVD application: PCA

for data matrix $X \in \mathbf{R}^{p \times N}$ with $X = U\Sigma V^T = \sum_{k=1}^{r} \sigma_k u_k v_k^T$

- the first $k$ **principal loadings** $u_1, u_2, \ldots, u_k$ accounts for a fraction of

$$(\sigma_1^2 + \cdots + \sigma_k^2)/T$$

- we can transform $X$ to a new data matrix using the first $k$ loadings

$$Y = \begin{bmatrix} u_1^T \\ \vdots \\ u_k^T \end{bmatrix} X$$

example:

$$X = \begin{bmatrix} 3 & -4 & 7 & 1 & -4 & -3 \\ 7 & -6 & 8 & -1 & -1 & 7 \end{bmatrix}, \quad \sigma_1 = 16.87, \sigma_2 = 3.92$$

supppose we reduce the data to 1-dimension using the first loading $u_1$

$$Y = u_1^T X = u_1^T \left(\sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T\right) = \sigma_1 v_1^T = \begin{bmatrix} -7.48 & 7.21 & -10.55 & 0.27 & 3.07 & 7.48 \end{bmatrix}$$

# Recall Gram-Schmidt (GS)

let $A \in \mathbf{R}^{m \times n}$ with independent columns $a_1, a_2, \ldots, a_n$ (hence, $A$ is **tall or square**)

- vectors $q_1, \ldots, q_n$ are orthonormal vectors produced by GS on $a_1, \ldots, a_n$
- $\tilde{q}_i$ is the vector after projecting $a_i$ on the previous orthogonal vectors

$$\tilde{q}_i = a_i - \left( \langle a_i, q_1 \rangle q_1 + \langle a_i, q_2 \rangle q_2 + \cdots + \langle a_i, q_{i-1} \rangle q_{i-1} \right), \quad \text{and} \quad q_i = \tilde{q}_i / \|\tilde{q}_i\|$$

- hence, we can write $a_i$ as linear combination of $q_1, \ldots, q_i$

$$\begin{aligned} a_i &= (q_1^T a_i) q_1 + (q_2^T a_i) q_2 + \cdots + (q_{i-1}^T a_i) q_{i-1} + \|\tilde{q}_i\| q_i, \quad i = 1, \ldots, n \\ a_1 &= \|\tilde{q}_1\| q_1 \\ a_2 &= (q_1^T a_2) q_1 + \|\tilde{q}_2\| q_2 \\ a_3 &= (q_1^T a_3) q_1 + (q_2^T a_3) q_2 + \|\tilde{q}_3\| q_3 \end{aligned}$$

- we can form $q_1, \ldots, q_n$ as columns of $Q$

# QR factorization

we can write $A = QR$ where

$$A = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_n \end{bmatrix} = \begin{bmatrix} q_1 & q_2 & q_3 & \cdots & q_n \end{bmatrix} \begin{bmatrix} \|\tilde{q}_1\| & q_1^T a_2 & q_1^T a_3 & \cdots & q_1^T a_n \\ & \|\tilde{q}_2\| & q_2^T a_3 & \cdots & q_2^T a_n \\ & & \|\tilde{q}_3\| & & \vdots \\ & & & \ddots & q_{n-1}^T a_n \\ & & & & \|\tilde{q}_n\| \end{bmatrix}$$

- $Q \in \mathbf{R}^{m \times n}$ contains columns as orthonormal vectors $q_1, \ldots, q_n$ with $Q^T Q = I_n$
- $R \in \mathbf{R}^{n \times n}$ is an upper triangular matrix with $R_{ii} = \|\tilde{q}_i\|$ and $R_{ij} = q_i^T a_j$ for $i < j$
- if $a_1, \ldots, a_n$ are all LI, $\tilde{q}_i$'s are not zero, so $R_{ii} \neq 0$
- if some $a_j$ is dependent of others, $R_{jj} = 0$

QR factorization can be found in computing orthogonal projection: numerical solution of least-square estimate, subspace identification

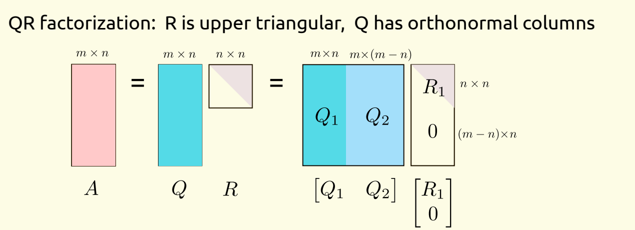
# Full QR factorization

for a full column rank $A \in \mathbf{R}^{m \times n}$, we have

- $q_1, q_2, \ldots, q_n$ that form bases vectors for $\mathcal{R}(A)$ and put them as columns in $Q_1$
- we can find the remaining $(m - n)$ orthonormal vectors: $q_{n+1}, \ldots, q_m$ so that $\{q_1, \ldots, q_m\}$ form a basis for $\mathbf{R}^m$; put these vectors as columns in $Q_2$

$$\mathcal{R}(A) = \mathcal{R}(Q_1), \quad \mathcal{R}(A)^\perp = \mathcal{R}(Q_2)$$

- hence, $\tilde{Q} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \in \mathbf{R}^{m \times m}$ is orthogonal: $\tilde{Q}^T \tilde{Q} = \tilde{Q}\tilde{Q}^T = I_m$
- we also have a full QR factorization: $A = \tilde{Q}\tilde{R}$ where $\tilde{R}$ has zero padding



# Factor-solve approach

to solve $Ax = b$, first write $A$ as a product of 'simple' matrices

$$A = A_1 A_2 \cdots A_k$$

then solve $(A_1 A_2 \cdots A_k)x = b$ by solving $k$ equations

$$A_1 z_1 = b, \qquad A_2 z_2 = z_1, \quad \ldots, \quad A_{k-1} z_{k-1} = z_{k-2}, \qquad A_k x = z_{k-1}$$

**complexity** of factor-solve method: flops $= f + s$

- $f$ is cost of factoring $A$ as $A = A_1 A_2 \cdots A_k$ (factorization step)
- $s$ is cost of solving the $k$ equations for $z_1$, $z_2$, ...$z_{k-1}$, $x$ (solve step)
- usually $f \gg s$

# Forward substitution

solve $Ax = b$ when $A$ is lower triangular with nonzero diagonal elements

$$\begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

**algorithm**:

$$\begin{aligned} x_1 &:= b_1/a_{11} \\ x_2 &:= (b_2 - a_{21}x_1)/a_{22} \\ x_3 &:= (b_3 - a_{31}x_1 - a_{32}x_2)/a_{33} \\ &\vdots \\ x_n &:= (b_n - a_{n1}x_1 - a_{n2}x_2 - \cdots - a_{n,n-1}x_{n-1})/a_{nn} \end{aligned}$$

**cost**: $1 + 3 + 5 + \cdots + (2n-1) = n^2$ flops

# Back substitution

solve $Ax = b$ when $A$ is upper triangular with nonzero diagonal elements

$$\begin{bmatrix} a_{11} & \cdots & a_{1,n-1} & a_{1n} \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & \cdots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}$$

**algorithm**:

$$\begin{aligned} x_n &:= b_n/a_{nn} \\ x_{n-1} &:= (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n-1} \\ x_{n-2} &:= (b_{n-2} - a_{n-2,n-1}x_{n-1} - a_{n-2,n}x_n)/a_{n-2,n-2} \\ &\vdots \\ x_1 &:= (b_1 - a_{12}x_2 - a_{13}x_3 - \cdots - a_{1n}x_n)/a_{11} \end{aligned}$$

**cost**: $n^2$ flops

# LU decomposition

for a nonsingular $A$, it can be factorized as (with row pivoting)

$$A = PLU$$

factorization:

- $P$ permutation matrix, $L$ unit lower triangular, $U$ upper triangular
- **factorization cost**: $(2/3)n^3$ if $A$ has order $n$
- not unique; there may be several possible choices for $P$, $L$, $U$
- interpretation: permute the rows of $A$ and factor $P^TA$ as $P^TA = LU$
- also known as *Gaussian elimination with partial pivoting* (GEPP)

# Not every matrix has an LU factor

without row pivoting, LU factor may not exist even when $A$ is invertible

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \quad \Rightarrow \quad LU = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$$

from this example,

- if $A$ could be factored as LU, it would require that $l_{11}u_{11} = a_{11} = 0$
- one of $L$ or $U$ would be singular, contradicting to the fact that $A = LU$ is nonsingular

# Solving a linear system with LU factor

solving linear system: $(PLU)x = b$ in three steps

- permutation: $z_1 = P^T b$ (0 flops)
- forward substitution: solve $Lz_2 = z_1$ ($n^2$ flops)
- back substitution: solve $Ux = z_2$ ($n^2$ flops)

**total cost**: $(2/3)n^3 + 2n^2$ flops, or roughly $(2/3)n^3$

# Cholesky factorization

every positive definite matrix $A$ can be factored as

$$A = LL^T$$

where $L$ is lower triangular with positive diagonal elements

- **cost**: $(1/3)n^3$ flops if $A$ is of order $n$
- $L$ is called the *Cholesky factor* of $A$
- can be interpreted as 'square root' of a positive define matrix
- $L$ is invertible (its diagonal elements are nonzero)
- $A$ is invertible and

$$A^{-1} = L^{-T}L^{-1}$$

# Cholesky factorization algorithm

partition matrices in $A = LL^T$ as

$$\begin{bmatrix} a_{11} & A_{21}^T \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ L_{21} & L_{22} \end{bmatrix} \begin{bmatrix} l_{11} & L_{21}^T \\ 0 & L_{22}^T \end{bmatrix} = \begin{bmatrix} l_{11}^2 & l_{11}L_{21}^T \\ l_{11}L_{21} & L_{21}L_{21}^T + L_{22}L_{22}^T \end{bmatrix}$$

**algorithm:**

1. determine $l_{11}$ and $L_{21}$:

$$l_{11} = \sqrt{a_{11}}, \qquad L_{21} = \frac{1}{l_{11}} A_{21}$$

2. compute $L_{22}$ from

$$A_{22} - L_{21}L_{21}^T = L_{22}L_{22}^T$$

   this is a Cholesky factorization of order $n-1$

# Proof of Cholesky algorithm

**proof** that the algorithm works for positive definite $A$ of order $n$

- step 1: if $A$ is positive definite then $a_{11} > 0$
- step 2: if $A$ is positive definite, then

$$A_{22} - L_{21}L_{21}^T = A_{22} - \frac{1}{a_{11}}A_{21}A_{21}^T$$

  is positive definite (by Schur complement)
- hence the algorithm works for $n = m$ if it works for $n = m - 1$
- it obviously works for $n = 1$; therefore it works for all $n$

# Example of Cholesky algorithm

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

- first column of $L$

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & l_{22} & 0 \\ -1 & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} 5 & 3 & -1 \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

- second column of $L$

$$\begin{bmatrix} 18 & 0 \\ 0 & 11 \end{bmatrix} - \begin{bmatrix} 3 \\ -1 \end{bmatrix} \begin{bmatrix} 3 & -1 \end{bmatrix} = \begin{bmatrix} l_{22} & 0 \\ l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{22} & l_{32} \\ 0 & l_{33} \end{bmatrix}$$

$$\begin{bmatrix} 9 & 3 \\ 3 & 10 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 1 & l_{33} \end{bmatrix} \begin{bmatrix} 3 & 1 \\ 0 & l_{33} \end{bmatrix}$$

- third column of $L$: $10 - 1 = l_{33}^2$, *i.e.*, $l_{33} = 3$

conclusion:

$$\begin{bmatrix} 25 & 15 & -5 \\ 15 & 18 & 0 \\ -5 & 0 & 11 \end{bmatrix} = \begin{bmatrix} 5 & 0 & 0 \\ 3 & 3 & 0 \\ -1 & 1 & 3 \end{bmatrix} \begin{bmatrix} 5 & 3 & -1 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}$$

# Solving equations with positive definite $A$

$$Ax = b \qquad (A \text{ positive definite of order } n)$$

**algorithm**

- factor $A$ as $A = LL^T$
- solve $LL^T x = b$
  - forward substitution $Lz = b$
  - back substitution $L^T x = z$

**cost**: $(1/3)n^3$ flops

- factorization: $(1/3)n^3$
- forward and backward substitution: $2n^2$

# References


1. G. Strang, *Linear Algebra and Learning from Data*, Wellesley-Cambridge Press, 2019
2. M.P. Deisenroth, A.A. Faisal, and C.S. Ong, *Mathematics for Machine Learning*, Cambridge University Press, 2020
3. S. Boyd and L. Vandenberghe, *Introduction to Applied Linear Algebra: Vectors, Matrices, and Least squares*, Cambridge, 2018
4. Lecture notes of EE133A, L. Vandenberhge, UCLA `https://www.seas.ucla.edu/~vandenbe/133A`

# Solving linear/nonlinear equations

# Topic

1. problem condition
2. solving large-scale linear systems
3. gradient and Hessian
4. solving nonlinear equations

# Sources of error in numerical computation

**example**: evaluate a function $f : \mathbf{R} \rightarrow \mathbf{R}$ at a given $x$ (*e.g.*, $f(x) = \sin x$)
sources of error in the result:

- $x$ is not exactly known
  - measurement errors
  - errors in previous computations

  $\longrightarrow$ how sensitive is $f(x)$ to errors in $x$?
- the algorithm for computing $f(x)$ is not exact
  - discretization (*e.g.*, the algorithm uses a table to look up $f(x)$)
  - truncation (*e.g.*, $f$ is computed by truncating a Taylor series)
  - rounding error during the computation

  $\longrightarrow$ how large is the error introduced by the algorithm?

# The condition of a problem

sensitivity of the solution with respect to errors in the data

- **well-conditioned:** if small errors in the data produce small errors in the result
- **ill-conditioned:** if small errors in the data may produce large errors in the result

**example:** function evaluation: $y = f(x), y + \Delta y = f(x + \Delta x)$

- absolute error

$$|\Delta y| \approx |f'(x)||\Delta x|$$

  ill-conditioned with respect to absolute error if $|f'(x)|$ is very large

- relative error

$$\frac{|\Delta y|}{|y|} \approx \frac{|f'(x)||x|}{|f(x)|} \frac{|\Delta x|}{|x|}$$

  ill-conditioned w.r.t relative error if $|f'(x)||x|/|f(x)|$ is very large

# Condition of a set of linear equations

assume $A$ is nonsingular and $Ax = b$

if we change $b$ to $b + \Delta b$, the new solution is $x + \Delta x$ with

$$A(x + \Delta x) = b + \Delta b$$

the change in $x$ is

$$\Delta x = A^{-1}\Delta b$$

**condition** of the equations: a technical term used to describe how sensitive the solution is to changes in the righthand side

- the equations are **well-conditioned** if small $\Delta b$ results in small $\Delta x$
- the equations are **ill-conditioned** if small $\Delta b$ can result in large $\Delta x$

# Example of ill-conditioned equations

$$A = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1+10^{-10} & 1-10^{-10} \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} 1-10^{10} & 10^{10} \\ 1+10^{10} & -10^{10} \end{bmatrix}$$

- solution for $b = (1,1)$ is $x = (1,1)$
- change in $x$ if we change $b$ to $b + \Delta b$:

$$\Delta x = A^{-1}\Delta b = \begin{bmatrix} \Delta b_1 - 10^{10}(\Delta b_1 - \Delta b_2) \\ \Delta b_1 + 10^{10}(\Delta b_1 - \Delta b_2) \end{bmatrix}$$

small $\Delta b$ can lead to extremely large $\Delta x$

# Bound on absolute error

suppose $A$ is nonsingular and $\Delta x = A^{-1}\Delta b$

**upper bound** on $\|\Delta x\|$

$$\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|$$

(follows from property of operator norm)

- small $\|A^{-1}\|$ means that $\|\Delta x\|$ is small when $\|\Delta b\|$ is small
- large $\|A^{-1}\|$ means that $\|\Delta x\|$ can be large, even when $\|\Delta b\|$ is small
- for any $A$, there exists $\Delta b$ such that $\|\Delta x\| = \|A^{-1}\|\|\Delta b\|$ ✎

# Bound on relative error

suppose $A$ is nonsingular, $Ax = b$ with $b \neq 0$, and $\Delta x = A^{-1}\Delta b$

**upper bound** on $\|\Delta x\|/\|x\|$:

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|}$$

(follows from $\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|$ and $\|b\| \leq \|A\|\|x\|$)

$\kappa(A) = \|A\|\|A^{-1}\|$ is called the **condition number** of $A$

- small $\kappa(A)$ means $\|\Delta x\|/\|x\|$ is small when $\|\Delta b\|/\|b\|$ is small
- large $\kappa(A)$ means $\|\Delta x\|/\|x\|$ can be large, even when $\|\Delta b\|/\|b\|$ is small
- for any $A$, there exist $b$, $\Delta b$ such that $\|\Delta x\|/\|x\| = \kappa(A)\|\Delta b\|/\|b\|$

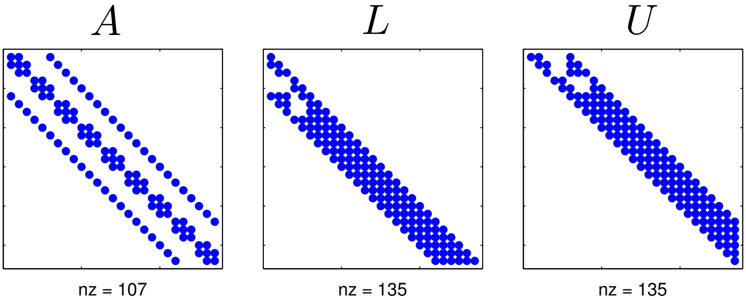
# Condition number

$$\kappa(A) = \|A\| \|A^{-1}\|$$

- defined for nonsingular $A$
- $\kappa(A) \geq 1$ for all $A$ ✎
- $A$ is a **well-conditioned** matrix if $\kappa(A)$ is small (close to 1):
  the relative error in $x$ is not much larger than the relative error in $b$
- $A$ is **badly conditioned** or **ill-conditioned** if $\kappa(A)$ is large:
  the relative error in $x$ can be much larger than the relative error in $b$

# Large sparse linear systems

consider solving $Ax = b$ when $A$ is **sparse** and the dimension of $A$ is **huge**




factorization methods are sometimes not a good technique because

- the number of non-zero entries in the factors is increased due to fill-in
- storing the factors $L$ and $U$ will require much more storage

# Application on solving PDE

large sparse matrices arise in the numerical solution of PDE/ODE

$$-u''(x) = f(x), \qquad 0 < x < 1, \qquad \text{where } u(0) \text{ and } u(1) \text{ are given}$$

discretize the system with step $h$ and obtain $Au = b$ with unknowns $u_1, \ldots, u_{n-1}$

$$A = \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & \ddots & & \\ & & \ddots & \ddots & -1 & \\ & & & 1 & 2 & -1 \\ & & & & -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} h^2 f(x_1) + u(0) \\ h^2 f(x_2) \\ h^2 f(x_3) \\ \vdots \\ h^2 f(x_{n-2}) \\ h^2 f(x_{n-1}) + u(1) \end{bmatrix}$$

- by making $h$ small, the solution is more accurate, but $\#$ of variables increases
- we can show that $A$ is nonsingular (and pdf), hence the solution is unique
- $A$ is tri-diagonal (extremely sparse)

# Solving large linear systems

outline of available methods

- splitting method: $A = M - N$ (split to easy $M$)

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b \quad \text{(until convergence which depends on } M^{-1}N)$$

- Jacobi iteration: $A = D - (D - A)$ (split to diagonal + residual)

$$x^{(k+1)} = (I - D^{-1}A)x^{(k)} + D^{-1}b$$

- Gauss-Seidal iteration: $A = L - (L - A)$ (split to lower triangular)

$$x^{(k+1)} = (I - L^{-1}A)x^{(k)} + L^{-1}b$$

convergence of Jacobi and Gauss-Seidal depends on $A$ (diagonally dominant, psdf)

further reading: D. Kincaid and W. Cheney, *Numerical analysis*, Brooks/Cole, 2022

# Derivative and Gradient

Suppose $f : \mathbf{R}^n \to \mathbf{R}^m$ and $x \in \mathbf{int\,dom}\, f$

the **derivative** (or **Jacobian**) of $f$ at $x$ is the matrix $Df(x) \in \mathbf{R}^{m \times n}$:

$$Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j}, \quad i = 1, \dots, m, \quad j = 1, \dots, n$$

- when $f$ is scalar-valued (*i.e.*, $f : \mathbf{R}^n \to \mathbf{R}$), the derivative $Df(x)$ is a row vector
- its transpose is called the **gradient** of the function:

$$\nabla f(x) = Df(x)^T, \qquad \nabla f(x)_i = \frac{\partial f(x)}{\partial x_i}, \quad i = 1, \dots, n$$

which is a column vector in $\mathbf{R}^n$

# Second Derivative

suppose $f$ is a scalar-valued function (*i.e.*, $f : \mathbf{R}^n \to \mathbf{R}$)

the second derivative or **Hessian matrix** of $f$ at $x$, denoted $\nabla^2 f(x)$ is

$$\nabla^2 f(x)_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, n$$

**example:** the quadratic function $f : \mathbf{R}^n \to \mathbf{R}$

$$f(x) = (1/2)x^T P x + q^T x + r,$$

where $P \in \mathbf{S}^n, q \in \mathbf{R}^n$, and $r \in \mathbf{R}$

- $\nabla f(x) = Px + q$
- $\nabla^2 f(x) = P$

# Chain rule

assumptions:

- $f : \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at $x \in \textbf{int dom}\, f$
- $g : \mathbf{R}^m \to \mathbf{R}^p$ is differentiable at $f(x) \in \textbf{int dom}\, g$
- define the composition $h : \mathbf{R}^n \to \mathbf{R}^p$ by

$$h(z) = g(f(z))$$

then $h$ is differentiable at $x$, with derivative

$$Dh(x) = Dg(f(x))Df(x)$$

special case: $f : \mathbf{R}^n \to \mathbf{R}$, $g : \mathbf{R} \to \mathbf{R}$, and $h(x) = g(f(x))$

$$\nabla h(x) = g'(f(x))\nabla f(x)$$

# Example of chain rule

1. $h(x) = f(Ax + b)$

$$Dh(x) = Df(Ax+b)A \quad \Rightarrow \quad \nabla h(x) = A^T \nabla f(Ax+b)$$

2. $h(x) = (1/2)(Ax - b)^T P(Ax - b)$

$$\nabla h(x) = A^T P(Ax - b)$$

3. $h(x) = (\max(0, a^T x + b))^2$

$$\nabla h(x) = \begin{cases} 2a\max(0, a^T x + b), & a^T x + b > 0 \\ 0, & a^T x + b < 0 \\ \text{not defined}, & a^T x + b = 0 \end{cases}$$

# Exercises

find the gradient of the following functions

1. probit log-likelihood: variable $= \theta$, $\Phi$ is Gaussian cdf, $(x, y)$ is data

$$f(\theta) = \sum_{i=1}^{N} y_i \log(\Phi(x_i^T\theta) + (1 - y_i)\log[1 - \Phi(x_i^T\theta)])$$

2. Poisson log-likelihood: variable $= \beta$, $(x, y)$ is data

$$f(\beta) = \sum_{i=1}^{N} -e^{x_i^T\beta} + y_i x_i^T\beta - \log y_i!$$

# Function of matrices

we typically encounter some scalar-valued functions of matrix $X \in \mathbf{R}^{m \times n}$

- $f(X) = \mathbf{tr}(A^T X)$ (linear in $X$)
- $f(X) = \mathbf{tr}(X^T A X)$ (quadratic in $X$)

**definition:** the derivative of $f$ (scalar-valued function) with respect to $X$ is

$$\frac{\partial f}{\partial X} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \vdots & & \ddots & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \frac{\partial f}{\partial x_{m2}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}$$

note that the differential of $f$ can be generalized to

$$f(X + dX) - f(X) = \langle \frac{\partial f}{\partial X}, dX \rangle + \text{higher order term}$$

see more on the matrix cookbook by Petersen and Pedersen, `https://ece.uwaterloo.ca/~ece602/MISC/matrixcookbook.pdf`

# Derivative of a trace function

let $f(X) = \mathbf{tr}(A^T X)$

$$\begin{aligned} f(X) &= \sum_i (A^T X)_{ii} = \sum_i \sum_k (A^T)_{ki} X_{ki} \\ &= \sum_i \sum_k A_{ki} X_{ki} \end{aligned}$$

then we can read that $\frac{\partial f}{\partial X} = A$ (by the definition of derivative)

we can also note that

$$f(X + dX) - f(X) = \mathbf{tr}(A^T(X + dX)) - \mathbf{tr}(A^T X) = \mathbf{tr}(A^T dX) = \langle dX, A \rangle$$

then we can read that $\frac{\partial f}{\partial X} = A$

# Examples

- $f(X) = \mathbf{tr}(X^T A X)$

$$\begin{aligned} f(X+dX) - f(X) &= \mathbf{tr}((X+dX)^T A (X+dX)) - \mathbf{tr}(X^T A X) \\ &\approx \mathbf{tr}(X^T A dX) + \mathbf{tr}(dX^T A X) \\ &= \langle dX, A^T X \rangle + \langle AX, dX \rangle \end{aligned}$$

then we can read that $\frac{\partial f}{\partial X} = A^T X + AX$

- $f(X) = \|Y - XH\|_F^2$ where $Y$ and $H$ are given

$$\begin{aligned} f(X+dX) &= \mathbf{tr}((Y - XH - dXH)^T (Y - XH - dXH)) \\ f(X+dX) - f(X) &\approx -\mathbf{tr}(H^T dX^T (Y - XH)) - \mathbf{tr}((Y-XH)^T dXH) \\ &= -\mathbf{tr}((Y-XH)H^T dX^T) - \mathbf{tr}(H(Y-XH)^T dX) \\ &= -2\langle (Y-XH)H^T, dX \rangle \end{aligned}$$

then we identifiy that $\frac{\partial f}{\partial X} = -2(Y - XH)H^T$

# Derivative of a $\log\det$ function

let $f : \mathbf{S}^n \to \mathbf{R}$ be defined by $f(X) = \log\det(X)$

$$\begin{aligned}
\log\det(X + dX) &= \log\det(X^{1/2}(I + X^{-1/2}dX X^{-1/2})X^{1/2}) \\
&= \log\det X + \log\det(I + X^{-1/2}dX X^{-1/2}) \\
&= \log\det X + \sum_{i=1}^{n} \log(1 + \lambda_i)
\end{aligned}$$

where $\lambda_i$ is an eigenvalue of $X^{-1/2}dX X^{-1/2}$

$$\begin{aligned}
f(X + dX) - f(X) &\approx \sum_{i=1}^{n} \lambda_i \quad (\log(1 + x) \approx x, \quad x \to 0) \\
&= \mathbf{tr}(X^{-1/2}dX X^{-1/2}) \\
&= \mathbf{tr}(X^{-1}dX)
\end{aligned}$$

we identify that $\frac{\partial f}{\partial X} = X^{-1}$

# Example: Gaussian log-likelihood

suppose $y_1, \ldots, y_N$ are Gaussian vectors $\mathcal{N}(\mu, \Sigma)$

$$
\begin{aligned}
\mathcal{L}(\mu, \Sigma) &= \frac{1}{2} \log\det \Sigma^{-1} + \frac{1}{2N} \sum_{k=1}^{N} (y_k - \mu)^T \Sigma^{-1} (y_k - \mu) \\
&\triangleq \log\det \Sigma^{-1} - \mathbf{tr}(C\Sigma^{-1}), \quad C = \frac{1}{N} \sum_{k=1}^{N} (y_k - \mu)(y_k - \mu)^T \\
&\triangleq \log\det X - \mathbf{tr}(CX)
\end{aligned}
$$

what is the gradient of $\mathcal{L}$ w.r.t. $X$ ?

# Notes on gradients

many machine learning and optimization problems use gradients for

- training model parameters
- finding solution that satisfies the optimality condition

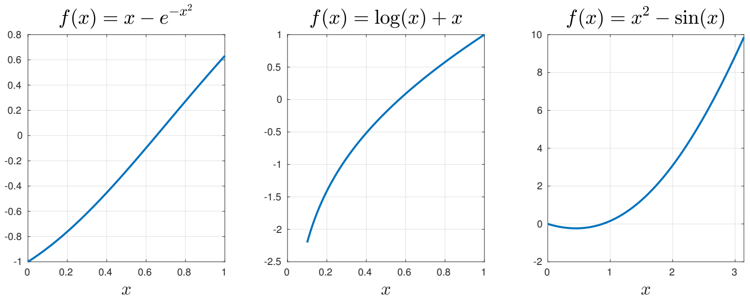
further reading on the topics

- backpropagation algorithm (apply chain rule) in deep NN
- automatic differentiation (a numerical technique to find $\nabla f$ by working with intermediate variables)

# Nonlinear equations

**root finding problem:** find $x \in \mathbf{R}$ such that $f(x) = 0$, *e.g.*,

- $f(x) = x - e^{-x^2}$
- $f(x) = \log(x) + x$
- $f(x) = x^2 - \sin(x)$



# Methods of finding roots

**example of methods:** bisection, Newton, secant, fixed point

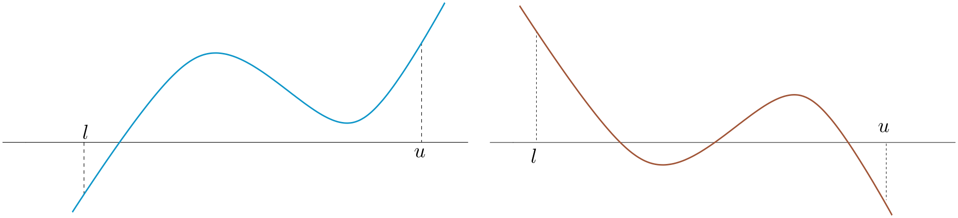
**methods are iterative**

- generate a sequence of points $x^{(k)}$, $k = 0, 1, 2, \ldots$ that converge to a solution; $x^{(k)}$ is called the $k$th *iterate*; $x^{(0)}$ is the *starting point*
- computing $x^{(k+1)}$ from $x^{(k)}$ is called one *iteration* of the algorithm
- each iteration typically requires one evaluation of $f$ (or $f$ and $f'$) at $x^{(k)}$
- algorithms need a stopping criterion, *e.g.*, terminate if

$$|f(x^{(k)})| \leq \text{specified tolerance}$$

- speed of the algorithm depends on:
  - the cost of evaluating $f(x)$ (and possibly, $f'(x)$)
  - the number of iterations

# Bisection



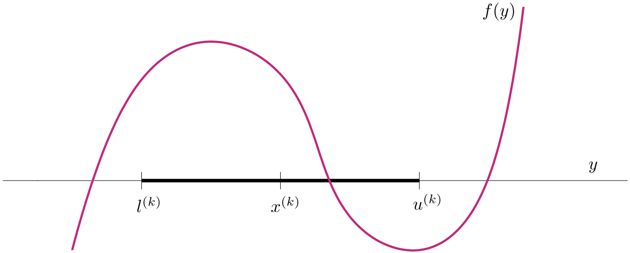

if $f(l)f(u) < 0$, then the interval $[l, u]$ contains at least one zero

**intermediate value theorem:** Let $f \in \mathbf{C}([a, b])$ and assume $p$ is a value between $f(a)$ and $f(b)$, that is

$$f(a) \leq p \leq f(b), \quad \text{or} \quad f(b) \leq p \leq f(a)$$

then there exists a point $c \in [a, b]$ for which $f(c) = p$

# Bisection algorithm



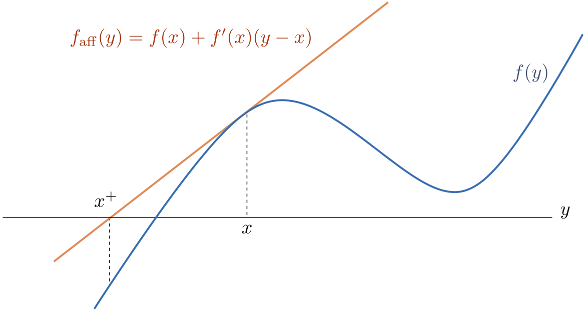

**given** $l$, $u$ with $l < u$ and $f(l)f(u) < 0$; a required tolerance $\delta, \epsilon > 0$

**repeat**

1. $x := (l + u)/2$.
2. Compute $f(x)$.
3. **if** $f(x) = 0$, **return** $x$.
4. **if** $f(x)f(l) < 0$, $u := x$, **else**, $l := x$.

**until** $u - l < \epsilon$ or $|f(x)| < \delta$

# Newton's method




- make affine approximation of $f$ around $x$ using Taylor series expansion:

$$f_{\text{aff}}(y) = f(x) + f'(x)(y - x)$$

- solve the linearized equation $f_{\text{aff}}(y) = 0$ and take the solution $y$ as $x^+$:

$$x^+ = x - f(x)/f'(x)$$

# Newton's algorithm

$f : \mathbf{R} \to \mathbf{R}$, differentiable

**given** initial $x$, required tolerance $\epsilon > 0$
**repeat**

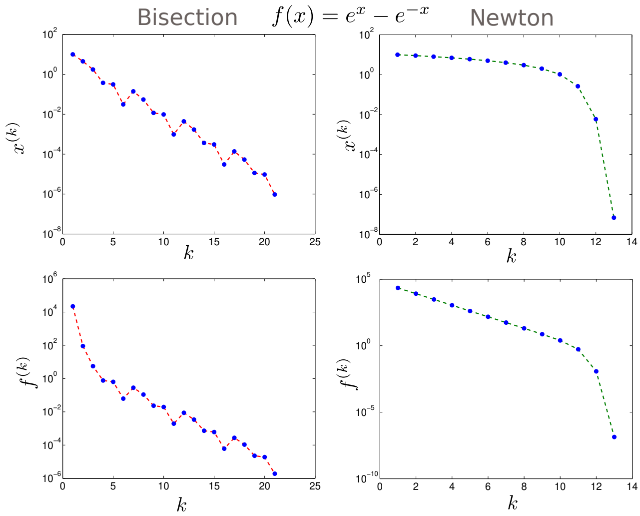
1. Compute $f(x)$ and $f'(x)$.
2. **if** $|f(x)| \leq \epsilon$, **return** $x$.
3. $x := x - f(x)/f'(x)$.

**until** maximum number of iterations is exceeded

properties:

- Newton's method has quadratic convergence
- require $f$ and $f'$
- it may not work if we start too far from a solution

# Numerical example




- $f(x) = e^x - e^{-x}$ which has a unique zero $x^\star = 0$
- start bisection method with $l = -1$, $u = 21$
- start Newton with $x^{(0)} = 10$

# Nonlinear systems

let $f : \mathbf{R}^n \to \mathbf{R}^m$, find $x \in \mathbf{R}^n$ such that $f(x) = 0$

example 1:

$$\begin{aligned} 2x_1 - x_2 + \frac{1}{9}e^{-x_1} &= -1 \\ -x_1 + 2x_2 + \frac{1}{9}e^{-x_2} &= 1 \end{aligned}$$

example 2:

$$\begin{aligned} 3x_1 - \cos(x_2 x_3) - 1/2 &= 0 \\ x_1^2 - 81(x_2 + 0.1)^2 + \sin(x_3) + 1.06 &= 0 \\ e^{-x_1 x_2} + 20x^3 + \frac{10\pi - 3}{3} &= 0 \end{aligned}$$

# Applications

most typical example is to solve uncontrained optimization

$$\underset{x}{\text{minimize}} \quad g(x) \quad \Longleftrightarrow \quad \text{find } x^\star \text{ such that } \nabla g(x^\star) = 0$$

1. zero gradient condition of nonlinear least-squares

$$\text{curv fitting:} \quad \underset{\beta}{\text{minimize}} \quad \sum_{i=1}^{N} (y_i - \beta_0 \sin(\beta_1 t + \beta_2))^2$$

2. zero gradient condition of maximum likelihood estimate

$$\text{Poisson likelihood:} \quad \text{maximize}_\beta \quad \mathcal{L}(\beta) = \sum_{i=1}^{N} -\exp(x_i^T \beta) + y_i x_i^T \beta - \log y_i!$$

where $\{x_i, y_i\}_{i=1}^N$ are data and variable is $\beta \in \mathbf{R}^n$

# Newton's method for nonlinear systems

consider a function $f : \mathbf{R}^n \rightarrow \mathbf{R}^n$

let $x^\star = x + h$ and use the affine approximation of $f$ about $x$

$$0 = f(x^\star) = f(x+h) \approx f(x) + Df(x)h$$

where $Df(x)$ is the Jacobian matrix of $f$, *i.e.*, $Df(x)_{ij} = \frac{\partial f_i(x)}{\partial x_j}$

then, solve $h$ from

$$h = -Df(x)^{-1} f(x)$$

provided that the Jacobian matrix is nonsingular

Newton's method is summarized by

$$x^{(k+1)} = x^{(k)} - [Df(x^{(k)})]^{-1} f(x^{(k)})$$

which follows the same treatment for single equation

# Softwares

- MATLAB: `fsolve`
  - algorithm: trust-region, Levenberg-Marquardt
  - input = function, initial point $x_0$
- python: `scipy.optimize.fsolve`
  - many other available methods for large scale problems
  - Broyden's method: approximate Jacobian matrix

# References


1. D. Kincaid and W. Cheney, *Numerical Analysis*, Brooks/Cole Publishing, 2002
2. J.F. Epperson, *An Introduction to Numerical Methods and Analysis*, John Wiley and Sons, 2002
3. M.P. Deisenroth, A.A. Faisal, and C.S. Ong, *Mathematics for Machine Learning*, Cambridge University Press, 2020
4. K.B. Petersen and M.S. Pedersen, *The Matrix Cookbook*, November 2012 `https://ece.uwaterloo.ca/~ece602/MISC/matrixcookbook.pdf`