# Applied Mathematical Methods

Bhaskar Dasgupta

Department of Mechanical Engineering
Indian Institute of Technology
Kanpur (INDIA)
dasgupta@iitk.ac.in

(Pearson Education 2006, 2007)

May 13, 2008



## Contents I





## Contents II





## Contents III

# Contents IV





# Contents V





# Contents VI





# Contents VII

# Outline







# Theme of the Course



To develop a firm mathematical background necessary for graduate studies and research

- a fast-paced recapitulation of UG mathematics
- extension with supplementary advanced ideas for a mature and forward orientation
- exposure and highlighting of interconnections

To *pre-empt* needs of the *future* challenges

- trade-off between *sufficient* and *reasonable*
- target mid-spectrum *majority* of students

Notable beneficiaries (at two ends)

- would-be researchers in analytical/computational areas
- students who are till now somewhat *afraid* of mathematics



# Course Contents



- Applied linear algebra
- Multivariate calculus and vector calculus
- Numerical methods
- Differential equations + +
- Complex analysis



# Sources for More Detailed Study



If you have the time, need and interest, then you may consult

- **individual books** on individual topics;
- another "umbrella" volume, like Kreyszig, McQuarrie, **O'Neil** or Wylie and Barrett;
- a good book of numerical analysis or scientific computing, like Acton, **Heath**, Hildebrand, Krishnamurthy and Sen, **Press et al**, Stoer and Bulirsch;
- friends, in **joint-study groups**.

## Logistic Strategy



- Study in the given sequence, to the extent possible.
- **Do not read mathematics.**
- Use lots of pen and paper.
  Read "mathematics books" and **do** mathematics.
- Exercises are **must**.
  - Use as many methods as you can think of, certainly including the one which is recommended.
  - Consult the Appendix after you work out the solution. Follow the comments, interpretations and suggested extensions.
  - Think. Get excited. Discuss. Bore everybody in your known circles.
  - Not enough time to attempt all? Want a selection?
- Program implementation is needed in algorithmic exercises.
  - Master a programming environment.
  - Use mathematical/numerical library/software.

*Take a MATLAB tutorial session?*



## Logistic Strategy



**Tutorial Plan**

| Chapter | Selection | Tutorial | Chapter | Selection | Tutorial |
|---|---|---|---|---|---|
| **2** | 2,3 | 3 | **26** | 1,2,4,6 | 4 |
| **3** | 2,4,5,6 | 4,5 | **27** | 1,2,3,4 | 3,4 |
| **4** | 1,2,4,5,7 | 4,5 | **28** | 2,5,6 | 6 |
| **5** | 1,4,5 | 4 | **29** | 1,2,5,6 | 6 |
| **6** | 1,2,4,7 | 4 | **30** | 1,2,3,4,5 | 4 |
| **7** | 1,2,3,4 | 2 | **31** | 1,2 | 1(d) |
| **8** | 1,2,3,4,6 | 4 | **32** | 1,3,5,7 | 7 |
| **9** | 1,2,4 | 4 | **33** | 1,2,3,7,8 | 8 |
| **10** | 2,3,4 | 4 | **34** | 1,3,5,6 | 5 |
| **11** | 2,4,5 | 5 | **35** | 1,3,4 | 3 |
| **12** | 1,3 | 3 | **36** | 1,2,4 | 4 |
| **13** | 1,2 | 1 | **37** | 1 | 1(c) |
| **14** | 2,4,5,6,7 | 4 | **38** | 1,2,3,4,5 | 5 |
| **15** | 6,7 | 7 | **39** | 2,3,4,5 | 4 |
| **16** | 2,3,4,8 | 8 | **40** | 1,2,4,5 | 4 |
| **17** | 1,2,3,6 | 6 | **41** | 1,3,6,8 | 8 |
| **18** | 1,2,3,6,7 | 3 | **42** | 1,3,6 | 6 |
| **19** | 1,3,4,6 | 6 | **43** | 2,3,4 | 3 |
| **20** | 1,2,3 | 2 | **44** | 1,2,4,7,9,10 | 7,10 |
| **21** | 1,2,5,7,8 | 7 | **45** | 1,2,3,4,7,9 | 4,9 |
| **22** | 1,2,3,4,5,6 | 3,4 | **46** | 1,2,5,7 | 7 |
| **23** | 1,2,3 | 3 | **47** | 1,2,3,5,8,9,10 | 9,10 |
| **24** | 1,2,3,4,5,6 | 1 | **48** | 1,2,4,5 | 5 |
| **25** | 1,2,3,4,5 | 5 | | | |



## Expected Background



- moderate background of undergraduate mathematics
- firm understanding of school mathematics and undergraduate calculus

Take the preliminary test.

Grade yourself sincerely.

**Prerequisite Problem Sets***



## Points to note



- Put in effort, keep pace.
- Stress concept as well as problem-solving.
- Follow methods diligently.
- Ensure background skills.

Necessary Exercises: **Prerequisite problem sets ??**

## Outline





## Matrices

**Question:** What is a "matrix"?
**Answers:**

- a rectangular array of numbers/elements ?
- a mapping $f : M \times N \to F$, where $M = \{1, 2, 3, \cdots, m\}$, $N = \{1, 2, 3, \cdots, n\}$ and $F$ is the set of real numbers or complex numbers ?

**Question:** What does a matrix **do**?
**Explore:** With an $m \times n$ matrix $\mathbf{A}$,

$$\left.\begin{array}{rcl} y_1 &=& a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ y_2 &=& a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots & \vdots & \vdots \qquad \vdots \qquad \vdots \\ y_m &=& a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n \end{array}\right\} \quad \text{or} \quad \mathbf{Ax} = \mathbf{y}$$



## Matrices

Consider these definitions:

- $y = f(x)$
- $y = f(\mathbf{x}) = f(x_1, x_2, \cdots, x_n)$
- $y_k = f_k(\mathbf{x}) = f_k(x_1, x_2, \cdots, x_n), \quad k = 1, 2, \cdots, m$
- $\mathbf{y} = \mathbf{f}(\mathbf{x})$
- $\mathbf{y} = \mathbf{Ax}$

**Further Answer:**

*A matrix is the* definition *of a linear vector function of a vector variable.*

*Anything deeper?*

**Caution:** Matrices *do not* define vector functions whose components are of the form

$$y_k = a_{k0} + a_{k1}x_1 + a_{k2}x_2 + \cdots + a_{kn}x_n.$$



## Geometry and Algebra

Let vector $\mathbf{x} = [x_1 \;\; x_2 \;\; x_3]^T$ denote a point $(x_1, x_2, x_3)$ in 3-dimensional space in frame of reference $OX_1X_2X_3$.
**Example:** With $m = 2$ and $n = 3$,

$$\left.\begin{array}{rcl} y_1 &=& a_{11}x_1 + a_{12}x_2 + a_{13}x_3 \\ y_2 &=& a_{21}x_1 + a_{22}x_2 + a_{23}x_3 \end{array}\right\}.$$

Plot $y_1$ and $y_2$ in the $OY_1Y_2$ plane.

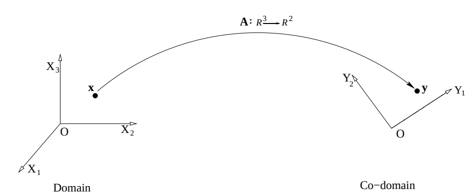



Figure: Linear transformation: schematic illustration

*What is matrix* $\mathbf{A}$ *doing?*

## Geometry and Algebra



*Operating* on point $\mathbf{x}$ in $R^3$, matrix $\mathbf{A}$ *transforms* it to $\mathbf{y}$ in $R^2$.

Point $\mathbf{y}$ is the *image* of point $\mathbf{x}$ under the mapping defined by matrix $\mathbf{A}$.

Note *domain* $R^3$, *co-domain* $R^2$ with reference to the figure and verify that $\mathbf{A} : R^3 \rightarrow R^2$ fulfils the requirements of a *mapping*, by definition.

*A matrix gives* **a** *definition of a* **linear transformation** *from one vector space to another.*



## Linear Transformations



Operate $\mathbf{A}$ on a large number of points $\mathbf{x}_i \in R^3$.
Obtain corresponding images $\mathbf{y}_i \in R^2$.

*The linear transformation represented by* $\mathbf{A}$ *implies the totality of these correspondences.*

We decide to use a different *frame of reference* $OX'_1X'_2X'_3$ for $R^3$.
[And, possibly $OY'_1Y'_2$ for $R^2$ at the same time.]

*Coordinates* change, i.e. $\mathbf{x}_i$ changes to $\mathbf{x}'_i$ (and possibly $\mathbf{y}_i$ to $\mathbf{y}'_i$).
Now, we need a different matrix, say $\mathbf{A}'$, to get back the correspondence as $\mathbf{y}' = \mathbf{A}'\mathbf{x}'$.

A matrix: just **one** description.

**Question:** How to get the new matrix $\mathbf{A}'$?



## Matrix Terminology



- ··· ···
- Matrix product
- Transpose
- Conjugate transpose
- Symmetric and skew-symmetric matrices
- Hermitian and skew-Hermitian matrices
- Determinant of a square matrix
- Inverse of a square matrix
- Adjoint of a square matrix
- ··· ···



## Points to note



- A matrix defines a linear transformation from one vector space to another.
- Matrix representation of a linear transformation depends on the selected bases (or frames of reference) of the source and target spaces.

**Important:** Revise matrix algebra basics as necessary tools.

Necessary Exercises: **2,3**

## Outline





## Range and Null Space: Rank and Nullity

Consider $\mathbf{A} \in R^{m \times n}$ as a mapping

$$\mathbf{A} : R^n \to R^m, \qquad \mathbf{A}\mathbf{x} = \mathbf{y}, \qquad \mathbf{x} \in R^n, \qquad \mathbf{y} \in R^m.$$

Observations

1. Every $\mathbf{x} \in R^n$ has an image $\mathbf{y} \in R^m$, but every $\mathbf{y} \in R^m$ need not have a pre-image in $R^n$.

   *Range (or range space) as subset/subspace of co-domain: containing images of* all $\mathbf{x} \in R^n$*.*

2. Image of $\mathbf{x} \in R^n$ in $R^m$ is unique, but pre-image of $\mathbf{y} \in R^m$ need not be.
   It may be non-existent, unique or infinitely many.

   *Null space as subset/subspace of domain: containing pre-images of* only $\mathbf{0} \in R^m$*.*



## Range and Null Space: Rank and Nullity

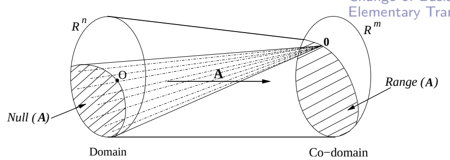



Figure: Range and null space: schematic representation

**Question:** What is the dimension of a vector space?
**Linear dependence and independence:** Vectors $\mathbf{x}_1$, $\mathbf{x}_2$, $\cdots$, $\mathbf{x}_r$ in a vector space are called linearly independent if

$$k_1\mathbf{x}_1 + k_2\mathbf{x}_2 + \cdots + k_r\mathbf{x}_r = \mathbf{0} \quad \Rightarrow \quad k_1 = k_2 = \cdots = k_r = 0.$$

$$\begin{aligned} Range(\mathbf{A}) &= \{\mathbf{y} : \mathbf{y} = \mathbf{A}\mathbf{x}, \ \ \mathbf{x} \in R^n\} \\ Null(\mathbf{A}) &= \{\mathbf{x} : \mathbf{x} \in R^n, \ \ \mathbf{A}\mathbf{x} = \mathbf{0}\} \\ Rank(\mathbf{A}) &= \dim Range(\mathbf{A}) \\ Nullity(\mathbf{A}) &= \dim Null(\mathbf{A}) \end{aligned}$$



## Basis

Take a set of vectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\cdots$, $\mathbf{v}_r$ in a vector space.
**Question:** Given a vector $\mathbf{v}$ in the vector space, can we describe it as

$$\mathbf{v} = k_1\mathbf{v}_1 + k_2\mathbf{v}_2 + \cdots + k_r\mathbf{v}_r = \mathbf{V}\mathbf{k},$$

where $\mathbf{V} = [\mathbf{v}_1 \ \ \mathbf{v}_2 \ \ \cdots \ \ \mathbf{v}_r]$ and $\mathbf{k} = [k_1 \ \ k_2 \ \ \cdots \ \ k_r]^T$?
**Answer:** Not necessarily.

**Span**, denoted as $< \mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_r >$: the subspace described/generated by a set of vectors.

**Basis:**

*A basis of a vector space is composed of an ordered* minimal *set of vectors spanning the entire space.*

The basis for an $n$-dimensional space will have exactly $n$ members, all linearly independent.

## Basis

Orthogonal basis: $\{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n\}$ with

$$\mathbf{v}_j^T \mathbf{v}_k = 0 \quad \forall\, j \neq k.$$

Orthonormal basis:

$$\mathbf{v}_j^T \mathbf{v}_k = \delta_{jk} = \begin{cases} 0 & \text{if} \quad j \neq k \\ 1 & \text{if} \quad j = k \end{cases}$$

Members of an **orthonormal** basis form an **orthogonal** matrix.
Properties of an orthogonal matrix:

$$\begin{aligned} \mathbf{V}^{-1} &= \mathbf{V}^T \text{ or } \mathbf{V}\mathbf{V}^T = \mathbf{I}, \text{ and} \\ \det \mathbf{V} &= +1 \text{ or } -1, \end{aligned}$$

Natural basis:

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \mathbf{e}_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}.$$



## Change of Basis

Suppose $\mathbf{x}$ represents a vector (point) in $R^n$ in some basis.
**Question:** If we change over to a new basis $\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n\}$, how does the representation of a vector change?

$$\begin{aligned} \mathbf{x} &= \bar{x}_1\mathbf{c}_1 + \bar{x}_2\mathbf{c}_2 + \cdots + \bar{x}_n\mathbf{c}_n \\ &= [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_n] \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{bmatrix}. \end{aligned}$$

With $\mathbf{C} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_n]$,

new to old coordinates: $\mathbf{C}\bar{\mathbf{x}} = \mathbf{x}$ *and*
old to new coordinates: $\bar{\mathbf{x}} = \mathbf{C}^{-1}\mathbf{x}$.

Note: Matrix $\mathbf{C}$ is invertible. *How?*
Special case with $\mathbf{C}$ orthogonal:
**orthogonal coordinate transformation.**



## Change of Basis

**Question:** And, how does basis change affect the representation of a linear transformation?

Consider the mapping $\mathbf{A} : R^n \rightarrow R^m, \quad \mathbf{A}\mathbf{x} = \mathbf{y}.$

Change the basis of the domain through $\mathbf{P} \in R^{n \times n}$ and that of the co-domain through $\mathbf{Q} \in R^{m \times m}$.

New and old vector representations are related as

$$\mathbf{P}\bar{\mathbf{x}} = \mathbf{x} \quad \text{and} \quad \mathbf{Q}\bar{\mathbf{y}} = \mathbf{y}.$$

Then, $\mathbf{A}\mathbf{x} = \mathbf{y} \Rightarrow \bar{\mathbf{A}}\bar{\mathbf{x}} = \bar{\mathbf{y}}$, with

$$\boxed{\bar{\mathbf{A}} = \mathbf{Q}^{-1}\mathbf{A}\mathbf{P}}$$

Special case: $m = n$ and $\mathbf{P} = \mathbf{Q}$ gives a **similarity transformation**

$$\bar{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$$



## Elementary Transformations

**Observation:** Certain reorganizations of equations in a system have no effect on the solution(s).

**Elementary Row Transformations:**

1. interchange of two rows,
2. scaling of a row, and
3. addition of a scalar multiple of a row to another.

**Elementary Column Transformations:** Similar operations with columns, equivalent to a corresponding *shuffling* of the *variables* (unknowns).

## Elementary Transformations

**Equivalence of matrices:** An elementary transformation defines an equivalence relation between two matrices.

**Reduction to normal form:**

$$\mathbf{A}_N = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

**Rank invariance:** Elementary transformations do not alter the rank of a matrix.

**Elementary transformation as matrix multiplication:**

> *an elementary row transformation on a matrix is equivalent to a* pre-multiplication *with an* elementary matrix, *obtained through the same row transformation on the identity matrix (of appropriate size).*

Similarly, an elementary column transformation is equivalent to *post-multiplication* with the corresponding elementary matrix.



## Points to note

- Concepts of range and null space of a linear transformation.
- Effects of change of basis on representations of vectors and linear transformations.
- Elementary transformations as tools to modify (simplify) systems of (simultaneous) linear equations.

Necessary Exercises: **2,4,5,6**



## Outline





## Nature of Solutions

$$\boxed{\mathbf{Ax} = \mathbf{b}}$$

Coefficient matrix: $\mathbf{A}$, augmented matrix: $[\mathbf{A} \mid \mathbf{b}]$.

**Existence of solutions** or **consistency**:

$$\begin{aligned} \mathbf{Ax} = \mathbf{b} &\quad \text{has a solution} \\ \Leftrightarrow &\quad \mathbf{b} \in Range(\mathbf{A}) \\ \Leftrightarrow &\quad Rank(\mathbf{A}) = Rank([\mathbf{A} \mid \mathbf{b}]) \end{aligned}$$

**Uniqueness of solutions**:

$$\begin{aligned} Rank(\mathbf{A}) &= Rank([\mathbf{A} \mid \mathbf{b}]) = n \\ &\Leftrightarrow \text{ Solution of } \mathbf{Ax} = \mathbf{b} \text{ is unique.} \\ &\Leftrightarrow \mathbf{Ax} = \mathbf{0} \text{ has only the trivial (zero) solution.} \end{aligned}$$

**Infinite solutions**: For $Rank(\mathbf{A}) = Rank([\mathbf{A}|\mathbf{b}]) = k < n$, solution

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{x}_N, \quad \text{with} \quad \mathbf{A}\bar{\mathbf{x}} = \mathbf{b} \quad \text{and} \quad \mathbf{x}_N \in Null(\mathbf{A})$$

## Basic Idea of Solution Methodology

To diagnose the non-existence of a solution,

To determine the unique solution, or

To describe infinite solutions;

*decouple the equations* using **elementary transformations**.

For solving $\mathbf{Ax} = \mathbf{b}$, apply suitable elementary row transformations on both sides, leading to

$$\begin{aligned} \mathbf{R}_q\mathbf{R}_{q-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{Ax} &= \mathbf{R}_q\mathbf{R}_{q-1}\cdots\mathbf{R}_2\mathbf{R}_1\mathbf{b}, \\ \text{or,}\quad [\mathbf{RA}]\mathbf{x} &= \mathbf{Rb}; \end{aligned}$$

such that matrix $[\mathbf{RA}]$ is greatly simplified.
In the best case, with complete reduction, $\mathbf{RA} = \mathbf{I}_n$, and components of $\mathbf{x}$ can be read off from $\mathbf{Rb}$.

For inverting matrix $\mathbf{A}$, treat $\mathbf{AA}^{-1} = \mathbf{I}_n$ similarly.



## Homogeneous Systems

To solve $\mathbf{Ax} = \mathbf{0}$ or to describe $Null(\mathbf{A})$,
apply a series of elementary row transformations on $\mathbf{A}$ to reduce it to the $\widetilde{\mathbf{A}}$,

the **row-reduced echelon form** or **RREF**.

Features of RREF:

1. The first non-zero entry in any row is a '1', the leading '1'.
2. In the same column as the leading '1', other entries are zero.
3. Non-zero entries in a lower row appear later.

*Variables corresponding to columns having leading '1's are expressed in terms of the remaining variables.*

Solution of $\mathbf{Ax} = \mathbf{0}$: $\mathbf{x} = \begin{bmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \cdots & \mathbf{z}_{n-k} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_{n-k} \end{bmatrix}$

Basis of $Null(\mathbf{A})$: $\{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_{n-k}\}$



## Pivoting

**Attempt:**
To get '1' at diagonal (or leading) position, with '0' elsewhere.
**Key step:** *division* by the diagonal (or leading) entry.
Consider

$$\bar{\mathbf{A}} = \begin{bmatrix} \mathbf{I}_k & . & . & . & . & . \\ . & \delta & . & . & . & . \\ . & . & . & . & \mathrm{BIG} & . \\ . & \mathrm{big} & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \end{bmatrix}.$$

Cannot divide by zero. Should not divide by $\delta$.

- **partial pivoting:** row interchange to get 'big' in place of $\delta$
- **complete pivoting:** row and column interchanges to get 'BIG' in place of $\delta$

Complete pivoting does not give a huge advantage over partial pivoting, but requires maintaining of variable permutation for later unscrambling.



## Partitioning and Block Operations

Equation $\mathbf{Ax} = \mathbf{y}$ can be written as

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} & \mathbf{A}_{13} \\ \mathbf{A}_{21} & \mathbf{A}_{22} & \mathbf{A}_{23} \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix},$$

with $\mathbf{x}_1$, $\mathbf{x}_2$ etc being themselves vectors (or matrices).

- For a valid partitioning, block sizes should be consistent.
- Elementary transformations can be applied over blocks.
- Block operations can be computationally economical at times.
- Conceptually, different blocks of contributions/equations can be *assembled* for mathematical modelling of complicated coupled systems.

## Points to note

- Solution(s) of $\mathbf{Ax} = \mathbf{b}$ may be non-existent, unique or infinitely many.
- Complete solution can be described by composing a particular solution with the null space of $\mathbf{A}$.
- Null space basis can be obtained conveniently from the row-reduced echelon form of $\mathbf{A}$.
- For a *strategy* of solution, pivoting is an important step.

Necessary Exercises: **1,2,4,5,7**



## Outline





## Gauss-Jordan Elimination

**Task:** Solve $\mathbf{Ax} = \mathbf{b}_1$, $\mathbf{Ax} = \mathbf{b}_2$ and $\mathbf{Ax} = \mathbf{b}_3$; find $\mathbf{A}^{-1}$ and evaluate $\mathbf{A}^{-1}\mathbf{B}$, where $\mathbf{A} \in R^{n\times n}$ and $\mathbf{B} \in R^{n\times p}$.

Assemble $\mathbf{C} = [\mathbf{A} \quad \mathbf{b}_1 \quad \mathbf{b}_2 \quad \mathbf{b}_3 \quad \mathbf{I}_n \quad \mathbf{B}] \quad \in R^{n\times(2n+3+p)}$ and follow the algorithm.

Collect solutions from the result

$$\mathbf{C} \longrightarrow \widetilde{\mathbf{C}} = [\mathbf{I}_n \quad \mathbf{A}^{-1}\mathbf{b}_1 \quad \mathbf{A}^{-1}\mathbf{b}_2 \quad \mathbf{A}^{-1}\mathbf{b}_3 \quad \mathbf{A}^{-1} \quad \mathbf{A}^{-1}\mathbf{B}].$$

Remarks:

- Premature termination: matrix $\mathbf{A}$ singular — decision?
- If you use complete pivoting, unscramble permutation.
- Identity matrix in both $\mathbf{C}$ and $\widetilde{\mathbf{C}}$? Store $\mathbf{A}^{-1}$ 'in place'.
- For *evaluating* $\mathbf{A}^{-1}\mathbf{b}$, do not develop $\mathbf{A}^{-1}$.
- Gauss-Jordan elimination an overkill? Want something cheaper?



## Gauss-Jordan Elimination

**Gauss-Jordan Algorithm**

- $\Delta = 1$
- For $k = 1, 2, 3, \cdots, (n-1)$
  1. Pivot : identify $l$ such that $|c_{lk}| = \max |c_{jk}|$ for $k \le j \le n$.
     If $c_{lk} = 0$, then $\Delta = 0$ and **exit**.
     Else, interchange row $k$ and row $l$.
  2. $\Delta \longleftarrow c_{kk}\Delta$,
     Divide row $k$ by $c_{kk}$.
  3. Subtract $c_{jk}$ times row $k$ from row $j$, $\forall j \ne k$.
- $\Delta \longleftarrow c_{nn}\Delta$
  If $c_{nn} = 0$, then **exit**.
  Else, divide row $n$ by $c_{nn}$.

In case of non-singular $\mathbf{A}$, default termination.

This outline is for partial pivoting.

## Gaussian Elimination with Back-Substitution

Gaussian elimination:

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ \longrightarrow \tilde{\mathbf{A}}\mathbf{x} &= \tilde{\mathbf{b}} \end{aligned}$$

$$\text{or,} \quad \begin{bmatrix} a'_{11} & a'_{12} & \cdots & a'_{1n} \\ & a'_{22} & \cdots & a'_{2n} \\ & & \ddots & \vdots \\ & & & a'_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b'_1 \\ b'_2 \\ \vdots \\ b'_n \end{bmatrix}$$

Back-substitutions:

$$\begin{aligned} x_n &= b'_n / a'_{nn}, \\ x_i &= \frac{1}{a'_{ii}} \left[ b'_i - \sum_{j=i+1}^{n} a'_{ij} x_j \right] \quad \text{for} \quad i = n-1, n-2, \cdots, 2, 1 \end{aligned}$$

Remarks

- Computational cost half compared to G-J elimination.
- Like G-J elimination, prior knowledge of RHS needed.



## Gaussian Elimination with Back-Substitution

**Anatomy of the Gaussian elimination:**

The process of Gaussian elimination (with no pivoting) leads to

$$\mathbf{U} = \mathbf{R}_q \mathbf{R}_{q-1} \cdots \mathbf{R}_2 \mathbf{R}_1 \mathbf{A} = \mathbf{RA}.$$

The steps given by

*for* $k = 1, 2, 3, \cdots, (n-1)$

$j$*-th row* $\longleftarrow$ $j$*-th row* $- \frac{a_{jk}}{a_{kk}} \times$ $k$*-th row* *for*
$j = k+1, k+2, \cdots, n$

involve elementary matrices

$$\mathbf{R}_k|_{k=1} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 & \cdots & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\frac{a_{n1}}{a_{11}} & 0 & 0 & \cdots & 1 \end{bmatrix} \quad etc.$$

With $\mathbf{L} = \mathbf{R}^{-1}$, $\quad \mathbf{A} = \mathbf{LU}$.



## LU Decomposition

A square matrix with non-zero leading minors is LU-decomposable.

No reference to a right-hand-side (RHS) vector!

To solve $\mathbf{Ax} = \mathbf{b}$, denote $\mathbf{y} = \mathbf{Ux}$ and split as

$$\begin{aligned} \mathbf{Ax} = \mathbf{b} \quad &\Rightarrow \quad \mathbf{LUx} = \mathbf{b} \\ &\Rightarrow \quad \mathbf{Ly} = \mathbf{b} \quad \text{and} \quad \mathbf{Ux} = \mathbf{y}. \end{aligned}$$

Forward substitutions:

$$y_i = \frac{1}{l_{ii}} \left( b_i - \sum_{j=1}^{i-1} l_{ij} y_j \right) \quad \text{for } i = 1, 2, 3, \cdots, n;$$

Back-substitutions:

$$x_i = \frac{1}{u_{ii}} \left( y_i - \sum_{j=i+1}^{n} u_{ij} x_j \right) \quad \text{for } i = n, n-1, n-2, \cdots, 1.$$



## LU Decomposition

**Question:** How to LU-decompose a given matrix?

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 & \cdots & 0 \\ l_{21} & l_{22} & 0 & \cdots & 0 \\ l_{31} & l_{32} & l_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & l_{nn} \end{bmatrix} \quad \text{and} \quad \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ 0 & u_{22} & u_{23} & \cdots & u_{2n} \\ 0 & 0 & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & u_{nn} \end{bmatrix}$$

Elements of the product give

$$\sum_{k=1}^{i} l_{ik} u_{kj} = a_{ij} \quad \text{for} \quad i \le j,$$

$$\text{and} \quad \sum_{k=1}^{j} l_{ik} u_{kj} = a_{ij} \quad \text{for} \quad i > j.$$

$n^2$ equations in $n^2 + n$ unknowns: choice of $n$ unknowns

# LU Decomposition

**Doolittle's algorithm**

- Choose $l_{ii} = 1$
- For $j = 1, 2, 3, \cdots, n$
  1. $u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}$ for $1 \le i \le j$
  2. $l_{ij} = \frac{1}{u_{jj}}(a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj})$ for $i > j$

Evaluation proceeds in column order of the matrix (for storage)

$$\mathbf{A}^* = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \cdots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \cdots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \cdots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \cdots & u_{nn} \end{bmatrix}$$



# LU Decomposition

**Question:** What about matrices which are *not* LU-decomposable?
**Question:** What about pivoting?

Consider the non-singular matrix

$$\begin{bmatrix} 0 & 1 & 2 \\ 3 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} = ? & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} = 0 & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}.$$

LU-decompose a permutation of its rows

$$\begin{aligned} \begin{bmatrix} 0 & 1 & 2 \\ 3 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix} &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 2 \\ 0 & 1 & 2 \\ 2 & 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{2}{3} & \frac{1}{3} & 1 \end{bmatrix} \begin{bmatrix} 3 & 1 & 2 \\ 0 & 1 & 2 \\ 0 & 0 & 1 \end{bmatrix}. \end{aligned}$$

In this **PLU** decomposition, permutation **P** is recorded in a vector.



# Points to note

For invertible coefficient matrices, use

- Gauss-Jordan elimination for large number of RHS vectors available all together and also for matrix inversion,
- Gaussian elimination with back-substitution for small number of RHS vectors available together,
- LU decomposition method to develop and maintain factors to be used as and when RHS vectors are available.

Pivoting is almost necessary (without further special structure).

Necessary Exercises: **1,4,5**



# Outline

## Quadratic Forms, Symmetry and Positive Definiteness

Quadratic form

$$q(\mathbf{x}) = \mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

*defined with respect to a symmetric matrix.*

Quadratic form $q(\mathbf{x})$, equivalently matrix $\mathbf{A}$, is called positive definite (p.d.) when

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0 \quad \forall\, \mathbf{x} \neq \mathbf{0}$$

and positive semi-definite (p.s.d.) when

$$\mathbf{x}^T \mathbf{A} \mathbf{x} \geq 0 \quad \forall\, \mathbf{x} \neq \mathbf{0}.$$

Sylvester's criteria:

$$a_{11} \geq 0, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \geq 0, \quad \cdots, \quad \det \mathbf{A} \geq 0;$$

i.e. all *leading minors* non-negative, for p.s.d.



## Cholesky Decomposition

If $\mathbf{A} \in R^{n \times n}$ is symmetric and positive definite, then there exists a non-singular lower triangular matrix $\mathbf{L} \in R^{n \times n}$ such that

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T.$$

**Algorithm** For $i = 1, 2, 3, \cdots, n$

- $L_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} L_{ik}^2}$
- $L_{ji} = \frac{1}{L_{ii}} \left( a_{ji} - \sum_{k=1}^{i-1} L_{jk} L_{ik} \right) \quad$ for $\; i < j \leq n$

For solving $\mathbf{A}\mathbf{x} = \mathbf{b}$,

Forward substitutions: $\mathbf{L}\mathbf{y} = \mathbf{b}$

Back-substitutions: $\mathbf{L}^T \mathbf{x} = \mathbf{y}$

Remarks

- Test of positive definiteness.
- Stable algorithm: no pivoting necessary!
- Economy of space and time.



## Sparse Systems*

- What is a sparse matrix?
- Bandedness and bandwidth
- Efficient storage and processing
- Updates
  - Sherman-Morrison formula

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

  - Woodbury formula
- Conjugate gradient method
  - efficiently implemented matrix-vector products



## Points to note

- Concepts and criteria of positive definiteness and positive semi-definiteness
- Cholesky decomposition method in symmetric positive definite systems
- Nature of sparsity and its exploitation

Necessary Exercises: **1,2,4,7**

# Outline







# Norms and Condition Numbers



**Norm of a vector:** a measure of size

- Euclidean norm or 2-norm

$$\|\mathbf{x}\| = \|\mathbf{x}\|_2 = \left[x_1^2 + x_2^2 + \cdots + x_n^2\right]^{\frac{1}{2}} = \sqrt{\mathbf{x}^T\mathbf{x}}$$

- The $p$-norm

$$\|\mathbf{x}\|_p = [|x_1|^p + |x_2|^p + \cdots + |x_n|^p]^{\frac{1}{p}}$$

- The 1-norm: $\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$
- The $\infty$-norm:

$$\|\mathbf{x}\|_\infty = \lim_{p\to\infty} [|x_1|^p + |x_2|^p + \cdots + |x_n|^p]^{\frac{1}{p}} = \max_j |x_j|$$

- Weighted norm

$$\|\mathbf{x}\|_{\mathbf{w}} = \sqrt{\mathbf{x}^T\mathbf{W}\mathbf{x}}$$

where weight matrix $\mathbf{W}$ is symmetric and positive definite.



# Norms and Condition Numbers



**Norm of a matrix:** magnitude or scale of the transformation

Matrix norm (induced by a vector norm) is given by the largest magnification it can produce on a vector

$$\|\mathbf{A}\| = \max_{\mathbf{x}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \max_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|$$

Direct consequence: $\|\mathbf{A}\mathbf{x}\| \le \|\mathbf{A}\|\ \|\mathbf{x}\|$

**Index of closeness to singularity:** Condition number

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\ \|\mathbf{A}^{-1}\|, \quad 1 \le \kappa(\mathbf{A}) \le \infty$$

** Isotropic, well-conditioned, ill-conditioned and singular matrices



# Ill-conditioning and Sensitivity



$$\begin{array}{rcrcl} 0.9999x_1 & - & 1.0001x_2 & = & 1 \\ x_1 & - & x_2 & = & 1+\epsilon \end{array}$$

Solution: $x_1 = \frac{10001\epsilon+1}{2}$, $x_2 = \frac{9999\epsilon-1}{2}$

- sensitive to small changes in the RHS
- insensitive to error in a guess

See illustration

For the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, solution is $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ and

$$\delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b} - \mathbf{A}^{-1}\delta\mathbf{A}\ \mathbf{x}$$

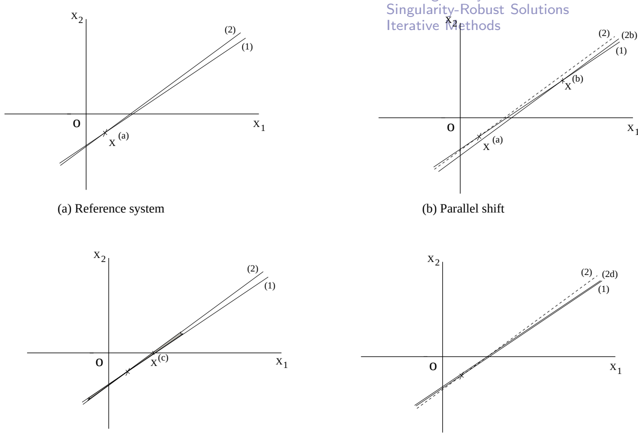
If the matrix $\mathbf{A}$ is exactly known, then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\|\ \|\mathbf{A}^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \kappa(\mathbf{A})\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

If the RHS is known exactly, then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \le \|\mathbf{A}\|\ \|\mathbf{A}^{-1}\|\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} = \kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

## Ill-conditioning and Sensitivity



(a) Reference system (b) Parallel shift

(c) Guess validation (d) Singularity

Figure: Ill-conditioning: a geometric perspective



## Rectangular Systems

Consider $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in R^{m\times n}$ and $Rank(\mathbf{A}) = n < m$.

$$\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b} \quad \Rightarrow \quad \mathbf{x} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$$

Square of error norm

$$\begin{aligned} U(\mathbf{x}) &= \frac{1}{2}\|\mathbf{Ax}-\mathbf{b}\|^2 = \frac{1}{2}(\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) \\ &= \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} + \frac{1}{2}\mathbf{b}^T\mathbf{b} \end{aligned}$$

Least square error solution:

$$\frac{\partial U}{\partial \mathbf{x}} = \mathbf{A}^T\mathbf{Ax} - \mathbf{A}^T\mathbf{b} = \mathbf{0}$$

Pseudoinverse or Moore-Penrose inverse or *left-inverse*

$$\mathbf{A}^{\#} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$$



## Rectangular Systems

Consider $\mathbf{Ax} = \mathbf{b}$ with $\mathbf{A} \in R^{m\times n}$ and $Rank(\mathbf{A}) = m < n$.
Look for $\boldsymbol{\lambda} \in R^m$ that satisfies $\mathbf{A}^T\boldsymbol{\lambda} = \mathbf{x}$ and

$$\mathbf{AA}^T\boldsymbol{\lambda} = \mathbf{b}$$

Solution

$$\mathbf{x} = \mathbf{A}^T\boldsymbol{\lambda} = \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{b}$$

Consider the problem

$$\textit{minimize} \;\; U(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{x} \quad \textit{subject to} \;\; \mathbf{Ax} = \mathbf{b}.$$

Extremum of the Lagrangian $\mathcal{L}(\mathbf{x},\boldsymbol{\lambda}) = \frac{1}{2}\mathbf{x}^T\mathbf{x} - \boldsymbol{\lambda}^T(\mathbf{Ax}-\mathbf{b})$ is given by

$$\frac{\partial \mathcal{L}}{\partial \mathbf{x}} = \mathbf{0},\; \frac{\partial \mathcal{L}}{\partial \boldsymbol{\lambda}} = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{A}^T\boldsymbol{\lambda},\; \mathbf{Ax} = \mathbf{b}.$$

Solution $\mathbf{x} = \mathbf{A}^T(\mathbf{AA}^T)^{-1}\mathbf{b}$ gives foot of the perpendicular on the solution 'plane' and the pseudoinverse

$$\mathbf{A}^{\#} = \mathbf{A}^T(\mathbf{AA}^T)^{-1}$$

here is a *right-inverse*!



## Singularity-Robust Solutions

**Ill-posed problems**: *Tikhonov regularization*

- recipe for *any* linear system ($m > n$, $m = n$ or $m < n$), with any condition!

$\mathbf{Ax} = \mathbf{b}$ may have conflict: form $\mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b}$.

$\mathbf{A}^T\mathbf{A}$ may be ill-conditioned: rig the system as

$$(\mathbf{A}^T\mathbf{A} + \nu^2\mathbf{I}_n)\mathbf{x} = \mathbf{A}^T\mathbf{b}$$

Coefficient matrix: symmetric and positive definite!
*The idea:* Immunize the system, paying a small price.

Issues:

- The choice of $\nu$?
- When $m < n$, computational advantage by

$$(\mathbf{AA}^T + \nu^2\mathbf{I}_m)\boldsymbol{\lambda} = \mathbf{b}, \quad \mathbf{x} = \mathbf{A}^T\boldsymbol{\lambda}$$

# Iterative Methods

Jacobi's iteration method:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1,\ j\neq i}^{n} a_{ij} x_j^{(k)} \right) \quad \text{for} \quad i = 1, 2, 3, \cdots, n.$$

Gauss-Seidel method:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left( b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \quad \text{for} \quad i = 1, 2, 3, \cdots, n.$$

The category of relaxation methods:

*diagonal dominance and availability of good initial approximations*



# Points to note

- Solutions are unreliable when the coefficient matrix is ill-conditioned.
- Finding pseudoinverse of a *full-rank* matrix is 'easy'.
- Tikhonov regularization provides singularity-robust solutions.
- Iterative methods may have an edge in certain situations!

Necessary Exercises: **1,2,3,4**



# Outline





# Eigenvalue Problem

In mapping $\mathbf{A} : R^n \to R^n$, special vectors of matrix $\mathbf{A} \in R^{n\times n}$

- mapped to scalar multiples, i.e. undergo pure scaling

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v}$$

Eigenvector ($\mathbf{v}$) and eigenvalue ($\lambda$): eigenpair ($\lambda, \mathbf{v}$)

algebraic eigenvalue problem

$$(\lambda \mathbf{I} - \mathbf{A})\mathbf{v} = \mathbf{0}$$

For non-trivial (non-zero) solution $\mathbf{v}$,

$$\det(\lambda \mathbf{I} - \mathbf{A}) = 0$$

Characteristic equation: characteristic polynomial: $n$ roots

- $n$ eigenvalues — for each, find eignevector(s)

Multiplicity of an eigenvalue: *algebraic* and *geometric*
Multiplicity mismatch: *diagonalizable* and *defective* matrices

# Generalized Eigenvalue Problem



1-dof mass-spring system: $m\ddot{x} + kx = 0$

*Natural frequency of vibration:* $\omega_n = \sqrt{\frac{k}{m}}$

Free vibration of n-dof system:

$$\mathbf{M}\ddot{\mathbf{x}} + \mathbf{K}\mathbf{x} = \mathbf{0},$$

Natural frequencies and corresponding modes?
Assuming a vibration mode $\mathbf{x} = \Phi \sin(\omega t + \alpha)$,

$$(-\omega^2 \mathbf{M}\Phi + \mathbf{K}\Phi)\sin(\omega t + \alpha) = \mathbf{0} \Rightarrow \boxed{\mathbf{K}\Phi = \omega^2 \mathbf{M}\Phi}$$

Reduce as $\left(\mathbf{M}^{-1}\mathbf{K}\right)\Phi = \omega^2\Phi$? *Why is it not a good idea?*

**K** *symmetric,* **M** *symmetric and positive definite!!*

With $\mathbf{M} = \mathbf{L}\mathbf{L}^T$, $\tilde{\mathbf{\Phi}} = \mathbf{L}^T\Phi$ and $\tilde{\mathbf{K}} = \mathbf{L}^{-1}\mathbf{K}\mathbf{L}^{-T}$,

$$\tilde{\mathbf{K}}\tilde{\mathbf{\Phi}} = \omega^2\tilde{\mathbf{\Phi}}$$



# Some Basic Theoretical Results



**Eigenvalues of transpose**

*Eigenvalues of* $\mathbf{A}^T$ *are the same as those of* $\mathbf{A}$.

Caution: Eigenvectors of $\mathbf{A}$ and $\mathbf{A}^T$ need not be same.

**Diagonal and block diagonal matrices**

Eigenvalues of a diagonal matrix are its diagonal entries.
Corresponding eigenvectors: natural basis members ($\mathbf{e}_1$, $\mathbf{e}_2$ etc).

Eigenvalues of a block diagonal matrix: those of diagonal blocks.
Eigenvectors: coordinate extensions of individual eigenvectors.
With $(\lambda_2, \mathbf{v}_2)$ as eigenpair of block $\mathbf{A}_2$,

$$\mathbf{A}\tilde{\mathbf{v}}_2 = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{v}_2 \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{A}_2\mathbf{v}_2 \\ \mathbf{0} \end{bmatrix} = \lambda_2 \begin{bmatrix} \mathbf{0} \\ \mathbf{v}_2 \\ \mathbf{0} \end{bmatrix}$$



# Some Basic Theoretical Results



**Triangular and block triangular matrices**

Eigenvalues of a triangular matrix are its diagonal entries.

Eigenvalues of a block triangular matrix are the collection of eigenvalues of its diagonal blocks.

Take

$$\mathbf{H} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}, \quad \mathbf{A} \in R^{r \times r} \text{ and } \mathbf{C} \in R^{s \times s}$$

If $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$, then

$$\mathbf{H}\begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{0} & \mathbf{C} \end{bmatrix}\begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{v} \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \lambda\mathbf{v} \\ \mathbf{0} \end{bmatrix} = \lambda\begin{bmatrix} \mathbf{v} \\ \mathbf{0} \end{bmatrix}$$

If $\mu$ is an eigenvalue of $\mathbf{C}$, then it is also an eigenvalue of $\mathbf{C}^T$ and

$$\mathbf{C}^T\mathbf{w} = \mu\mathbf{w} \Rightarrow \mathbf{H}^T\begin{bmatrix} \mathbf{0} \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^T & \mathbf{0} \\ \mathbf{B}^T & \mathbf{C}^T \end{bmatrix}\begin{bmatrix} \mathbf{0} \\ \mathbf{w} \end{bmatrix} = \mu\begin{bmatrix} \mathbf{0} \\ \mathbf{w} \end{bmatrix}$$



# Some Basic Theoretical Results



**Shift theorem**

Eigenvectors of $\mathbf{A} + \mu\mathbf{I}$ are the same as those of $\mathbf{A}$.
Eigenvalues: shifted by $\mu$.

**Deflation**

For a symmetric matrix $\mathbf{A}$, with mutually orthogonal eigenvectors, having $(\lambda_j, \mathbf{v}_j)$ as an eigenpair,

$$\mathbf{B} = \mathbf{A} - \lambda_j \frac{\mathbf{v}_j\mathbf{v}_j^T}{\mathbf{v}_j^T\mathbf{v}_j}$$

has the same eigenstructure as $\mathbf{A}$, *except that the eigenvalue corresponding to* $\mathbf{v}_j$ *is zero.*

## Some Basic Theoretical Results



**Eigenspace**

If $\mathbf{v}_1$, $\mathbf{v}_2$, $\cdots$, $\mathbf{v}_k$ are eigenvectors of $\mathbf{A}$ corresponding to the *same* eigenvalue $\lambda$, then

*eigenspace:* $<\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k>$

**Similarity transformation**

$\mathbf{B} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S}$: same transformation expressed in new basis.

$$\det(\lambda\mathbf{I} - \mathbf{A}) = \det \mathbf{S}^{-1} \ \det(\lambda\mathbf{I} - \mathbf{A}) \ \det \mathbf{S} = \det(\lambda\mathbf{I} - \mathbf{B})$$

Same characteristic polynomial!

*Eigenvalues are the property of a linear transformation, not of the basis.*

An eigenvector $\mathbf{v}$ of $\mathbf{A}$ transforms to $\mathbf{S}^{-1}\mathbf{v}$, as the corresponding eigenvector of $\mathbf{B}$.



## Power Method



Consider matrix $\mathbf{A}$ with

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_{n-1}| > |\lambda_n|$$

and a full set of $n$ eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\cdots$, $\mathbf{v}_n$.

For vector $\mathbf{x} = \alpha_1\mathbf{v}_1 + \alpha_2\mathbf{v}_2 + \cdots + \alpha_n\mathbf{v}_n$,

$$\mathbf{A}^p\mathbf{x} = \lambda_1^p \left[\alpha_1\mathbf{v}_1 + \left(\frac{\lambda_2}{\lambda_1}\right)^p \alpha_2\mathbf{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^p \alpha_3\mathbf{v}_3 + \cdots + \left(\frac{\lambda_n}{\lambda_1}\right)^p \alpha_n\mathbf{v}_n\right]$$

As $p \to \infty$, $\mathbf{A}^p\mathbf{x} \to \lambda_1^p\alpha_1\mathbf{v}_1$, and

$$\lambda_1 = \lim_{p\to\infty} \frac{(\mathbf{A}^p\mathbf{x})_r}{(\mathbf{A}^{p-1}\mathbf{x})_r}, \quad r = 1, 2, 3, \cdots, n.$$

At convergence, $n$ ratios will be the same.

**Question:** *How to find the least magnitude eigenvalue?*



## Points to note



- Meaning and context of the algebraic eigenvalue problem
- Fundamental deductions and vital relationships
- Power method as an inexpensive procedure to determine extremal magnitude eigenvalues

Necessary Exercises: **1,2,3,4,6**



## Outline

# Diagonalizability



Consider $\mathbf{A} \in R^{n\times n}$, having $n$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$; with corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$.

$$
\begin{aligned}
\mathbf{AS} &= \mathbf{A}[\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] = [\lambda_1\mathbf{v}_1 \quad \lambda_2\mathbf{v}_2 \quad \cdots \quad \lambda_n\mathbf{v}_n] \\
&= [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} = \mathbf{S}\Lambda
\end{aligned}
$$

$$\Rightarrow \mathbf{A} = \mathbf{S}\Lambda\mathbf{S}^{-1} \quad \text{and} \quad \mathbf{S}^{-1}\mathbf{AS} = \Lambda$$

Diagonalization: The process of changing the basis of a linear transformation so that its new matrix representation is diagonal, i.e. so that it is decoupled among its coordinates.



# Diagonalizability



**Diagonalizability:**

*A matrix having a complete set of n linearly independent eigenvectors is diagonalizable.*

**Existence of a complete set of eigenvectors:**

*A diagonalizable matrix possesses a complete set of n linearly independent eigenvectors.*

- All distinct eigenvalues implies *diagonalizability*.
- But, diagonalizability does **not** imply distinct eigenvalues!
- However, a *lack* of diagonalizability certainly implies a *multiplicity mismatch*.



# Canonical Forms



- *Jordan canonical form (JCF)*
- *Diagonal (canonical) form*
- *Triangular (canonical) form*

*Other convenient forms*

- *Tridiagonal form*
- *Hessenberg form*



# Canonical Forms



**Jordan canonical form (JCF):** composed of Jordan blocks

$$\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & & & \\ & \mathbf{J}_2 & & \\ & & \ddots & \\ & & & \mathbf{J}_k \end{bmatrix}, \quad \mathbf{J}_r = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \lambda & \ddots & \\ & & & \ddots & 1 \\ & & & & \lambda \end{bmatrix}$$

The key equation $\mathbf{AS} = \mathbf{SJ}$ in extended form gives

$$\mathbf{A}[\cdots \quad \mathbf{S}_r \quad \cdots] = [\cdots \quad \mathbf{S}_r \quad \cdots] \begin{bmatrix} \ddots & & \\ & \mathbf{J}_r & \\ & & \ddots \end{bmatrix},$$

where Jordan block $\mathbf{J}_r$ is associated with the subspace of

$$\mathbf{S}_r = [\mathbf{v} \quad \mathbf{w}_2 \quad \mathbf{w}_3 \quad \cdots]$$

# Canonical Forms

Equating blocks as $\mathbf{A}\mathbf{S}_r = \mathbf{S}_r\mathbf{J}_r$ gives

$$[\mathbf{A}\mathbf{v} \quad \mathbf{A}\mathbf{w}_2 \quad \mathbf{A}\mathbf{w}_3 \quad \cdots] = [\mathbf{v} \quad \mathbf{w}_2 \quad \mathbf{w}_3 \quad \cdots] \begin{bmatrix} \lambda & 1 & & \\ & \lambda & 1 & \\ & & \lambda & \ddots \\ & & & \ddots \end{bmatrix}$$

Columnwise equality leads to

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v}, \quad \mathbf{A}\mathbf{w}_2 = \mathbf{v} + \lambda\mathbf{w}_2, \quad \mathbf{A}\mathbf{w}_3 = \mathbf{w}_2 + \lambda\mathbf{w}_3, \quad \cdots$$

Generalized eigenvectors $\mathbf{w}_2$, $\mathbf{w}_3$ etc:

$$\begin{aligned} (\mathbf{A} - \lambda\mathbf{I})\mathbf{v} &= \mathbf{0}, \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{w}_2 &= \mathbf{v} \quad \text{and} \quad (\mathbf{A} - \lambda\mathbf{I})^2\mathbf{w}_2 = \mathbf{0}, \\ (\mathbf{A} - \lambda\mathbf{I})\mathbf{w}_3 &= \mathbf{w}_2 \quad \text{and} \quad (\mathbf{A} - \lambda\mathbf{I})^3\mathbf{w}_3 = \mathbf{0}, \quad \cdots \end{aligned}$$



# Canonical Forms

**Diagonal form**

- Special case of Jordan form, with each Jordan block of $1 \times 1$ size
- Matrix is diagonalizable
- Similarity transformation matrix $\mathbf{S}$ is composed of $n$ linearly independent eigenvectors as columns
- None of the eigenvectors admits any *generalized eigenvector*
- Equal geometric and algebraic multiplicities for every eigenvalue



# Canonical Forms

**Triangular form**

Triangularization: Change of basis of a linear tranformation so as to get its matrix in the triangular form

- For real eigenvalues, always possible to accomplish with orthogonal similarity transformation
- Always possible to accomplish with unitary similarity transformation, with complex arithmetic
- Determination of eigenvalues

Note: The case of complex eigenvalues: $2 \times 2$ real diagonal block

$$\begin{bmatrix} \alpha & -\beta \\ \beta & \alpha \end{bmatrix} \sim \begin{bmatrix} \alpha + i\beta & 0 \\ 0 & \alpha - i\beta \end{bmatrix}$$



# Canonical Forms

Forms that can be obtained with pre-determined number of arithmetic operations (without iteration):

Tridiagonal form: non-zero entries only in the (leading) diagonal, sub-diagonal and super-diagonal

- useful for symmetric matrices

Hessenberg form: A slight generalization of a triangular matrix

$$\mathbf{H}_u = \begin{bmatrix} * & * & * & \cdots & * & * \\ * & * & * & \cdots & * & * \\ & * & * & \cdots & * & * \\ & & \ddots & \ddots & \vdots & \vdots \\ & & & \ddots & \ddots & \vdots \\ & & & & * & * \end{bmatrix}$$

**Note:** Tridiagonal and Hessenberg forms do not fall in the category of canonical forms.

## Symmetric Matrices



*A real symmetric matrix has all real eigenvalues and is diagonalizable through an orthogonal similarity transformation.*

- Eigenvalues must be real.
- A complete set of eigenvectors exists.
- Eigenvectors corresponding to distinct eigenvalues are necessarily orthogonal.
- Corresponding to repeated eigenvalues, orthogonal eigenvectors are available.

*In all cases of a symmetric matrix, we can form an orthogonal matrix* $\mathbf{V}$*, such that* $\mathbf{V}^T\mathbf{A}\mathbf{V} = \Lambda$ *is a real diagonal matrix.*

Further, $\mathbf{A} = \mathbf{V}\Lambda\mathbf{V}^T$.

Similar results for complex *Hermitian* matrices.



## Symmetric Matrices



**Proposition:** Eigenvalues of a real symmetric matrix must be real.

Take $\mathbf{A} \in R^{n\times n}$ such that $\mathbf{A} = \mathbf{A}^T$, with eigenvalue $\lambda = h + ik$.

Since $\lambda\mathbf{I} - \mathbf{A}$ is singular, so is

$$\begin{aligned} \mathbf{B} &= (\lambda\mathbf{I} - \mathbf{A})\,(\bar{\lambda}\mathbf{I} - \mathbf{A}) = (h\mathbf{I} - \mathbf{A} + ik\mathbf{I})(h\mathbf{I} - \mathbf{A} - ik\mathbf{I}) \\ &= (h\mathbf{I} - \mathbf{A})^2 + k^2 I \end{aligned}$$

For some $\mathbf{x} \neq \mathbf{0}$, $\mathbf{B}\mathbf{x} = \mathbf{0}$, and

$$\mathbf{x}^T\mathbf{B}\mathbf{x} = 0 \Rightarrow \mathbf{x}^T(h\mathbf{I} - \mathbf{A})^T(h\mathbf{I} - \mathbf{A})\mathbf{x} + k^2\mathbf{x}^T\mathbf{x} = 0$$

Thus, $\|(h\mathbf{I} - \mathbf{A})\mathbf{x}\|^2 + \|k\mathbf{x}\|^2 = 0$

$$\boxed{k = 0 \text{ and } \lambda = h}$$



## Symmetric Matrices



**Proposition:** A symmetric matrix possesses a complete set of eigenvectors.

Consider a repeated real eigenvalue $\lambda$ of $\mathbf{A}$ and examine its Jordan block(s).

Suppose $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$.
The first generalized eigenvector $\mathbf{w}$ satisfies $(\mathbf{A} - \lambda\mathbf{I})\mathbf{w} = \mathbf{v}$, giving

$$\begin{aligned} \mathbf{v}^T(\mathbf{A} - \lambda\mathbf{I})\mathbf{w} = \mathbf{v}^T\mathbf{v} &\Rightarrow \mathbf{v}^T\mathbf{A}^T\mathbf{w} - \lambda\mathbf{v}^T\mathbf{w} = \mathbf{v}^T\mathbf{v} \\ &\Rightarrow (\mathbf{A}\mathbf{v})^T\mathbf{w} - \lambda\mathbf{v}^T\mathbf{w} = \|\mathbf{v}\|^2 \\ &\Rightarrow \|\mathbf{v}\|^2 = 0 \end{aligned}$$

which is absurd.

An eigenvector will not admit a generalized eigenvector.

$$\boxed{\text{All Jordan blocks will be of } 1 \times 1 \text{ size.}}$$



## Symmetric Matrices



**Proposition:** Eigenvectors of a symmetric matrix corresponding to distinct eigenvalues are necessarily orthogonal.

Take two eigenpairs $(\lambda_1, \mathbf{v}_1)$ and $(\lambda_2, \mathbf{v}_2)$, with $\lambda_1 \neq \lambda_2$.

$$\begin{aligned} \mathbf{v}_1^T\mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1^T(\lambda_2\mathbf{v}_2) = \lambda_2\mathbf{v}_1^T\mathbf{v}_2 \\ \mathbf{v}_1^T\mathbf{A}\mathbf{v}_2 &= \mathbf{v}_1^T\mathbf{A}^T\mathbf{v}_2 = (\mathbf{A}\mathbf{v}_1)^T\mathbf{v}_2 = (\lambda_1\mathbf{v}_1)^T\mathbf{v}_2 = \lambda_1\mathbf{v}_1^T\mathbf{v}_2 \end{aligned}$$

From the two expressions, $(\lambda_1 - \lambda_2)\mathbf{v}_1^T\mathbf{v}_2 = 0$

$$\boxed{\mathbf{v}_1^T\mathbf{v}_2 = 0}$$

**Proposition:** Corresponding to a repeated eigenvalue of a symmetric matrix, an appropriate number of orthogonal eigenvectors can be selected.

If $\lambda_1 = \lambda_2$, then the entire subspace $< \mathbf{v}_1, \mathbf{v}_2 >$ is an eigenspace.
Select any two mutually orthogonal eigenvectors for the basis.

## Symmetric Matrices

Facilities with the 'omnipresent' symmetric matrices:

- Expression

$$\begin{aligned}
\mathbf{A} &= \mathbf{V}\Lambda\mathbf{V}^T \\
&= [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_n] \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_n^T \end{bmatrix} \\
&= \lambda_1 \mathbf{v}_1 \mathbf{v}_1^T + \lambda_2 \mathbf{v}_2 \mathbf{v}_2^T + \cdots + \lambda_n \mathbf{v}_n \mathbf{v}_n^T = \sum_{i=1}^{n} \lambda_i \mathbf{v}_i \mathbf{v}_i^T
\end{aligned}$$

- Reconstruction from a sum of rank-one components
- Efficient storage with only large eigenvalues and corresponding eigenvectors
- Deflation technique
- Stable and effective methods: easier to solve the eigenvalue problem



## Similarity Transformations

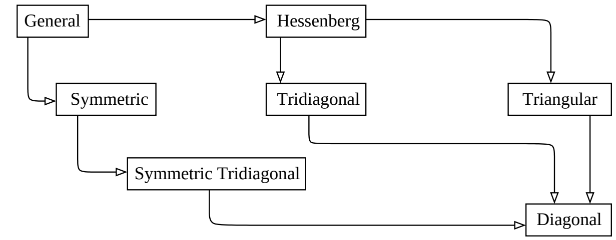



Figure: Eigenvalue problem: forms and steps

How to find suitable similarity transformations?

1. rotation
2. reflection
3. matrix decomposition or factorization
4. elementary transformation



## Points to note

- Generally possible reduction: Jordan canonical form
- Condition of diagonalizability and the diagonal form
- Possible with orthogonal similarity transformations: triangular form
- Useful non-canonical forms: tridiagonal and Hessenberg
- *Orthogonal diagonalization of symmetric matrices*

**Caution:** Each step in this context to be effected through similarity transformations

Necessary Exercises: **1,2,4**



## Outline

# Plane Rotations

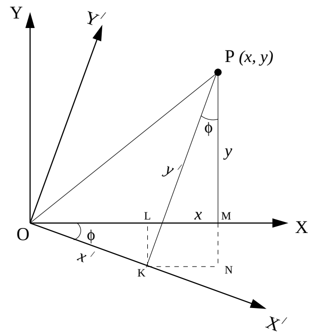



Figure: Rotation of axes and change of basis

$$\begin{aligned} x &= OL + LM = OL + KN = x' \cos\phi + y' \sin\phi \\ y &= PN - MN = PN - LK = y' \cos\phi - x' \sin\phi \end{aligned}$$



# Plane Rotations

Orthogonal change of basis:

$$\mathbf{r} = \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} x' \\ y' \end{bmatrix} = \Re \mathbf{r}'$$

Mapping of position vectors with

$$\Re^{-1} = \Re^T = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix}$$

In three-dimensional (ambient) space,

$$\Re_{xy} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix}, \ \Re_{xz} = \begin{bmatrix} \cos\phi & 0 & \sin\phi \\ 0 & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{bmatrix} \text{ etc.}$$



# Plane Rotations

Generalizing to $n$-dimensional Euclidean space ($R^n$),

$$\mathbf{P}_{pq} = \begin{bmatrix} 1 & & & & 0 & & & & 0 & & \\ & 1 & & & 0 & & & & 0 & & \\ & & \ddots & & \vdots & & & & \vdots & & \\ & & & 1 & 0 & & & & 0 & & \\ 0 & 0 & \cdots & 0 & c & 0 & \cdots & 0 & s & \cdots & 0 \\ & & & & 0 & 1 & & & 0 & & \\ & & & & \vdots & & \ddots & & \vdots & & \\ & & & & 0 & & & 1 & 0 & & \\ 0 & 0 & \cdots & 0 & -s & 0 & \cdots & 0 & c & \cdots & 0 \\ & & & & \vdots & & & & \vdots & \ddots & \\ & & & & 0 & & & & 0 & & 1 \end{bmatrix}$$

Matrix $\mathbf{A}$ is transformed as

$$\mathbf{A}' = \mathbf{P}_{pq}^{-1} \mathbf{A} \mathbf{P}_{pq} = \mathbf{P}_{pq}^T \mathbf{A} \mathbf{P}_{pq},$$

only the $p$-th and $q$-th rows and columns being affected.



# Jacobi Rotation Method

$$\begin{aligned} a'_{pr} = a'_{rp} &= c a_{rp} - s a_{rq} \text{ for } p \neq r \neq q, \\ a'_{qr} = a'_{rq} &= c a_{rq} + s a_{rp} \text{ for } p \neq r \neq q, \\ a'_{pp} &= c^2 a_{pp} + s^2 a_{qq} - 2sc a_{pq}, \\ a'_{qq} &= s^2 a_{pp} + c^2 a_{qq} + 2sc a_{pq}, \text{ and} \\ a'_{pq} = a'_{qp} &= (c^2 - s^2) a_{pq} + sc(a_{pp} - a_{qq}) \end{aligned}$$

In a Jacobi rotation,

$$a'_{pq} = 0 \Rightarrow \frac{c^2 - s^2}{2sc} = \frac{a_{qq} - a_{pp}}{2a_{pq}} = k \ \text{(say)}.$$

Left side is $\cot 2\phi$: solve this equation for $\phi$.

Jacobi rotation transformations $\mathbf{P}_{12}$, $\mathbf{P}_{13}$, $\cdots$, $\mathbf{P}_{1n}$; $\mathbf{P}_{23}$, $\cdots$, $\mathbf{P}_{2n}$; $\cdots$; $\mathbf{P}_{n-1,n}$ complete a full sweep.

**Note:** The resulting matrix is far from diagonal!

# Jacobi Rotation Method

Sum of squares of off-diagonal terms before the transformation

$$S = \sum_{r \neq s} |a_{rs}|^2 = 2\left[\sum_{r \neq p} a_{rp}^2 + \sum_{p \neq r \neq q} a_{rq}^2\right]$$
$$= 2\left[\sum_{p \neq r \neq q} (a_{rp}^2 + a_{rq}^2) + a_{pq}^2\right]$$

and that afterwards

$$S' = 2\left[\sum_{p \neq r \neq q} (a_{rp}'^2 + a_{rq}'^2) + a_{pq}'^2\right]$$
$$= 2\sum_{p \neq r \neq q} (a_{rp}^2 + a_{rq}^2)$$

differ by

$$\Delta S = S' - S = -2a_{pq}^2 \leq 0; \quad \text{and } S \to 0.$$



# Givens Rotation Method

While applying the rotation $\mathbf{P}_{pq}$, demand $a'_{rq} = 0$: $\tan\phi = -\frac{a_{rq}}{a_{rp}}$

$r = p - 1$: Givens rotation

- Once $a_{p-1,q}$ is annihilated, it is never updated again!

Sweep $\mathbf{P}_{23}, \mathbf{P}_{24}, \cdots, \mathbf{P}_{2n}$; $\mathbf{P}_{34}, \cdots, \mathbf{P}_{3n}$; $\cdots$; $\mathbf{P}_{n-1,n}$ to annihilate $a_{13}, a_{14}, \cdots, a_{1n}$; $a_{24}, \cdots, a_{2n}$; $\cdots$; $a_{n-2,n}$.

*Symmetric tridiagonal matrix*

How do eigenvectors transform through Jacobi/Givens rotation steps?

$$\tilde{\mathbf{A}} = \cdots \mathbf{P}^{(2)^T} \mathbf{P}^{(1)^T} \mathbf{A} \mathbf{P}^{(1)} \mathbf{P}^{(2)} \cdots$$

Product matrix $\mathbf{P}^{(1)}\mathbf{P}^{(2)} \cdots$ gives the basis.

To record it, initialize $\mathbf{V}$ by identity and keep multiplying new rotation matrices on the right side.



# Givens Rotation Method

Contrast between Jacobi and Givens rotation methods

- What happens to intermediate zeros?
- What do we get after a complete sweep?
- How many sweeps are to be applied?
- What is the *intended* final form of the matrix?
- How is size of the matrix relevant in the choice of the method?

Fast forward ...

- Housholder method accomplishes ‘tridiagonalization’ more efficiently than Givens rotation method.
- But, with a half-processed matrix, there come situations in which Givens rotation method turns out to be more efficient!



# Points to note

Rotation transformation on symmetric matrices

- Plane rotations provide orthogonal change of basis that can be used for diagonalization of matrices.
- For small matrices (say $4 \leq n \leq 8$), Jacobi rotation sweeps are competitive enough for diagonalization upto a reasonable tolerance.
- For large matrices, one sweep of Givens rotations can be applied to get a symmetric tridiagonal matrix, for efficient further processing.

Necessary Exercises: **2,3,4**

# Outline

Householder Transformation and Tridiagonal Matrices
- Householder Reflection Transformation
- Householder Method
- Eigenvalues of Symmetric Tridiagonal Matrices



# Householder Reflection Transformation

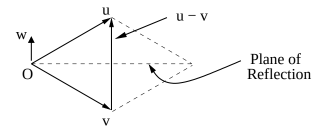



Figure: Vectors in Householder reflection

Consider $\mathbf{u}, \mathbf{v} \in R^k$, $\|\mathbf{u}\| = \|\mathbf{v}\|$ and $\mathbf{w} = \frac{\mathbf{u}-\mathbf{v}}{\|\mathbf{u}-\mathbf{v}\|}$.

**Householder reflection matrix**

$$\mathbf{H}_k = \mathbf{I}_k - 2\mathbf{w}\mathbf{w}^T$$

is symmetric and orthogonal.

For any vector $\mathbf{x}$ orthogonal to $\mathbf{w}$,

$$\mathbf{H}_k\mathbf{x} = (\mathbf{I}_k - 2\mathbf{w}\mathbf{w}^T)\mathbf{x} = \mathbf{x} \quad \text{and} \quad \mathbf{H}_k\mathbf{w} = (\mathbf{I}_k - 2\mathbf{w}\mathbf{w}^T)\mathbf{w} = -\mathbf{w}.$$

Hence, $\mathbf{H}_k\mathbf{y} = \mathbf{H}_k(\mathbf{y}_\mathbf{w} + \mathbf{y}_\perp) = -\mathbf{y}_\mathbf{w} + \mathbf{y}_\perp$, $\mathbf{H}_k\mathbf{u} = \mathbf{v}$ and $\mathbf{H}_k\mathbf{v} = \mathbf{u}$.



# Householder Method

Consider $n \times n$ symmetric matrix $\mathbf{A}$.
Let $\mathbf{u} = [a_{21} \;\; a_{31} \;\; \cdots \;\; a_{n1}]^T \in R^{n-1}$ and $\mathbf{v} = \|\mathbf{u}\|\mathbf{e}_1 \in R^{n-1}$.

Construct $\mathbf{P}_1 = \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{n-1} \end{bmatrix}$ and operate as

$$\begin{aligned} \mathbf{A}^{(1)} = \mathbf{P}_1\mathbf{A}\mathbf{P}_1 &= \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{n-1} \end{bmatrix} \begin{bmatrix} a_{11} & \mathbf{u}^T \\ \mathbf{u} & \mathbf{A}_1 \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} & \mathbf{v}^T \\ \mathbf{v} & \mathbf{H}_{n-1}\mathbf{A}_1\mathbf{H}_{n-1} \end{bmatrix}. \end{aligned}$$

Reorganizing and re-naming,

$$\mathbf{A}^{(1)} = \begin{bmatrix} d_1 & e_2 & \mathbf{0} \\ e_2 & d_2 & \mathbf{u}_2^T \\ \mathbf{0} & \mathbf{u}_2 & \mathbf{A}_2 \end{bmatrix}.$$



# Householder Method

Next, with $\mathbf{v}_2 = \|\mathbf{u}_2\|\mathbf{e}_1$, we form

$$\mathbf{P}_2 = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{n-2} \end{bmatrix}$$

and operate as $\mathbf{A}^{(2)} = \mathbf{P}_2\mathbf{A}^{(1)}\mathbf{P}_2$.
After $j$ steps,

$$\mathbf{A}^{(j)} = \begin{bmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & \ddots & & \\ & \ddots & \ddots & e_{j+1} & \\ & & e_{j+1} & d_{j+1} & \mathbf{u}_{j+1}^T \\ & & & \mathbf{u}_{j+1} & \mathbf{A}_{j+1} \end{bmatrix}$$

By $n-2$ steps, with $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\mathbf{P}_3 \cdots \mathbf{P}_{n-2}$,

$$\mathbf{A}^{(n-2)} = \mathbf{P}^T\mathbf{A}\mathbf{P}$$

is **symmetric tridiagonal**.

## Eigenvalues of Symmetric Tridiagonal Matrices

$$\mathbf{T} = \begin{bmatrix} d_1 & e_2 & & & \\ e_2 & d_2 & \ddots & & \\ & \ddots & \ddots & e_{n-1} & \\ & & e_{n-1} & d_{n-1} & e_n \\ & & & e_n & d_n \end{bmatrix}$$

Characteristic polynomial

$$p(\lambda) = \begin{vmatrix} \lambda - d_1 & -e_2 & & & \\ -e_2 & \lambda - d_2 & \ddots & & \\ & \ddots & \ddots & -e_{n-1} & \\ & & -e_{n-1} & \lambda - d_{n-1} & -e_n \\ & & & -e_n & \lambda - d_n \end{vmatrix}.$$



## Eigenvalues of Symmetric Tridiagonal Matrices

Characteristic polynomial of the leading $k \times k$ sub-matrix: $p_k(\lambda)$

$$\begin{aligned} p_0(\lambda) &= 1, \\ p_1(\lambda) &= \lambda - d_1, \\ p_2(\lambda) &= (\lambda - d_2)(\lambda - d_1) - e_2^2, \\ \cdots & \quad \cdots \quad \cdots, \\ p_{k+1}(\lambda) &= (\lambda - d_{k+1})p_k(\lambda) - e_{k+1}^2 p_{k-1}(\lambda). \end{aligned}$$

$P(\lambda) = \{p_0(\lambda), p_1(\lambda), \cdots, p_n(\lambda)\}$

- a Sturmian sequence if $e_j \neq 0 \ \forall j$

**Question:** What if $e_j = 0$ for some $j$?!
**Answer:** That is good news. Split the matrix.



## Eigenvalues of Symmetric Tridiagonal Matrices

**Sturmian sequence property** of $P(\lambda)$ with $e_j \neq 0$:

> **Interlacing property:** *Roots of $p_{k+1}(\lambda)$ interlace the roots of $p_k(\lambda)$. That is, if the roots of $p_{k+1}(\lambda)$ are $\lambda_1 > \lambda_2 > \cdots > \lambda_{k+1}$ and those of $p_k(\lambda)$ are $\mu_1 > \mu_2 > \cdots > \mu_k$; then*
>
> $$\lambda_1 > \mu_1 > \lambda_2 > \mu_2 > \cdots \quad \cdots > \lambda_k > \mu_k > \lambda_{k+1}.$$

This property leads to a convenient procedure.

**Proof**

$p_1(\lambda)$ has a single root, $d_1$.

$$p_2(d_1) = -e_2^2 < 0,$$

Since $p_2(\pm\infty) = \infty > 0$, roots $t_1$ and $t_2$ of $p_2(\lambda)$ are separated as $\infty > t_1 > d_1 > t_2 > -\infty$.

*The statement is true for $k = 1$.*



## Eigenvalues of Symmetric Tridiagonal Matrices

Next, we assume that the statement is true for $k = i$.
Roots of $p_i(\lambda)$: $\alpha_1 > \alpha_2 > \cdots > \alpha_i$
Roots of $p_{i+1}(\lambda)$: $\beta_1 > \beta_2 > \cdots > \beta_i > \beta_{i+1}$
Roots of $p_{i+2}(\lambda)$: $\gamma_1 > \gamma_2 > \cdots > \gamma_i > \gamma_{i+1} > \gamma_{i+2}$

**Assumption:** $\beta_1 > \alpha_1 > \beta_2 > \alpha_2 > \cdots \quad \cdots > \beta_i > \alpha_i > \beta_{i+1}$

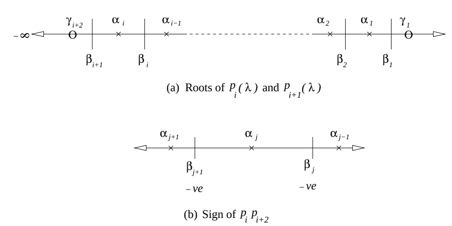



Figure: Interlacing of roots of characteristic polynomials

**To show:** $\gamma_1 > \beta_1 > \gamma_2 > \beta_2 > \cdots \quad \cdots > \gamma_{i+1} > \beta_{i+1} > \gamma_{i+2}$

## Eigenvalues of Symmetric Tridiagonal Matrices

Since $\beta_1 > \alpha_1$, $p_i(\beta_1)$ is of the same sign as $p_i(\infty)$, i.e. positive.

Therefore, $p_{i+2}(\beta_1) = -e_{i+2}^2 p_i(\beta_1)$ is negative.
But, $p_{i+2}(\infty)$ is clearly positive.

Hence, $\gamma_1 \in (\beta_1, \infty)$.
Similarly, $\gamma_{i+2} \in (-\infty, \beta_{i+1})$.

**Question:** Where are the rest of the $i$ roots of $p_{i+2}(\lambda)$?

$$\begin{aligned} p_{i+2}(\beta_j) &= (\beta_j - d_{i+2})p_{i+1}(\beta_j) - e_{i+2}^2 p_i(\beta_j) = -e_{i+2}^2 p_i(\beta_j) \\ p_{i+2}(\beta_{j+1}) &= -e_{i+2}^2 p_i(\beta_{j+1}) \end{aligned}$$

That is, $p_i$ and $p_{i+2}$ are of opposite signs at each $\beta$.

Refer figure.

Over $[\beta_{i+1}, \beta_1]$, $p_{i+2}(\lambda)$ changes sign over each sub-interval $[\beta_{j+1}, \beta_j]$, along with $p_i(\lambda)$, to maintain opposite signs at each $\beta$.

**Conclusion:** $p_{i+2}(\lambda)$ has *exactly one root* in $(\beta_{j+1}, \beta_j)$.



## Eigenvalues of Symmetric Tridiagonal Matrices

Examine sequence $P(w) = \{p_0(w), p_1(w), p_2(w), \cdots, p_n(w)\}$.
If $p_k(w)$ and $p_{k+1}(w)$ have opposite signs then $p_{k+1}(\lambda)$ has one root more than $p_k(\lambda)$ in the interval $(w, \infty)$.

> *Number of roots of $p_n(\lambda)$ above $w$ = number of sign changes in the sequence $P(w)$.*

**Consequence:** Number of roots of $p_n(\lambda)$ in $(a, b)$ = difference between numbers of sign changes in $P(a)$ and $P(b)$.

**Bisection method:** Examine the sequence at $\frac{a+b}{2}$.

> *Separate roots, bracket each of them and then squeeze the interval!*

Any way to start with an interval to include *all* eigenvalues?

$$|\lambda_i| \le \lambda_{bnd} = \max_{1 \le j \le n} \{|e_j| + |d_j| + |e_{j+1}|\}$$



## Eigenvalues of Symmetric Tridiagonal Matrices

**Algorithm**

- Identify the interval $[a, b]$ of interest.
- For a degenerate case (some $e_j = 0$), split the given matrix.
- For each of the non-degenerate matrices,
  - by repeated use of bisection and study of the sequence $P(\lambda)$, bracket individual eigenvalues within small sub-intervals, and
  - by further use of the bisection method (or a substitute) within each such sub-interval, determine the individual eigenvalues to the desired accuracy.

**Note:** The algorithm is based on Sturmian sequence property.



## Points to note

- A Householder matrix is symmetric and orthogonal. It effects a reflection transformation.
- A sequence of Householder transformations can be used to convert a symmetric matrix into a symmetric tridiagonal form.
- Eigenvalues of the leading square sub-matrices of a symmetric tridiagonal matrix exhibit a useful interlacing structure.
- This property can be used to separate and bracket eigenvalues.
- Method of bisection is useful in the separation as well as subsequent determination of the eigenvalues.

Necessary Exercises: **2,4,5**

# Outline







# QR Decomposition



Decomposition (or factorization) $\mathbf{A} = \mathbf{QR}$ into two factors, orthogonal $\mathbf{Q}$ and upper-triangular $\mathbf{R}$:

(a) It always exists.
(b) Performing this decomposition is pretty straightforward.
(c) It has a number of properties useful in the solution of the eigenvalue problem.

$$[\mathbf{a}_1 \quad \cdots \quad \mathbf{a}_n] = [\mathbf{q}_1 \quad \cdots \quad \mathbf{q}_n] \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ & \ddots & \vdots \\ & & r_{nn} \end{bmatrix}$$

A simple method based on Gram-Schmidt orthogonalization:
Considering columnwise equality $\mathbf{a}_j = \sum_{i=1}^{j} r_{ij}\mathbf{q}_i$,
for $j = 1, 2, 3, \cdots, n$;

$$r_{ij} = \mathbf{q}_i^T \mathbf{a}_j \quad \forall i < j, \quad \mathbf{a}'_j = \mathbf{a}_j - \sum_{i=1}^{j-1} r_{ij}\mathbf{q}_i, \quad r_{jj} = \|\mathbf{a}'_j\|;$$

$$\mathbf{q}_j = \begin{cases} \mathbf{a}'_j / r_{jj}, & \text{if } r_{jj} \neq 0; \\ \text{any vector satisfying } \mathbf{q}_i^T \mathbf{q}_j = \delta_{ij} \text{ for } 1 \leq i \leq j, & \text{if } r_{jj} = 0. \end{cases}$$



# QR Decomposition



**Practical method:** one-sided Householder transformations, starting with

$$\mathbf{u}_0 = \mathbf{a}_1, \quad \mathbf{v}_0 = \|\mathbf{u}_0\|\mathbf{e}_1 \in R^n \quad \text{and} \quad \mathbf{w}_0 = \frac{\mathbf{u}_0 - \mathbf{v}_0}{\|\mathbf{u}_0 - \mathbf{v}_0\|}$$

and $\mathbf{P}_0 = \mathbf{H}_n = \mathbf{I}_n - 2\mathbf{w}_0\mathbf{w}_0^T$.

$$\mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_2\mathbf{P}_1\mathbf{P}_0\mathbf{A} = \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_2\mathbf{P}_1 \begin{bmatrix} \|\mathbf{a}_1\| & ** \\ \mathbf{0} & \mathbf{A}_0 \end{bmatrix}$$

$$= \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_2 \begin{bmatrix} r_{11} & * & ** \\ & r_{22} & ** \\ & & \mathbf{A}_1 \end{bmatrix} = \cdots \quad \cdots = \mathbf{R}$$

With

$$\mathbf{Q} = (\mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_2\mathbf{P}_1\mathbf{P}_0)^T = \mathbf{P}_0\mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2},$$

we have $\mathbf{Q}^T\mathbf{A} = \mathbf{R} \Rightarrow \mathbf{A} = \mathbf{QR}$.



# QR Decomposition



**Alternative method useful for tridiagonal and Hessenberg matrices:** One-sided plane rotations

- rotations $\mathbf{P}_{12}$, $\mathbf{P}_{23}$ etc to annihilate $a_{21}$, $a_{32}$ etc in that sequence

Givens rotation matrices!

**Application in solution of a linear system:** $\mathbf{Q}$ and $\mathbf{R}$ factors of a matrix $\mathbf{A}$ come handy in the solution of $\mathbf{Ax} = \mathbf{b}$

$$\mathbf{QRx} = \mathbf{b} \Rightarrow \mathbf{Rx} = \mathbf{Q}^T\mathbf{b}$$

needs only a sequence of back-substitutions.

# QR Iterations



Multiplying **Q** and **R** factors in reverse,

$$\mathbf{A}' = \mathbf{RQ} = \mathbf{Q}^T\mathbf{AQ},$$

an orthogonal similarity transformation.

1. If **A** is symmetric, then so is $\mathbf{A}'$.
2. If **A** is in upper Hessenberg form, then so is $\mathbf{A}'$.
3. If **A** is symmetric tridiagonal, then so is $\mathbf{A}'$.

**Complexity of QR iteration:** $\mathcal{O}(n)$ for a symmetric tridiagonal matrix, $\mathcal{O}(n^2)$ operation for an upper Hessenberg matrix and $\mathcal{O}(n^3)$ for the general case.

**Algorithm:** Set $\mathbf{A}_1 = \mathbf{A}$ and for $k = 1, 2, 3, \cdots$,

- decompose $\mathbf{A}_k = \mathbf{Q}_k\mathbf{R}_k$,
- reassemble $\mathbf{A}_{k+1} = \mathbf{R}_k\mathbf{Q}_k$.

As $k \to \infty$, $\mathbf{A}_k$ approaches the *quasi-upper-triangular form*.



# QR Iterations



**Quasi-upper-triangular form:**

$$\begin{bmatrix} \lambda_1 & * & \cdots & * & \star\star & \cdots & * & * \\ & \lambda_2 & \cdots & * & \star\star & \cdots & * & * \\ & & \ddots & * & \star\star & \cdots & * & * \\ & & & \lambda_r & \star\star & \cdots & * & * \\ & & & & \mathbf{B}_k & \cdots & * & * \\ & & & & & \ddots & \vdots & \vdots \\ & & & & & & \begin{bmatrix} \alpha & -\omega \\ \omega & \beta \end{bmatrix} \end{bmatrix},$$

with $|\lambda_1| > |\lambda_2| > \cdots$.

- Diagonal blocks $\mathbf{B}_k$ correspond to eigenspaces of equal/close (magnitude) eigenvalues.
- $2 \times 2$ diagonal blocks often correspond to pairs of complex eigenvalues (for non-symmetric matrices).
- For symmetric matrices, the quasi-upper-triangular form reduces to quasi-diagonal form.



# Conceptual Basis of QR Method*



QR decomposition algorithm operates on the basis of the *relative magnitudes* of eigenvalues and segregates subspaces.

With $k \to \infty$,

$$\mathbf{A}^k Range\{\mathbf{e}_1\} = Range\{\mathbf{q}_1\} \to Range\{\mathbf{v}_1\}$$

and $(\mathbf{a}_1)_k \to \mathcal{Q}_k^T\mathbf{A}\mathbf{q}_1 = \lambda_1\mathcal{Q}_k^T\mathbf{q}_1 = \lambda_1\mathbf{e}_1$.

Further,

$$\mathbf{A}^k Range\{\mathbf{e}_1, \mathbf{e}_2\} = Range\{\mathbf{q}_1, \mathbf{q}_2\} \to Range\{\mathbf{v}_1, \mathbf{v}_2\}.$$

and $(\mathbf{a}_2)_k \to \mathcal{Q}_k^T\mathbf{A}\mathbf{q}_2 = \begin{bmatrix} (\lambda_1 - \lambda_2)\alpha_1 \\ \lambda_2 \\ \mathbf{0} \end{bmatrix}$.

And, so on ...



# QR Algorithm with Shift*



For $\lambda_i < \lambda_j$, entry $a_{ij}$ decays through iterations as $\left(\frac{\lambda_i}{\lambda_j}\right)^k$.

With shift,

$$\bar{\mathbf{A}}_k = \mathbf{A}_k - \mu_k\mathbf{I};$$
$$\bar{\mathbf{A}}_k = \mathbf{Q}_k\mathbf{R}_k, \quad \bar{\mathbf{A}}_{k+1} = \mathbf{R}_k\mathbf{Q}_k;$$
$$\mathbf{A}_{k+1} = \bar{\mathbf{A}}_{k+1} + \mu_k\mathbf{I}.$$

Resulting transformation is

$$\begin{aligned} \mathbf{A}_{k+1} &= \mathbf{R}_k\mathbf{Q}_k + \mu_k\mathbf{I} = \mathbf{Q}_k^T\bar{\mathbf{A}}_k\mathbf{Q}_k + \mu_k\mathbf{I} \\ &= \mathbf{Q}_k^T(\mathbf{A}_k - \mu_k\mathbf{I})\mathbf{Q}_k + \mu_k\mathbf{I} = \mathbf{Q}_k^T\mathbf{A}_k\mathbf{Q}_k. \end{aligned}$$

For the iteration,

$$\textit{convergence ratio} = \frac{\lambda_i - \mu_k}{\lambda_j - \mu_k}.$$

**Question:** How to find a suitable value for $\mu_k$?

## Points to note

- QR decomposition can be effected on any square matrix.
- Practical methods of QR decomposition use Householder transformations or Givens rotations.
- A QR iteration effects a similarity transformation on a matrix, preserving symmetry, Hessenberg structure and also a symmetric tridiagonal form.
- A sequence of QR iterations converge to an almost upper-triangular form.
- Operations on symmetric tridiagonal and Hessenberg forms are computationally efficient.
- QR iterations tend to order subspaces according to the relative magnitudes of eigenvalues.
- Eigenvalue shifting is useful as an expediting strategy.

Necessary Exercises: **1,3**



## Outline





## Introductory Remarks

- A general (non-symmetric) matrix may not be diagonalizable. We attempt to triangularize it.
- With real arithmetic, $2 \times 2$ diagonal blocks are inevitable — signifying complex pair of eigenvalues.
- Higher computational complexity, slow convergence and lack of numerical stability.

A non-symmetric matrix is usually unbalanced and is prone to higher round-off errors.

**Balancing** as a pre-processing step: multiplication of a row and division of the corresponding column with the same number, ensuring similarity.

*Note:* A balanced matrix may get unbalanced again through similarity transformations that are not orthogonal!



## Reduction to Hessenberg Form*

Methods to find appropriate similarity transformations

1. a full sweep of Givens rotations,
2. a sequence of $n-2$ steps of Householder transformations, and
3. a cycle of coordinated Gaussian elimination.

Method based on Gaussian elimination or elementary transformations:

> *The pre-multiplying matrix corresponding to the elementary row transformation and the post-multiplying matrix corresponding to the matching column transformation* **must be** *inverses of each other.*

Two kinds of steps

- Pivoting
- Elimination

## Reduction to Hessenberg Form*

**Pivoting step:** $\bar{\mathbf{A}} = \mathbf{P}_{rs}\mathbf{A}\mathbf{P}_{rs} = \mathbf{P}_{rs}^{-1}\mathbf{A}\mathbf{P}_{rs}$.

- Permutation $\mathbf{P}_{rs}$: interchange of $r$-th and $s$-th columns.
- $\mathbf{P}_{rs}^{-1} = \mathbf{P}_{rs}$: interchange of $r$-th and $s$-th rows.
- Pivot locations: $a_{21}, a_{32}, \cdots, a_{n-1,n-2}$.

**Elimination step:** $\bar{\mathbf{A}} = \mathbf{G}_r^{-1}\mathbf{A}\mathbf{G}_r$ with elimination matrix

$$\mathbf{G}_r = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & \mathbf{k} & \mathbf{I}_{n-r-1} \end{bmatrix} \quad \text{and} \quad \mathbf{G}_r^{-1} = \begin{bmatrix} \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 1 & \mathbf{0} \\ \mathbf{0} & -\mathbf{k} & \mathbf{I}_{n-r-1} \end{bmatrix}.$$

- $\mathbf{G}_r^{-1}$: Row $(r+1+i) \leftarrow$ Row $(r+1+i) - k_i \times$ Row $(r+1)$ for $i = 1, 2, 3, \cdots, n-r-1$
- $\mathbf{G}_r$: Column $(r+1) \leftarrow$ Column $(r+1) + \sum_{i=1}^{n-r-1} [k_i \times$ Column $(r+1+i)$ ]



## QR Algorithm on Hessenberg Matrices*

QR iterations: $\mathcal{O}(n^2)$ operations for upper Hessenberg form.

*Whenever a sub-diagonal zero appears, the matrix is split into two smaller upper Hessenberg blocks, and they are processed separately, thereby reducing the cost drastically.*

Particular cases:

- $a_{n,n-1} \to 0$: Accept $a_{nn} = \lambda_n$ as an eigenvalue, continue with the leading $(n-1) \times (n-1)$ sub-matrix.
- $a_{n-1,n-2} \to 0$: Separately find the eigenvalues $\lambda_{n-1}$ and $\lambda_n$ from $\begin{bmatrix} a_{n-1,n-1} & a_{n-1,n} \\ a_{n,n-1} & a_{n,n} \end{bmatrix}$, continue with the leading $(n-2) \times (n-2)$ sub-matrix.

**Shift strategy:** Double QR steps.



## Inverse Iteration

**Assumption:** Matrix $\mathbf{A}$ has a complete set of eigenvectors.

$(\lambda_i)_0$: a good estimate of an eigenvalue $\lambda_i$ of $\mathbf{A}$.

**Purpose:** To find $\lambda_i$ precisely and also to find $\mathbf{v}_i$.

**Step:** Select a random vector $\mathbf{y}_0$ (with $\|\mathbf{y}_0\| = 1$) and solve

$$[\mathbf{A} - (\lambda_i)_0\mathbf{I}]\mathbf{y} = \mathbf{y}_0.$$

**Result:** $\mathbf{y}$ is a good estimate of $\mathbf{v}_i$ and

$$(\lambda_i)_1 = (\lambda_i)_0 + \frac{1}{\mathbf{y}_0^T\mathbf{y}}$$

is an improvement in the estimate of the eigenvalue.

*How to establish the result and work out an* algorithm*?*



## Inverse Iteration

With $\mathbf{y}_0 = \sum_{j=1}^n \alpha_j\mathbf{v}_j$ and $\mathbf{y} = \sum_{j=1}^n \beta_j\mathbf{v}_j$, $[\mathbf{A} - (\lambda_i)_0\mathbf{I}]\mathbf{y} = \mathbf{y}_0$ gives

$$\sum_{j=1}^n \beta_j[\mathbf{A} - (\lambda_i)_0\mathbf{I}]\mathbf{v}_j = \sum_{j=1}^n \alpha_j\mathbf{v}_j$$
$$\Rightarrow \beta_j[\lambda_j - (\lambda_i)_0] = \alpha_j \Rightarrow \beta_j = \frac{\alpha_j}{\lambda_j - (\lambda_i)_0}.$$

$\beta_i$ is typically large and eigenvector $\mathbf{v}_i$ dominates $\mathbf{y}$.

$\mathbf{A}\mathbf{v}_i = \lambda_i\mathbf{v}_i$ gives $[\mathbf{A} - (\lambda_i)_0\mathbf{I}]\mathbf{v}_i = [\lambda_i - (\lambda_i)_0]\mathbf{v}_i$. Hence,

$$[\lambda_i - (\lambda_i)_0]\mathbf{y} \approx [\mathbf{A} - (\lambda_i)_0\mathbf{I}]\mathbf{y} = \mathbf{y}_0.$$

Inner product with $\mathbf{y}_0$ gives

$$[\lambda_i - (\lambda_i)_0]\mathbf{y}_0^T\mathbf{y} \approx 1 \Rightarrow \lambda_i \approx (\lambda_i)_0 + \frac{1}{\mathbf{y}_0^T\mathbf{y}}.$$

# Inverse Iteration

**Algorithm:**

Start with estimate $(\lambda_i)_0$, guess $\mathbf{y}_0$ (normalized).
For $k = 0, 1, 2, \cdots$

- Solve $[\mathbf{A} - (\lambda_i)_k \mathbf{I}]\mathbf{y} = \mathbf{y}_k$.
- Normalize $\mathbf{y}_{k+1} = \frac{\mathbf{y}}{\|\mathbf{y}\|}$.
- Improve $(\lambda_i)_{k+1} = (\lambda_i)_k + \frac{1}{\mathbf{y}_k^T \mathbf{y}}$.
- If $\|\mathbf{y}_{k+1} - \mathbf{y}_k\| < \epsilon$, terminate.

Important issues

- Update eigenvalue once in a while, not at every iteration.
- Use some acceptable small number as artificial pivot.
- The method may not converge for defective matrix or for one having complex eigenvalues.
- Repeated eigenvalues may inhibit the process.



# Recommendation

Table: Eigenvalue problem: summary of methods

| Type | Size | Reduction | Algorithm | Post-processing |
|---|---|---|---|---|
| General | Small (up to 4) | Definition: Characteristic polynomial | Polynomial root finding (eigenvalues) | Solution of linear systems (eigenvectors) |
| Symmetric | Intermediate (say, 4–12) | Jacobi sweeps | Selective Jacobi rotations | |
| | | Tridiagonalization (Givens rotation or Householder method) | Sturm sequence property: Bracketing and bisection (rough eigenvalues) | Inverse iteration (eigenvalue improvement and eigenvectors) |
| | Large | Tridiagonalization (usually Householder method) | QR decomposition iterations | |
| Non-symmetric | Intermediate Large | Balancing, and then Reduction to Hessenberg form (Above methods or Gaussian elimination) | QR decomposition iterations (eigenvalues) | Inverse iteration (eigenvectors) |
| General | Very large (selective requirement) | | Power method, shift and deflation | |



# Points to note

- Eigenvalue problem of a non-symmetric matrix is difficult!
- Balancing and reduction to Hessenberg form are desirable pre-processing steps.
- QR decomposition algorithm is typically used for reduction to an upper-triangular form.
- Use inverse iteration to polish eigenvalue and find eigenvectors.
- In algebraic eigenvalue problems, different methods or combinations are suitable for different cases; regarding matrix size, symmetry and the requirements.

Necessary Exercises: **1,2**



# Outline

## SVD Theorem and Construction



Eigenvalue problem: $\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^{-1}$ where $\mathbf{U} = \mathbf{V}$

Do not ask for similarity. Focus on the *form* of the decomposition.

**Guaranteed** decomposition with **orthogonal** $\mathbf{U}$, $\mathbf{V}$, and **non-negative** diagonal entries in $\Lambda$.

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \quad \text{such that} \quad \mathbf{U}^T\mathbf{A}\mathbf{V} = \Sigma$$

**SVD Theorem** *For any real matrix* $\mathbf{A} \in R^{m\times n}$, *there exist orthogonal matrices* $\mathbf{U} \in R^{m\times m}$ *and* $\mathbf{V} \in R^{n\times n}$ *such that*

$$\mathbf{U}^T\mathbf{A}\mathbf{V} = \Sigma \in R^{m\times n}$$

*is a diagonal matrix, with diagonal entries* $\sigma_1, \sigma_2, \cdots \geq 0$, *obtained by appending the square diagonal matrix* $\mathrm{diag}\,(\sigma_1, \sigma_2, \cdots, \sigma_p)$ *with* $(m-p)$ *zero rows or* $(n-p)$ *zero columns, where* $p = \min(m, n)$.

Singular values: $\sigma_1, \sigma_2, \cdots, \sigma_p$.

Similar result for complex matrices



## SVD Theorem and Construction



**Question:** How to construct $\mathbf{U}$, $\mathbf{V}$ and $\Sigma$?

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

$$\mathbf{A}^T\mathbf{A} = (\mathbf{V}\Sigma^T\mathbf{U}^T)(\mathbf{U}\Sigma\mathbf{V}^T) = \mathbf{V}\Sigma^T\Sigma\mathbf{V}^T = \mathbf{V}\Lambda\mathbf{V}^T,$$

where $\Lambda = \Sigma^T\Sigma$ is an $n \times n$ diagonal matrix.

$$\Sigma = \left[\begin{array}{cccc|c} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \mathbf{0} \\ & & & \sigma_p & \\ \hline & & \mathbf{0} & & \times \end{array}\right]$$

Determine $\mathbf{V}$ and $\Lambda$. Work out $\Sigma$ and we have

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T \;\Rightarrow\; \mathbf{A}\mathbf{V} = \mathbf{U}\Sigma$$

This provides a proof as well!



## SVD Theorem and Construction



From $\mathbf{A}\mathbf{V} = \mathbf{U}\Sigma$, determine columns of $\mathbf{U}$.

1. Column $\mathbf{A}\mathbf{v}_k = \sigma_k\mathbf{u}_k$, with $\sigma_k \neq 0$: determine column $\mathbf{u}_k$.

   *Columns developed are* bound *to be mutually orthonormal!*

   Verify $\mathbf{u}_i^T\mathbf{u}_j = \left(\frac{1}{\sigma_i}\mathbf{A}\mathbf{v}_i\right)^T \left(\frac{1}{\sigma_j}\mathbf{A}\mathbf{v}_j\right) = \delta_{ij}$.
2. Column $\mathbf{A}\mathbf{v}_k = \sigma_k\mathbf{u}_k$, with $\sigma_k = 0$: $\mathbf{u}_k$ is left indeterminate (free).
3. In the case of $m < n$, identically zero columns $\mathbf{A}\mathbf{v}_k = \mathbf{0}$ for $k > m$: no corresponding columns of $\mathbf{U}$ to determine.
4. In the case of $m > n$, there will be $(m-n)$ columns of $\mathbf{U}$ left indeterminate.

Extend columns of $\mathbf{U}$ to an orthonormal basis.

All three factors in the decomposition are constructed, as desired.



## Properties of SVD



For a given matrix, the SVD is unique up to

(a) the same permutations of columns of $\mathbf{U}$, columns of $\mathbf{V}$ and diagonal elements of $\Sigma$;

(b) the same orthonormal linear combinations among columns of $\mathbf{U}$ and columns of $\mathbf{V}$, corresponding to equal singular values; and

(c) arbitrary orthonormal linear combinations among columns of $\mathbf{U}$ or columns of $\mathbf{V}$, corresponding to zero or non-existent singular values.

Ordering of the singular values:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0, \quad \text{and} \quad \sigma_{r+1} = \sigma_{r+2} = \cdots = \sigma_p = 0.$$

$Rank(\mathbf{A}) = Rank(\Sigma) = r$

*Rank of a matrix is the same as the number of its non-zero singular values.*

## Properties of SVD



$$\mathbf{Ax} = \mathbf{U\Sigma V}^T\mathbf{x} = \mathbf{U\Sigma y} = [\mathbf{u}_1 \quad \cdots \quad \mathbf{u}_r \quad \mathbf{u}_{r+1} \quad \cdots \quad \mathbf{u}_m]\begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_r y_r \\ \mathbf{0} \end{bmatrix}$$

$$= \sigma_1 y_1 \mathbf{u}_1 + \sigma_2 y_2 \mathbf{u}_2 + \cdots + \sigma_r y_r \mathbf{u}_r$$

has non-zero components along only the first $r$ columns of $\mathbf{U}$.

$\mathbf{U}$ *gives an orthonormal basis for the co-domain such that*

$$Range(\mathbf{A}) = < \mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_r > .$$

With $\mathbf{V}^T\mathbf{x} = \mathbf{y}$, $\mathbf{v}_k^T\mathbf{x} = y_k$, and

$$\mathbf{x} = y_1\mathbf{v}_1 + y_2\mathbf{v}_2 + \cdots + y_r\mathbf{v}_r + y_{r+1}\mathbf{v}_{r+1} + \cdots y_n\mathbf{v}_n.$$

$\mathbf{V}$ *gives an orthonormal basis for the domain such that*

$$Null(\mathbf{A}) = < \mathbf{v}_{r+1}, \mathbf{v}_{r+2}, \cdots, \mathbf{v}_n > .$$



## Properties of SVD



In basis $\mathbf{V}$, $\mathbf{v} = c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + \cdots + c_n\mathbf{v}_n = \mathbf{Vc}$ and the norm is given by

$$\|\mathbf{A}\|^2 = \max_{\mathbf{v}} \frac{\|\mathbf{Av}\|^2}{\|\mathbf{v}\|^2} = \max_{\mathbf{v}} \frac{\mathbf{v}^T\mathbf{A}^T\mathbf{Av}}{\mathbf{v}^T\mathbf{v}}$$
$$= \max_{\mathbf{c}} \frac{\mathbf{c}^T\mathbf{V}^T\mathbf{A}^T\mathbf{AVc}}{\mathbf{c}^T\mathbf{V}^T\mathbf{Vc}} = \max_{\mathbf{c}} \frac{\mathbf{c}^T\Sigma^T\Sigma\mathbf{c}}{\mathbf{c}^T\mathbf{c}} = \max_{\mathbf{c}} \frac{\sum_k \sigma_k^2 c_k^2}{\sum_k c_k^2}.$$

$$\|\mathbf{A}\| = \sqrt{\max_{\mathbf{c}} \frac{\sum_k \sigma_k^2 c_k^2}{\sum_k c_k^2}} = \sigma_{\max}$$

For a non-singular square matrix,

$$\mathbf{A}^{-1} = (\mathbf{U\Sigma V}^T)^{-1} = \mathbf{V}\Sigma^{-1}\mathbf{U}^T = \mathbf{V}\ \mathrm{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \cdots, \frac{1}{\sigma_n}\right)\mathbf{U}^T.$$

Then, $\|\mathbf{A}^{-1}\| = \frac{1}{\sigma_{\min}}$ and the condition number is

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\ \|\mathbf{A}^{-1}\| = \frac{\sigma_{\max}}{\sigma_{\min}}.$$



## Properties of SVD



Revision of definition of **norm** and **condition number**:

*The norm of a matrix is the same as its largest singular value, while its condition number is given by the ratio of the largest singular value to the least.*

Arranging singular values in decreasing order, with $Rank(\mathbf{A}) = r$,

$$\mathbf{U} = [\mathbf{U}_r \quad \bar{\mathbf{U}}] \quad \text{and} \quad \mathbf{V} = [\mathbf{V}_r \quad \bar{\mathbf{V}}],$$

$$\mathbf{A} = \mathbf{U\Sigma V}^T = [\mathbf{U}_r \quad \bar{\mathbf{U}}]\begin{bmatrix} \Sigma_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\begin{bmatrix} \mathbf{V}_r^T \\ \bar{\mathbf{V}}^T \end{bmatrix},$$

or,

$$\mathbf{A} = \mathbf{U}_r\Sigma_r\mathbf{V}_r^T = \sum_{k=1}^{r} \sigma_k \mathbf{u}_k \mathbf{v}_k^T.$$

**Efficient storage and reconstruction!**



## Pseudoinverse and Solution of Linear Systems



**Generalized inverse:** $\mathbf{G}$ is called a generalized inverse or g-inverse of $\mathbf{A}$ if, for $\mathbf{b} \in Range(\mathbf{A})$, $\mathbf{Gb}$ is a solution of $\mathbf{Ax} = \mathbf{b}$.

The Moore-Penrose inverse or the pseudoinverse:

$$\mathbf{A}^{\#} = (\mathbf{U\Sigma V}^T)^{\#} = (\mathbf{V}^T)^{\#}\Sigma^{\#}\mathbf{U}^{\#} = \mathbf{V}\Sigma^{\#}\mathbf{U}^T$$

With $\Sigma = \begin{bmatrix} \Sigma_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$, $\Sigma^{\#} = \begin{bmatrix} \Sigma_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$.

$$\text{Or,} \quad \Sigma^{\#} = \left[\begin{array}{cccc|c} \rho_1 & & & & \\ & \rho_2 & & & \\ & & \ddots & & \mathbf{0} \\ & & & \rho_p & \\ \hline & & \mathbf{0} & & \times \end{array}\right],$$

where $\rho_k = \begin{cases} \frac{1}{\sigma_k}, & \text{for } \sigma_k \neq 0 \text{ or for } |\sigma_k| > \epsilon; \\ 0, & \text{for } \sigma_k = 0 \text{ or for } |\sigma_k| \leq \epsilon. \end{cases}$

## Pseudoinverse and Solution of Linear Systems



Inverse-like facets and beyond

- $(\mathbf{A}^{\#})^{\#} = \mathbf{A}$.
- If $\mathbf{A}$ is invertible, then $\mathbf{A}^{\#} = \mathbf{A}^{-1}$.
  - $\mathbf{A}^{\#}\mathbf{b}$ gives the correct unique solution.
- If $\mathbf{Ax} = \mathbf{b}$ is an under-determined consistent system, then $\mathbf{A}^{\#}\mathbf{b}$ selects the solution $\mathbf{x}^*$ with the minimum norm.
- If the system is inconsistent, then $\mathbf{A}^{\#}\mathbf{b}$ minimizes the least square error $\|\mathbf{Ax} - \mathbf{b}\|$.
  - If the minimizer of $\|\mathbf{Ax} - \mathbf{b}\|$ is not unique, then it picks up that minimizer which has the minimum norm $\|\mathbf{x}\|$ among such minimizers.

Contrast with Tikhonov regularization:

*Pseudoinverse solution for precision and diagnosis.*
*Tikhonov's solution for continuity of solution over variable $\mathbf{A}$ and computational efficiency.*



## Optimality of Pseudoinverse Solution



Pseudoinverse solution of $\mathbf{Ax} = \mathbf{b}$:

$$\mathbf{x}^* = \mathbf{V}\Sigma^{\#}\mathbf{U}^T\mathbf{b} = \sum_{k=1}^{r} \rho_k \mathbf{v}_k \mathbf{u}_k^T \mathbf{b} = \sum_{k=1}^{r} (\mathbf{u}_k^T\mathbf{b}/\sigma_k)\mathbf{v}_k$$

Minimize

$$E(\mathbf{x}) = \frac{1}{2}(\mathbf{Ax}-\mathbf{b})^T(\mathbf{Ax}-\mathbf{b}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{Ax} - \mathbf{x}^T\mathbf{A}^T\mathbf{b} + \frac{1}{2}\mathbf{b}^T\mathbf{b}$$

Condition of vanishing gradient:

$$\begin{aligned}
\frac{\partial E}{\partial \mathbf{x}} = \mathbf{0} \quad &\Rightarrow \quad \mathbf{A}^T\mathbf{Ax} = \mathbf{A}^T\mathbf{b} \\
&\Rightarrow \quad \mathbf{V}(\Sigma^T\Sigma)\mathbf{V}^T\mathbf{x} = \mathbf{V}\Sigma^T\mathbf{U}^T\mathbf{b} \\
&\Rightarrow \quad (\Sigma^T\Sigma)\mathbf{V}^T\mathbf{x} = \Sigma^T\mathbf{U}^T\mathbf{b} \\
&\Rightarrow \quad \sigma_k^2\mathbf{v}_k^T\mathbf{x} = \sigma_k\mathbf{u}_k^T\mathbf{b} \\
&\Rightarrow \quad \mathbf{v}_k^T\mathbf{x} = \mathbf{u}_k^T\mathbf{b}/\sigma_k \quad \text{for } k = 1, 2, 3, \cdots, r.
\end{aligned}$$



## Optimality of Pseudoinverse Solution



With $\bar{\mathbf{V}} = [\mathbf{v}_{r+1} \quad \mathbf{v}_{r+2} \quad \cdots \quad \mathbf{v}_n]$, then

$$\mathbf{x} = \sum_{k=1}^{r} (\mathbf{u}_k^T\mathbf{b}/\sigma_k)\mathbf{v}_k + \bar{\mathbf{V}}\mathbf{y} = \mathbf{x}^* + \bar{\mathbf{V}}\mathbf{y}.$$

How to minimize $\|\mathbf{x}\|^2$ subject to $E(\mathbf{x})$ minimum?

*Minimize* $E_1(\mathbf{y}) = \|\mathbf{x}^* + \bar{\mathbf{V}}\mathbf{y}\|^2$.

Since $\mathbf{x}^*$ and $\bar{\mathbf{V}}\mathbf{y}$ are mutually orthogonal,

$$E_1(\mathbf{y}) = \|\mathbf{x}^* + \bar{\mathbf{V}}\mathbf{y}\|^2 = \|\mathbf{x}^*\|^2 + \|\bar{\mathbf{V}}\mathbf{y}\|^2$$

is minimum when $\bar{\mathbf{V}}\mathbf{y} = 0$, i.e. $\mathbf{y} = 0$.



## Optimality of Pseudoinverse Solution



**Anatomy of the optimization through SVD**

Using basis $\mathbf{V}$ for domain and $\mathbf{U}$ for co-domain, the variables are transformed as

$$\mathbf{V}^T\mathbf{x} = \mathbf{y} \quad \text{and} \quad \mathbf{U}^T\mathbf{b} = \mathbf{c}.$$

Then,

$$\mathbf{Ax} = \mathbf{b} \;\Rightarrow\; \mathbf{U}\Sigma\mathbf{V}^T\mathbf{x} = \mathbf{b} \;\Rightarrow\; \Sigma\mathbf{V}^T\mathbf{x} = \mathbf{U}^T\mathbf{b} \;\Rightarrow\; \Sigma\mathbf{y} = \mathbf{c}.$$

A completely decoupled system!
Usable components: $y_k = c_k/\sigma_k$ for $k = 1, 2, 3, \cdots, r$.
For $k > r$,

- completely redundant information ($c_k = 0$)
- purely unresolvable conflict ($c_k \neq 0$)

*SVD extracts this pure redundancy/inconsistency.*
*Setting $\rho_k = 0$ for $k > r$ rejects it wholesale!*
*At the same time, $\|\mathbf{y}\|$ is minimized, and hence $\|\mathbf{x}\|$ too.*

## Points to note

- SVD provides a complete orthogonal decomposition of the domain and co-domain of a linear transformation, separating out functionally distinct subspaces.
- If offers a complete diagnosis of the pathologies of systems of linear equations.
- Pseudoinverse solution of linear systems satisfy meaningful optimality requirements in several contexts.
- With the existence of SVD guaranteed, many important results can be established in a straightforward manner.

Necessary Exercises: **2,4,5,6,7**



## Outline





## Group

A set $G$ and a binary operation, say '+', fulfilling

Closure: $a + b \in G \; \forall a, b \in G$

Associativity: $a + (b + c) = (a + b) + c, \; \forall a, b, c \in G$

Existence of identity: $\exists 0 \in G$ such that $\forall a \in G, a + 0 = a = 0 + a$

Existence of inverse: $\forall a \in G, \; \exists (-a) \in G$ such that
$$a + (-a) = 0 = (-a) + a$$

Examples: $(Z, +)$, $(Z, +)$, $(Q - \{0\}, \cdot)$, $2 \times 5$ real matrices, Rotations etc.

- Commutative group
- Subgroup



## Field

A set $F$ and two binary operations, say '+' and '·', satisfying

Group property for addition: $(F, +)$ is a commutative group. (Denote the identity element of this group as '0'.)

Group property for multiplication: $(F - \{0\}, \cdot)$ is a commutative group. (Denote the identity element of this group as '1'.)

Distributivity: $a \cdot (b + c) = a \cdot b + a \cdot c, \quad \forall a, b, c \in F.$

Concept of field: abstraction of a number system

Examples: $(Q, +, \cdot)$, $(R, +, \cdot)$, $(C, +, \cdot)$ etc.

- Subfield

## Vector Space

A vector space is defined by

- a field $F$ of 'scalars',
- a commutative group $\mathbf{V}$ of 'vectors', and
- a binary operation between $F$ and $\mathbf{V}$, that may be called 'scalar multiplication', such that $\forall \alpha, \beta \in F$, $\forall \mathbf{a}, \mathbf{b} \in \mathbf{V}$; the following conditions hold.

  Closure: $\alpha \mathbf{a} \in \mathbf{V}$.
  Identity: $1\mathbf{a} = \mathbf{a}$.
  Associativity: $(\alpha\beta)\mathbf{a} = \alpha(\beta\mathbf{a})$.
  Scalar distributivity: $\alpha(\mathbf{a}+\mathbf{b}) = \alpha\mathbf{a} + \alpha\mathbf{b}$.
  Vector distributivity: $(\alpha+\beta)\mathbf{a} = \alpha\mathbf{a} + \beta\mathbf{a}$.

Examples: $R^n$, $C^n$, $m \times n$ real matrices etc.

Field ↔ Number system
Vector space ↔ Space



## Vector Space

Suppose $\mathbf{V}$ is a vector space.
Take a vector $\xi_1 \neq \mathbf{0}$ in it.

*Then, vectors* linearly dependent *on* $\xi_1$*:*
$\alpha_1\xi_1 \in \mathbf{V}\ \forall \alpha_1 \in F$.

**Question:** Are the elements of $\mathbf{V}$ exhausted?

If not, then take $\xi_2 \in \mathbf{V}$: *linearly independent* from $\xi_1$.

*Then,* $\alpha_1\xi_1 + \alpha_2\xi_2 \in \mathbf{V}\ \forall \alpha_1, \alpha_2 \in F$.

**Question:** Are the elements of $\mathbf{V}$ exhausted *now*?

... ... ...

**Question:** Will this process ever end?

Suppose it does.

*finite dimensional* vector space



## Vector Space

**Finite dimensional vector space**

Suppose the above process ends after $n$ choices of *linearly independent* vectors.

$$\chi = \alpha_1\xi_1 + \alpha_2\xi_2 + \cdots + \alpha_n\xi_n$$

Then,

- $n$: *dimension* of the vector space
- ordered set $\xi_1, \xi_2, \cdots, \xi_n$: a basis
- $\alpha_1, \alpha_2, \cdots, \alpha_n \in F$: *coordinates* of $\chi$ in that basis

$R^n$, $R^m$ etc: vector spaces over the field of real numbers

- Subspace



## Linear Transformation

A mapping $\mathbf{T} : \mathbf{V} \to \mathbf{W}$ satisfying

$$\mathbf{T}(\alpha\mathbf{a} + \beta\mathbf{b}) = \alpha\mathbf{T}(\mathbf{a}) + \beta\mathbf{T}(\mathbf{b})\ \ \forall \alpha, \beta \in F \ \text{ and } \ \forall \mathbf{a}, \mathbf{b} \in \mathbf{V}$$

where $\mathbf{V}$ and $\mathbf{W}$ are vector spaces over the field $F$.

**Question:** How to describe the linear transformation $\mathbf{T}$?

- For $\mathbf{V}$, basis $\xi_1, \xi_2, \cdots, \xi_n$
- For $\mathbf{W}$, basis $\eta_1, \eta_2, \cdots, \eta_m$

$\xi_1 \in \mathbf{V}$ gets mapped to $\mathbf{T}(\xi_1) \in \mathbf{W}$.

$$\mathbf{T}(\xi_1) = a_{11}\eta_1 + a_{21}\eta_2 + \cdots + a_{m1}\eta_m$$

Similarly, enumerate $\mathbf{T}(\xi_j) = \sum_{i=1}^m a_{ij}\eta_i$.

Matrix $\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_n]$ codes this description!

## Linear Transformation

A general element $\chi$ of $\mathbf{V}$ can be expressed as

$$\chi = x_1\xi_1 + x_2\xi_2 + \cdots + x_n\xi_n$$

Coordinates in a column: $\mathbf{x} = [x_1 \;\; x_2 \;\; \cdots \;\; x_n]^T$

Mapping:

$$\mathbf{T}(\chi) = x_1\mathbf{T}(\xi_1) + x_2\mathbf{T}(\xi_2) + \cdots + x_n\mathbf{T}(\xi_n),$$

with coordinates $\mathbf{Ax}$, as we know!

Summary:

- basis vectors of $\mathbf{V}$ get mapped to vectors in $\mathbf{W}$ whose coordinates are listed in columns of $\mathbf{A}$, and
- a vector of $\mathbf{V}$, having its coordinates in $\mathbf{x}$, gets mapped to a vector in $\mathbf{W}$ whose coordinates are obtained from $\mathbf{Ax}$.



## Linear Transformation

**Understanding:**

- Vector $\chi$ is an actual object in the set $\mathbf{V}$ and the column $\mathbf{x} \in R^n$ is merely a list of its coordinates.
- $\mathbf{T} : \mathbf{V} \to \mathbf{W}$ is the linear transformation and the matrix $\mathbf{A}$ simply stores coefficients needed to describe it.
- By changing bases of $\mathbf{V}$ and $\mathbf{W}$, the same vector $\chi$ and the same linear transformation are now expressed by different $\mathbf{x}$ and $\mathbf{A}$, respectively.

*Matrix representation emerges as the natural description of a linear transformation between two vector spaces.*

**Exercise:** Set of all $\mathbf{T} : \mathbf{V} \to \mathbf{W}$ form a vector space of their own!! Analyze and describe *that* vector space.



## Isomorphism

Consider $\mathbf{T} : \mathbf{V} \to \mathbf{W}$ that establishes a *one-to-one* correspondence.

- Linear transformation $\mathbf{T}$ defines a one-one onto mapping, which is *invertible*.
- $\dim \mathbf{V} = \dim \mathbf{W}$
- Inverse linear transformation $\mathbf{T}^{-1} : \mathbf{W} \to \mathbf{V}$
- $\mathbf{T}$ defines (is) an *isomorphism*.
- Vector spaces $\mathbf{V}$ and $\mathbf{W}$ are *isomorphic* to each other.
- Isomorphism is an equivalence relation. $\mathbf{V}$ and $\mathbf{W}$ are *equivalent!*

If we need to perform some operations on vectors in one vector space, we may as well

1. transform the vectors to another vector space through an isomorphism,
2. conduct the required operations there, and
3. map the results back to the original space through the inverse.



## Isomorphism

Consider vector spaces $\mathbf{V}$ and $\mathbf{W}$ over the same field $F$ and of the same dimension $n$.

**Question:** Can we define an isomorphism between them?

**Answer:** Of course. As many as we want!

*The underlying field and the dimension together completely specify a vector space,* up to an isomorphism.

- All $n$-dimensional vector spaces over the field $F$ are isomorphic to one another.
- In particular, they are all isomorphic to $F^n$.
- The representation (columns) can be considered as the objects (vectors) themselves.

## Inner Product Space

**Inner product** $(\mathbf{a}, \mathbf{b})$ in a *real* or *complex* vector space: a scalar function $p : \mathbf{V} \times \mathbf{V} \rightarrow F$ satisfying

Closure: $\forall\ \mathbf{a}, \mathbf{b} \in \mathbf{V},\ (\mathbf{a}, \mathbf{b}) \in F$

Associativity: $(\alpha\mathbf{a}, \mathbf{b}) = \alpha(\mathbf{a}, \mathbf{b})$

Distributivity: $(\mathbf{a} + \mathbf{b}, \mathbf{c}) = (\mathbf{a}, \mathbf{c}) + (\mathbf{b}, \mathbf{c})$

Conjugate commutativity: $(\mathbf{b}, \mathbf{a}) = \overline{(\mathbf{a}, \mathbf{b})}$

Positive definiteness: $(\mathbf{a}, \mathbf{a}) \geq 0$; and $(\mathbf{a}, \mathbf{a}) = 0$ iff $\mathbf{a} = \mathbf{0}$

*Note:* Property of conjugate commutativity forces $(\mathbf{a}, \mathbf{a})$ to be real.

Examples: $\mathbf{a}^T\mathbf{b}$, $\mathbf{a}^T\mathbf{W}\mathbf{b}$ in $R$, $\mathbf{a}^*\mathbf{b}$ in $C$ etc.

**Inner product space:** a vector space possessing an inner product

- Euclidean space: over $R$
- Unitary space: over $C$



## Inner Product Space

*Inner products bring in ideas of angle and length in the geometry of vector spaces.*

**Orthogonality:** $(\mathbf{a}, \mathbf{b}) = 0$

**Norm:** $\|\cdot\| : \mathbf{V} \rightarrow R$, such that $\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})}$

Associativity: $\|\alpha\mathbf{a}\| = |\alpha|\ \|\mathbf{a}\|$

Positive definiteness: $\|\mathbf{a}\| > 0$ for $\mathbf{a} \neq 0$ and $\|\mathbf{0}\| = 0$

Triangle inequality: $\|\mathbf{a} + \mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$

Cauchy-Schwarz inequality: $(\mathbf{a}, \mathbf{b}) \leq \|\mathbf{a}\|\ \|\mathbf{b}\|$

A distance function or *metric*: $d_\mathbf{V} : \mathbf{V} \times \mathbf{V} \rightarrow R$ such that

$$d_\mathbf{V}(\mathbf{a}, \mathbf{b}) = \|\mathbf{a} - \mathbf{b}\|$$



## Function Space

Suppose we decide to represent a continuous function $f : [a, b] \rightarrow R$ by the listing

$$\mathbf{v}_f = [f(x_1) \quad f(x_2) \quad f(x_3) \quad \cdots \quad f(x_N)]^T$$

with $a = x_1 < x_2 < x_3 < \cdots < x_N = b$.

*Note:* The 'true' representation will require $N$ to be infinite!

Here, $\mathbf{v}_f$ is a real column vector.
Do such vectors form a **vector space**?

*Correspondingly, does the set $\mathcal{F}$ of continuous functions over $[a, b]$ form a vector space?*

*infinite dimensional* vector space



## Function Space

**Vector space of continuous functions**

First, $(\mathcal{F}, +)$ is a commutative group.

Next, with $\alpha, \beta \in R$, $\forall x \in [a, b]$,

- if $f(x) \in R$, then $\alpha f(x) \in R$
- $1 \cdot f(x) = f(x)$
- $(\alpha\beta)f(x) = \alpha[\beta f(x)]$
- $\alpha[f_1(x) + f_2(x)] = \alpha f_1(x) + \alpha f_2(x)$
- $(\alpha + \beta)f(x) = \alpha f(x) + \beta f(x)$

- Thus, $\mathcal{F}$ forms a vector space over $R$.
- Every function in this space is an (infinite dimensional) vector.
- Listing of values is just an obvious basis.

## Function Space

**Linear dependence** of (non-zero) functions $f_1$ and $f_2$

- $f_2(x) = kf_1(x)$ for all $x$ in the domain
- $k_1 f_1(x) + k_2 f_2(x) = 0$, $\forall x$ with $k_1$ and $k_2$ not both zero.

**Linear independence**: $k_1 f_1(x) + k_2 f_2(x) = 0 \ \forall x \ \Rightarrow \ k_1 = k_2 = 0$

In general,

- Functions $f_1, f_2, f_3, \cdots, f_n \in \mathcal{F}$ are linearly dependent if $\exists \ k_1, k_2, k_3, \cdots, k_n$, not all zero, such that $k_1 f_1(x) + k_2 f_2(x) + k_3 f_3(x) + \cdots + k_n f_n(x) = 0 \ \forall x \in [a, b]$.
- $k_1 f_1(x) + k_2 f_2(x) + k_3 f_3(x) + \cdots + k_n f_n(x) = 0 \ \forall x \in [a, b] \ \Rightarrow$ $k_1, k_2, k_3, \cdots, k_n = 0$ means that functions $f_1, f_2, f_3, \cdots, f_n$ are linearly independent.

**Example:** functions $1, x, x^2, x^3, \cdots$ are a set of linearly independent functions.

Incidentally, this set is a commonly used **basis**.



## Function Space

**Inner product:** For functions $f(x)$ and $g(x)$ in $\mathcal{F}$, the usual inner product between corresponding vectors:

$$(\mathbf{v}_f, \mathbf{v}_g) = \mathbf{v}_f^T \mathbf{v}_g = f(x_1)g(x_1) + f(x_2)g(x_2) + f(x_3)g(x_3) + \cdots$$

Weighted inner product: $(\mathbf{v}_f, \mathbf{v}_g) = \mathbf{v}_f^T \mathbf{W} \mathbf{v}_g = \sum_i w_i f(x_i) g(x_i)$

For the functions,

$$(f, g) = \int_a^b w(x) f(x) g(x) dx$$

- **Orthogonality:** $(f, g) = \int_a^b w(x) f(x) g(x) dx = 0$
- **Norm:** $\|f\| = \sqrt{\int_a^b w(x)[f(x)]^2 dx}$
- **Orthonormal basis:** $(f_j, f_k) = \int_a^b w(x) f_j(x) f_k(x) dx = \delta_{jk} \ \forall j, k$



## Points to note

- Matrix algebra provides a *natural* description for vector spaces and linear transformations.
- Through isomorphisms, $R^n$ can represent all $n$-dimensional real vector spaces.
- Through the definition of an inner product, a vector space incorporates key geometric features of physical space.
- Continuous functions over an interval constitute an infinite dimensional vector space, complete with the usual notions.

Necessary Exercises: **6,7**



## Outline

## Derivatives in Multi-Dimensional Spaces



**Gradient**

$$\nabla f(\mathbf{x}) \equiv \frac{\partial f}{\partial \mathbf{x}}(\mathbf{x}) = \left[\frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \cdots \quad \frac{\partial f}{\partial x_n}\right]^T$$

Up to the first order, $\delta f \approx [\nabla f(\mathbf{x})]^T \delta \mathbf{x}$

**Directional derivative**

$$\frac{\partial f}{\partial \mathbf{d}} = \lim_{\alpha \to 0} \frac{f(\mathbf{x}+\alpha \mathbf{d}) - f(\mathbf{x})}{\alpha}$$

Relationships:

$$\frac{\partial f}{\partial \mathbf{e}_j} = \frac{\partial f}{\partial x_j}, \quad \frac{\partial f}{\partial \mathbf{d}} = \mathbf{d}^T \nabla f(\mathbf{x}) \quad \text{and} \quad \frac{\partial f}{\partial \hat{\mathbf{g}}} = \|\nabla f(\mathbf{x})\|$$

Among all unit vectors, taken as directions,

- the rate of change of a function in a direction is the same as the component of its gradient along that direction, and
- the rate of change along the direction of the gradient is the greatest and is equal to the magnitude of the gradient.



## Derivatives in Multi-Dimensional Spaces



**Hessian**

$$\mathbf{H}(\mathbf{x}) = \frac{\partial^2 f}{\partial \mathbf{x}^2} = \begin{bmatrix} \frac{\partial^2 f}{\partial {x_1}^2} & \frac{\partial^2 f}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_1} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial {x_2}^2} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_n} & \frac{\partial^2 f}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f}{\partial {x_n}^2} \end{bmatrix}$$

Meaning: $\nabla f(\mathbf{x}+\delta\mathbf{x}) - \nabla f(\mathbf{x}) \approx \left[\frac{\partial^2 f}{\partial \mathbf{x}^2}(\mathbf{x})\right] \delta \mathbf{x}$

For a vector function $\mathbf{h}(\mathbf{x})$, **Jacobian**

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}(\mathbf{x}) = \left[\frac{\partial \mathbf{h}}{\partial x_1} \quad \frac{\partial \mathbf{h}}{\partial x_2} \quad \cdots \quad \frac{\partial \mathbf{h}}{\partial x_n}\right]$$

Underlying notion: $\delta \mathbf{h} \approx [\mathbf{J}(\mathbf{x})]\delta \mathbf{x}$



## Taylor's Series



Taylor's formula in the remainder form:

$$f(x+\delta x) = f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \cdots + \frac{1}{(n-1)!}f^{(n-1)}(x)\delta x^{n-1} + \frac{1}{n!}f^{(n)}(x_c)\delta x^n$$

where $x_c = x + t\delta x$ with $0 \le t \le 1$
Mean value theorem: existence of $x_c$
Taylor's series:

$$f(x+\delta x) = f(x) + f'(x)\delta x + \frac{1}{2!}f''(x)\delta x^2 + \cdots$$

For a multivariate function,

$$\begin{aligned} f(\mathbf{x}+\delta\mathbf{x}) &= f(\mathbf{x}) + [\delta\mathbf{x}^T\nabla]f(\mathbf{x}) + \frac{1}{2!}[\delta\mathbf{x}^T\nabla]^2 f(\mathbf{x}) + \cdots \\ &\quad + \frac{1}{(n-1)!}[\delta\mathbf{x}^T\nabla]^{n-1}f(\mathbf{x}) + \frac{1}{n!}[\delta\mathbf{x}^T\nabla]^n f(\mathbf{x}+t\delta\mathbf{x}) \\ f(\mathbf{x}+\delta\mathbf{x}) &\approx f(\mathbf{x}) + [\nabla f(\mathbf{x})]^T\delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T\left[\frac{\partial^2 f}{\partial \mathbf{x}^2}(\mathbf{x})\right]\delta\mathbf{x} \end{aligned}$$



## Chain Rule and Change of Variables



For $f(\mathbf{x})$, the total differential:

$$df = [\nabla f(\mathbf{x})]^T d\mathbf{x} = \frac{\partial f}{\partial x_1}dx_1 + \frac{\partial f}{\partial x_2}dx_2 + \cdots + \frac{\partial f}{\partial x_n}dx_n$$

Ordinary derivative or total derivative:

$$\frac{df}{dt} = [\nabla f(\mathbf{x})]^T \frac{d\mathbf{x}}{dt}$$

For $f(t, \mathbf{x}(t))$, total derivative: $\frac{df}{dt} = \frac{\partial f}{\partial t} + [\nabla f(\mathbf{x})]^T \frac{d\mathbf{x}}{dt}$
For $f(\mathbf{v}, \mathbf{x}(\mathbf{v})) = f(v_1, v_2, \cdots, v_m, \ x_1(\mathbf{v}), x_2(\mathbf{v}), \cdots, x_n(\mathbf{v}))$,

$$\frac{\partial f}{\partial v_i}(\mathbf{v}, \mathbf{x}(\mathbf{v})) = \left(\frac{\partial f}{\partial v_i}\right)_x + \left[\frac{\partial f}{\partial \mathbf{x}}(\mathbf{v}, \mathbf{x})\right]^T \frac{\partial \mathbf{x}}{\partial v_i} = \left(\frac{\partial f}{\partial v_i}\right)_x + [\nabla_x f(\mathbf{v}, \mathbf{x})]^T \frac{\partial \mathbf{x}}{\partial v_i}$$

$$\Rightarrow \nabla f(\mathbf{v}, \mathbf{x}(\mathbf{v})) = \nabla_v f(\mathbf{v}, \mathbf{x}) + \left[\frac{\partial \mathbf{x}}{\partial \mathbf{v}}(\mathbf{v})\right]^T \nabla_x f(\mathbf{v}, \mathbf{x})$$

## Chain Rule and Change of Variables



Let $\mathbf{x} \in R^{m+n}$ and $\mathbf{h}(\mathbf{x}) \in R^m$.

Partition $\mathbf{x} \in R^{m+n}$ into $\mathbf{z} \in R^n$ and $\mathbf{w} \in R^m$.

System of equations $\mathbf{h}(\mathbf{x}) = \mathbf{0}$ means $\mathbf{h}(\mathbf{z}, \mathbf{w}) = \mathbf{0}$.

**Question:** Can we work out the function $\mathbf{w} = \mathbf{w}(\mathbf{z})$?

Solution of $m$ equations in $m$ unknowns?

**Question:** If we have one valid pair $(\mathbf{z}, \mathbf{w})$, then is it possible to develop $\mathbf{w} = \mathbf{w}(\mathbf{z})$ in the local neighbourhood?
**Answer:** Yes, if Jacobian $\frac{\partial \mathbf{h}}{\partial \mathbf{w}}$ is non-singular.

Implicit function theorem

$$\frac{\partial \mathbf{h}}{\partial \mathbf{z}} + \frac{\partial \mathbf{h}}{\partial \mathbf{w}} \frac{\partial \mathbf{w}}{\partial \mathbf{z}} = \mathbf{0} \Rightarrow \frac{\partial \mathbf{w}}{\partial \mathbf{z}} = -\left[\frac{\partial \mathbf{h}}{\partial \mathbf{w}}\right]^{-1} \left[\frac{\partial \mathbf{h}}{\partial \mathbf{z}}\right]$$

Upto first order, $\mathbf{w}_1 = \mathbf{w} + \left[\frac{\partial \mathbf{w}}{\partial \mathbf{z}}\right](\mathbf{z}_1 - \mathbf{z})$.



## Chain Rule and Change of Variables



For a multiple integral

$$I = \int \int_A \int f(x, y, z)\, dx\, dy\, dz,$$

change of variables $x = x(u, v, w)$, $y = y(u, v, w)$, $z = z(u, v, w)$ gives

$$I = \int \int_{\bar{A}} \int f(x(u,v,w), y(u,v,w), z(u,v,w))\, |J(u,v,w)|\, du\, dv\, dw,$$

where Jacobian determinant $|J(u,v,w)| = \left|\frac{\partial(x,y,z)}{\partial(u,v,w)}\right|$.
For the differential

$$P_1(\mathbf{x})dx_1 + P_2(\mathbf{x})dx_2 + \cdots + P_n(\mathbf{x})dx_n,$$

we ask: does there exist a function $f(\mathbf{x})$,

- of which this is the differential;
- or equivalently, the gradient of which is $\mathbf{P}(\mathbf{x})$?

*Perfect* or *exact differential*: can be integrated to find $f$.



## Chain Rule and Change of Variables



**Differentiation under the integral sign**

How To differentiate $\phi(x) = \phi(x, u(x), v(x)) = \int_{u(x)}^{v(x)} f(x,t)\, dt$?
In the expression

$$\phi'(x) = \frac{\partial \phi}{\partial x} + \frac{\partial \phi}{\partial u}\frac{du}{dx} + \frac{\partial \phi}{\partial v}\frac{dv}{dx},$$

we have $\frac{\partial \phi}{\partial x} = \int_u^v \frac{\partial f}{\partial x}(x,t)dt$.
Now, considering function $F(x,t)$ such that $f(x,t) = \frac{\partial F(x,t)}{\partial t}$,

$$\phi(x) = \int_u^v \frac{\partial F}{\partial t}(x,t)dt = F(x,v) - F(x,u) \equiv \phi(x,u,v).$$

Using $\frac{\partial \phi}{\partial v} = f(x,v)$ and $\frac{\partial \phi}{\partial u} = -f(x,u)$,

$$\phi'(x) = \int_{u(x)}^{v(x)} \frac{\partial f}{\partial x}(x,t)dt + f(x,v)\frac{dv}{dx} - f(x,u)\frac{du}{dx}.$$

Leibnitz rule



## Numerical Differentiation



Forward difference formula

$$f'(x) = \frac{f(x+\delta x) - f(x)}{\delta x} + \mathcal{O}(\delta x)$$

Central difference formulae

$$f'(x) = \frac{f(x+\delta x) - f(x-\delta x)}{2\delta x} + \mathcal{O}(\delta x^2)$$

$$f''(x) = \frac{f(x+\delta x) - 2f(x) + f(x-\delta x)}{\delta x^2} + \mathcal{O}(\delta x^2)$$

For gradient $\nabla f(\mathbf{x})$ and Hessian,

$$\frac{\partial f}{\partial x_i}(\mathbf{x}) = \frac{1}{2\delta}[f(\mathbf{x}+\delta\mathbf{e}_i) - f(\mathbf{x}-\delta\mathbf{e}_i)],$$

$$\frac{\partial^2 f}{\partial {x_i}^2}(\mathbf{x}) = \frac{f(\mathbf{x}+\delta\mathbf{e}_i) - 2f(\mathbf{x}) + f(\mathbf{x}-\delta\mathbf{e}_i)}{\delta^2}, \quad \text{and}$$

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{f(\mathbf{x}+\delta\mathbf{e}_i+\delta\mathbf{e}_j) - f(\mathbf{x}+\delta\mathbf{e}_i-\delta\mathbf{e}_j) - f(\mathbf{x}-\delta\mathbf{e}_i+\delta\mathbf{e}_j) + f(\mathbf{x}-\delta\mathbf{e}_i-\delta\mathbf{e}_j)}{4\delta^2}$$

# An Introduction to Tensors*

- Indicial notation and summation convention
- Kronecker delta and Levi-Civita symbol
- Rotation of reference axes
- Tensors of order zero, or scalars
- Contravariant and covariant tensors of order one, or vectors
- Cartesian tensors
- Cartesian tensors of order two
- Higher order tensors
- Elementary tensor operations
- Symmetric tensors
- Tensor fields
- $\cdots \quad \cdots \quad \cdots$



# Points to note

- Gradient, Hessian, Jacobian and the Taylor's series
- Partial and total gradients
- Implicit functions
- Leibnitz rule
- Numerical derivatives

Necessary Exercises: **2,3,4,8**



# Outline





# Recapitulation of Basic Notions

Dot and cross products: their implications

Scalar and vector triple products

Differentiation rules

Interface with matrix algebra:

$$\begin{aligned} \mathbf{a} \cdot \mathbf{x} &= \mathbf{a}^T \mathbf{x}, \\ (\mathbf{a} \cdot \mathbf{x})\mathbf{b} &= (\mathbf{b}\mathbf{a}^T)\mathbf{x}, \quad \text{and} \\ \mathbf{a} \times \mathbf{x} &= \begin{cases} \mathbf{a}_\perp^T \mathbf{x}, & \text{for 2-d vectors} \\ \tilde{\mathbf{a}}\mathbf{x}, & \text{for 3-d vectors} \end{cases} \end{aligned}$$

where

$$\mathbf{a}_\perp = \begin{bmatrix} -a_y \\ a_x \end{bmatrix} \quad \text{and} \quad \tilde{\mathbf{a}} = \begin{bmatrix} 0 & -a_z & a_y \\ a_z & 0 & -a_x \\ -a_y & a_x & 0 \end{bmatrix}$$

## Curves in Space



Explicit equation: $y = y(x)$ and $z = z(x)$
Implicit equation: $F(x, y, z) = 0 = G(x, y, z)$

**Parametric equation:**

$$\mathbf{r}(t) = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \equiv [x(t) \;\; y(t) \;\; z(t)]^T$$

- Tangent vector: $\mathbf{r}'(t)$
- Speed: $\|\mathbf{r}'\|$
- Unit tangent: $\mathbf{u}(t) = \frac{\mathbf{r}'}{\|\mathbf{r}'\|}$
- Length of the curve: $l = \int_a^b \|d\mathbf{r}\| = \int_a^b \sqrt{\mathbf{r}' \cdot \mathbf{r}'}\, dt$

Arc length function

$$s(t) = \int_a^t \sqrt{\mathbf{r}'(\tau) \cdot \mathbf{r}'(\tau)}\, d\tau$$

with $ds = \|d\mathbf{r}\| = \sqrt{dx^2 + dy^2 + dz^2}$ and $\frac{ds}{dt} = \|\mathbf{r}'\|$



## Curves in Space



Curve $\mathbf{r}(t)$ is *regular* if $\mathbf{r}'(t) \neq \mathbf{0}\ \forall t$.

- **Reparametrization** with respect to parameter $t^*$, some strictly increasing function of $t$

Observations

- Arc length $s(t)$ is obviously a monotonically increasing function.
- For a regular curve, $\frac{ds}{dt} \neq 0$.
- Then, $s(t)$ has an inverse function.
- Inverse $t(s)$ reparametrizes the curve as $\mathbf{r}(t(s))$.

For a **unit speed curve** $\mathbf{r}(s)$, $\|\mathbf{r}'(s)\| = 1$ and the unit tangent is

$$\mathbf{u}(s) = \mathbf{r}'(s).$$



## Curves in Space



**Curvature:** The rate at which the direction changes with arc length.

$$\kappa(s) = \|\mathbf{u}'(s)\| = \|\mathbf{r}''(s)\|$$

Unit principal normal:

$$\mathbf{p} = \frac{1}{\kappa}\mathbf{u}'(s)$$

With general parametrization,

$$\mathbf{r}''(t) = \frac{d\|\mathbf{r}'\|}{dt}\mathbf{u}(t) + \|\mathbf{r}'(t)\|\frac{d\mathbf{u}}{dt} = \frac{d\|\mathbf{r}'\|}{dt}\mathbf{u}(t) + \kappa(t)\|\mathbf{r}'\|^2\mathbf{p}(t)$$

- Osculating plane
- Centre of curvature
- Radius of curvature

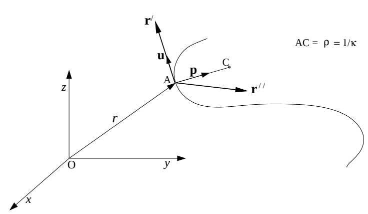



Figure: Tangent and normal to a curve



## Curves in Space



**Binormal:** $\mathbf{b} = \mathbf{u} \times \mathbf{p}$

**Serret-Frenet frame:** Right-handed triad $\{\mathbf{u}, \mathbf{p}, \mathbf{b}\}$

- Osculating, rectifying and normal planes

**Torsion:** Twisting out of the osculating plane

- rate of change of $\mathbf{b}$ with respect to arc length $s$

$$\mathbf{b}' = \mathbf{u}' \times \mathbf{p} + \mathbf{u} \times \mathbf{p}' = \kappa(s)\mathbf{p} \times \mathbf{p} + \mathbf{u} \times \mathbf{p}' = \mathbf{u} \times \mathbf{p}'$$

What is $\mathbf{p}'$?

Taking $\mathbf{p}' = \sigma\mathbf{u} + \tau\mathbf{b}$,

$$\mathbf{b}' = \mathbf{u} \times (\sigma\mathbf{u} + \tau\mathbf{b}) = -\tau\mathbf{p}.$$

**Torsion** of the curve

$$\tau(s) = -\mathbf{p}(s) \cdot \mathbf{b}'(s)$$

## Curves in Space

We have $\mathbf{u}'$ and $\mathbf{b}'$. What is $\mathbf{p}'$?

From $\mathbf{p} = \mathbf{b} \times \mathbf{u}$,

$$\mathbf{p}' = \mathbf{b}' \times \mathbf{u} + \mathbf{b} \times \mathbf{u}' = -\tau\mathbf{p} \times \mathbf{u} + \mathbf{b} \times \kappa\mathbf{p} = -\kappa\mathbf{u} + \tau\mathbf{b}.$$

**Serret-Frenet formulae**

$$\left.\begin{array}{rcl} \mathbf{u}' &=& \kappa\mathbf{p}, \\ \mathbf{p}' &=& -\kappa\mathbf{u} + \tau\mathbf{b}, \\ \mathbf{b}' &=& -\tau\mathbf{p} \end{array}\right\}$$

Intrinsic representation of a curve is complete with $\kappa(s)$ and $\tau(s)$.

> *The arc-length parametrization of a curve is completely determined by its curvature $\kappa(s)$ and torsion $\tau(s)$ functions, except for a rigid body motion.*



## Surfaces*

Parametric surface equation:

$$\mathbf{r}(u,v) = x(u,v)\mathbf{i} + y(u,v)\mathbf{j} + z(u,v)\mathbf{k} \equiv [x(u,v) \;\; y(u,v) \;\; z(u,v)]^T$$

Tangent vectors $\mathbf{r}_u$ and $\mathbf{r}_v$ define a tangent plane $\mathcal{T}$.

$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$ is normal to the surface and the unit normal is

$$\mathbf{n} = \frac{\mathbf{N}}{\|\mathbf{N}\|} = \frac{\mathbf{r}_u \times \mathbf{r}_v}{\|\mathbf{r}_u \times \mathbf{r}_v\|}.$$

**Question:** How does $\mathbf{n}$ vary over the surface?

Information on local geometry: *curvature* tensor

- Normal and principal curvatures
- Local shape: convex, concave, saddle, cylindrical, planar



## Points to note

- Parametric equation is the general and most convenient representation of curves and surfaces.
- Arc length is the natural parameter and the Serret-Frenet frame offers the natural frame of reference.
- Curvature and torsion are the only inherent properties of a curve.
- The local shape of a surface patch can be understood through an analysis of its curvature tensor.

Necessary Exercises: **1,2,3,6**



## Outline

## Differential Operations on Field Functions



Scalar point function or scalar field $\phi(x, y, z)$: $R^3 \to R$
Vector point function or vector field $\mathbf{V}(x, y, z)$: $R^3 \to R^3$

**The del or nabla ($\nabla$) operator**

$$\nabla \equiv \mathbf{i}\frac{\partial}{\partial x} + \mathbf{j}\frac{\partial}{\partial y} + \mathbf{k}\frac{\partial}{\partial z}$$

- $\nabla$ is a vector,
- it signifies a differentiation, and
- it operates from the left side.

Laplacian operator:

$$\nabla^2 \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \qquad = \nabla \cdot \nabla \quad ??$$

Laplace's equation:

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial y^2} + \frac{\partial^2 \phi}{\partial z^2} = 0$$

Solution of $\nabla^2 \phi = 0$: *harmonic function*



## Differential Operations on Field Functions



**Gradient**

$$\text{grad } \phi \equiv \nabla\phi = \frac{\partial \phi}{\partial x}\mathbf{i} + \frac{\partial \phi}{\partial y}\mathbf{j} + \frac{\partial \phi}{\partial z}\mathbf{k}$$

is orthogonal to the level surfaces.

Flow fields: $-\nabla\phi$ gives the velocity vector.

**Divergence**

For $\mathbf{V}(x, y, z) \equiv V_x(x, y, z)\mathbf{i} + V_y(x, y, z)\mathbf{j} + V_z(x, y, z)\mathbf{k}$,

$$\text{div } \mathbf{V} \equiv \nabla \cdot \mathbf{V} = \frac{\partial V_x}{\partial x} + \frac{\partial V_y}{\partial y} + \frac{\partial V_z}{\partial z}$$

Divergence of $\rho\mathbf{V}$: flow rate of mass per unit volume out of the control volume.

Similar relation between field and flux in electromagnetics.



## Differential Operations on Field Functions



**Curl**

$$\begin{aligned}\text{curl } \mathbf{V} &\equiv \nabla \times \mathbf{V} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \frac{\partial}{\partial x} & \frac{\partial}{\partial y} & \frac{\partial}{\partial z} \\ V_x & V_y & V_z \end{vmatrix} \\ &= \left(\frac{\partial V_z}{\partial y} - \frac{\partial V_y}{\partial z}\right)\mathbf{i} + \left(\frac{\partial V_x}{\partial z} - \frac{\partial V_z}{\partial x}\right)\mathbf{j} + \left(\frac{\partial V_y}{\partial x} - \frac{\partial V_x}{\partial y}\right)\mathbf{k}\end{aligned}$$

If $\mathbf{V} = \omega \times \mathbf{r}$ represents the velocity field, then angular velocity

$$\omega = \frac{1}{2} \text{curl } \mathbf{V}.$$

Curl represents rotationality.

Connections between electric and magnetic fields!



## Differential Operations on Field Functions



**Composite operations**

Operator $\nabla$ is linear.

$$\begin{aligned}\nabla(\phi + \psi) &= \nabla\phi + \nabla\psi, \\ \nabla \cdot (\mathbf{V} + \mathbf{W}) &= \nabla \cdot \mathbf{V} + \nabla \cdot \mathbf{W}, \quad \text{and} \\ \nabla \times (\mathbf{V} + \mathbf{W}) &= \nabla \times \mathbf{V} + \nabla \times \mathbf{W}.\end{aligned}$$

Considering the products $\phi\psi$, $\phi\mathbf{V}$, $\mathbf{V} \cdot \mathbf{W}$, and $\mathbf{V} \times \mathbf{W}$;

$$\nabla(\phi\psi) = \psi\nabla\phi + \phi\nabla\psi$$
$$\nabla \cdot (\phi\mathbf{V}) = \nabla\phi \cdot \mathbf{V} + \phi\nabla \cdot \mathbf{V}$$
$$\nabla \times (\phi\mathbf{V}) = \nabla\phi \times \mathbf{V} + \phi\nabla \times \mathbf{V}$$
$$\nabla(\mathbf{V} \cdot \mathbf{W}) = (\mathbf{W} \cdot \nabla)\mathbf{V} + (\mathbf{V} \cdot \nabla)\mathbf{W} + \mathbf{W} \times (\nabla \times \mathbf{V}) + \mathbf{V} \times (\nabla \times \mathbf{W})$$
$$\nabla \cdot (\mathbf{V} \times \mathbf{W}) = \mathbf{W} \cdot (\nabla \times \mathbf{V}) - \mathbf{V} \cdot (\nabla \times \mathbf{W})$$
$$\nabla \times (\mathbf{V} \times \mathbf{W}) = (\mathbf{W} \cdot \nabla)\mathbf{V} - \mathbf{W}(\nabla \cdot \mathbf{V}) - (\mathbf{V} \cdot \nabla)\mathbf{W} + \mathbf{V}(\nabla \cdot \mathbf{W})$$

*Note:* the expression $\mathbf{V} \cdot \nabla \equiv V_x\frac{\partial}{\partial x} + V_y\frac{\partial}{\partial y} + V_z\frac{\partial}{\partial z}$ is an operator!

## Differential Operations on Field Functions



**Second order differential operators**

$$\begin{aligned} \text{div grad } \phi &\equiv \nabla \cdot (\nabla \phi) \\ \text{curl grad } \phi &\equiv \nabla \times (\nabla \phi) \\ \text{div curl } \mathbf{V} &\equiv \nabla \cdot (\nabla \times \mathbf{V}) \\ \text{curl curl } \mathbf{V} &\equiv \nabla \times (\nabla \times \mathbf{V}) \\ \text{grad div } \mathbf{V} &\equiv \nabla (\nabla \cdot \mathbf{V}) \end{aligned}$$

Important identities:

$$\begin{aligned} \text{div grad } \phi &\equiv \nabla \cdot (\nabla \phi) = \nabla^2 \phi \\ \text{curl grad } \phi &\equiv \nabla \times (\nabla \phi) = \mathbf{0} \\ \text{div curl } \mathbf{V} &\equiv \nabla \cdot (\nabla \times \mathbf{V}) = 0 \\ \text{curl curl } \mathbf{V} &\equiv \nabla \times (\nabla \times \mathbf{V}) \\ &= \nabla(\nabla \cdot \mathbf{V}) - \nabla^2 \mathbf{V} = \text{ grad div } \mathbf{V} - \nabla^2 \mathbf{V} \end{aligned}$$



## Integral Operations on Field Functions



**Line integral** along curve $C$:

$$I = \int_C \mathbf{V} \cdot d\mathbf{r} = \int_C (V_x dx + V_y dy + V_z dz)$$

For a parametrized curve $\mathbf{r}(t)$, $t \in [a, b]$,

$$I = \int_C \mathbf{V} \cdot d\mathbf{r} = \int_a^b \mathbf{V} \cdot \frac{d\mathbf{r}}{dt} dt.$$

For simple (non-intersecting) paths contained in a simply connected region, equivalent statements:

- $V_x dx + V_y dy + V_z dz$ is an exact differential.
- $\mathbf{V} = \nabla \phi$ for some $\phi(\mathbf{r})$.
- $\int_C \mathbf{V} \cdot d\mathbf{r}$ is independent of path.
- Circulation $\oint \mathbf{V} \cdot d\mathbf{r} = 0$ around any closed path.
- curl $\mathbf{V} = \mathbf{0}$.
- Field $\mathbf{V}$ is conservative.



## Integral Operations on Field Functions



**Surface integral** over an orientable surface $S$:

$$J = \int_S \int \mathbf{V} \cdot d\mathbf{S} = \int_S \int \mathbf{V} \cdot \mathbf{n} dS$$

For $\mathbf{r}(u, w)$, $dS = \|\mathbf{r}_u \times \mathbf{r}_w\| \, du \, dw$ and

$$J = \int_S \int \mathbf{V} \cdot \mathbf{n} dS = \int_R \int \mathbf{V} \cdot (\mathbf{r}_u \times \mathbf{r}_w) \, du \, dw.$$

**Volume integrals** of point functions over a region $T$:

$$M = \int \int_T \int \phi dv \quad \text{and} \quad \mathbf{F} = \int \int_T \int \mathbf{V} dv$$



## Integral Theorems



**Green's theorem in the plane**

*$R$: closed bounded region in the xy-plane*
*$C$: boundary, a piecewise smooth closed curve*
*$F_1(x, y)$ and $F_2(x, y)$: first order continuous functions*

$$\oint_C (F_1 dx + F_2 dy) = \int_R \int \left( \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dx \, dy$$

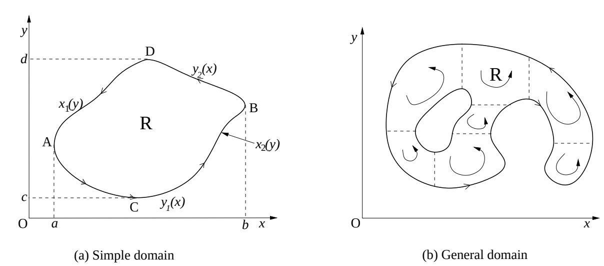



Figure: Regions for proof of Green's theorem in the plane

## Integral Theorems



**Proof:**

$$\begin{aligned}\int_R\int \frac{\partial F_1}{\partial y}dxdy &= \int_a^b \int_{y_1(x)}^{y_2(x)} \frac{\partial F_1}{\partial y}dydx\\ &= \int_a^b [F_1\{x,y_2(x)\} - F_1\{x,y_1(x)\}]dx\\ &= -\int_b^a F_1\{x,y_2(x)\}dx - \int_a^b F_1\{x,y_1(x)\}dx\\ &= -\oint_C F_1(x,y)dx\end{aligned}$$

$$\int_R\int \frac{\partial F_2}{\partial x}dxdy = \int_c^d \int_{x_1(y)}^{x_2(y)} \frac{\partial F_2}{\partial x}dxdy = \oint_C F_2(x,y)dy$$

Difference: $\oint_C (F_1 dx + F_2 dy) = \int_R\int \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dx\, dy$

In alternative form, $\oint_C \mathbf{F}\cdot d\mathbf{r} = \int_R\int \text{ curl } \mathbf{F}\cdot\mathbf{k}\ dx\, dy$.



## Integral Theorems



**Gauss's divergence theorem**

*T: a closed bounded region*
*S: boundary, a piecewise smooth closed orientable surface*
$\mathbf{F}(x,y,z)$*: a first order continuous vector function*

$$\int\int_T\int \ div\, \mathbf{F} dv = \int_S\int \mathbf{F}\cdot\mathbf{n}dS$$

Interpretation of the definition extended to finite domains.

$$\int\int_T\int \left(\frac{\partial F_x}{\partial x} + \frac{\partial F_y}{\partial y} + \frac{\partial F_z}{\partial z}\right) dx\, dy\, dz = \int_S\int (F_x n_x + F_y n_y + F_z n_z)dS$$

To show: $\int\int_T\int \frac{\partial F_z}{\partial z}dx\, dy\, dz = \int_S\int F_z n_z dS$
*First consider a region, the boundary of which is intersected at most twice by any line parallel to a coordinate axis.*



## Integral Theorems



Lower and upper segments of $S$: $z = z_1(x,y)$ and $z = z_2(x,y)$.

$$\begin{aligned}\int\int_T\int \frac{\partial F_z}{\partial z}dx\, dy\, dz &= \int_R\int \left[\int_{z_1}^{z_2} \frac{\partial F_z}{\partial z}dz\right] dx\, dy\\ &= \int_R\int [F_z\{x,y,z_2(x,y)\} - F_z\{x,y,z_1(x,y)\}]dx\, dy\end{aligned}$$

$R$: projection of $T$ on the $xy$-plane

Projection of area element of the upper segment: $n_z dS = dx\, dy$
Projection of area element of the lower segment: $n_z dS = -dx\, dy$

Thus, $\int\int_T\int \frac{\partial F_z}{\partial z}dx\, dy\, dz = \int_S\int F_z n_z dS$.

Sum of three such components leads to the result.

Extension to arbitrary regions by a suitable subdivision of domain!



## Integral Theorems



**Green's identities (theorem)**

*Region $T$ and boundary $S$: as required in premises of Gauss's theorem*
$\phi(x,y,z)$ *and* $\psi(x,y,z)$*: second order continuous scalar functions*

$$\begin{aligned}\int_S\int \phi\nabla\psi\cdot\mathbf{n}dS &= \int\int_T\int (\phi\nabla^2\psi + \nabla\phi\cdot\nabla\psi)dv\\ \int_S\int (\phi\nabla\psi - \psi\nabla\phi)\cdot\mathbf{n}dS &= \int\int_T\int (\phi\nabla^2\psi - \psi\nabla^2\phi)dv\end{aligned}$$

Direct consequences of Gauss's theorem

To establish, apply Gauss's divergence theorem on $\phi\nabla\psi$, and then on $\psi\nabla\phi$ as well.

# Integral Theorems

**Stokes's theorem**

*S: a piecewise smooth surface*
*C: boundary, a piecewise smooth simple closed curve*
$\mathbf{F}(x, y, z)$*: first order continuous vector function*

$$\oint_C \mathbf{F} \cdot d\mathbf{r} = \int_S \int \, curl \, \mathbf{F} \cdot \mathbf{n} dS$$

$\mathbf{n}$*: unit normal given by the right hand clasp rule on C*

For $\mathbf{F}(x, y, z) = F_x(x, y, z)\mathbf{i}$,

$$\oint_C F_x dx = \int_S \int \left( \frac{\partial F_x}{\partial z}\mathbf{j} - \frac{\partial F_x}{\partial y}\mathbf{k} \right) \cdot \mathbf{n} dS = \int_S \int \left( \frac{\partial F_x}{\partial z} n_y - \frac{\partial F_x}{\partial y} n_z \right) dS.$$

First, consider a surface $S$ intersected at most once by any line parallel to a coordinate axis.



# Integral Theorems

Represent $S$ as $z = z(x, y) \equiv f(x, y)$.

Unit normal $\mathbf{n} = [n_x \;\; n_y \;\; n_z]^T$ is proportional to $[\frac{\partial f}{\partial x} \;\; \frac{\partial f}{\partial y} \;\; -1]^T$.

$$n_y = -n_z \frac{\partial z}{\partial y}$$

$$\int_S \int \left( \frac{\partial F_x}{\partial z} n_y - \frac{\partial F_x}{\partial y} n_z \right) dS = -\int_S \int \left( \frac{\partial F_x}{\partial y} + \frac{\partial F_x}{\partial z}\frac{\partial z}{\partial y} \right) n_z dS$$

Over projection $R$ of $S$ on $xy$-plane, $\phi(x, y) = F_x(x, y, z(x, y))$.

$$\text{LHS} = -\int_R \int \frac{\partial \phi}{\partial y} dx\, dy = \oint_{C'} \phi(x, y) dx = \oint_C F_x dx$$

Similar results for $F_y(x, y, z)\mathbf{j}$ and $F_z(x, y, z)\mathbf{k}$.



# Points to note

- The 'del' operator $\nabla$
- Gradient, divergence and curl
- Composite and second order operators
- Line, surface and volume intergals
- Green's, Gauss's and Stokes's theorems
- Applications in physics (and engineering)

Necessary Exercises: **1,2,3,6,7**



# Outline

# Basic Principles



**Fundamental theorem of algebra**

$$p(x) = a_0x^n + a_1x^{n-1} + a_2x^{n-2} + \cdots + a_{n-1}x + a_n$$

has exactly $n$ roots $x_1, x_2, \cdots, x_n$; with

$$p(x) = a_0(x - x_1)(x - x_2)(x - x_3)\cdots(x - x_n).$$

In general, roots are complex.
**Multiplicity:** A root of $p(x)$ with multiplicity $k$ satisfies

$$p(x) = p'(x) = p''(x) = \cdots = p^{(k-1)}(x) = 0.$$

- **Descartes' rule of signs**
- **Bracketing and separation**
- **Synthetic division and deflation**

$$p(x) = f(x)q(x) + r(x)$$



# Analytical Solution



**Quadratic equation**

$$ax^2 + bx + c = 0 \;\Rightarrow\; x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Method of completing the square:

$$x^2 + \frac{b}{a}x + \left(\frac{b}{2a}\right)^2 = \frac{b^2}{4a^2} - \frac{c}{a} \;\Rightarrow\; \left(x + \frac{b}{2a}\right)^2 = \frac{b^2 - 4ac}{4a^2}$$

**Cubic equations (Cardano)**:

$$x^3 + ax^2 + bx + c = 0$$

Completing the cube?
Substituting $y = x + k$,

$$y^3 + (a - 3k)y^2 + (b - 2ak + 3k^2)y + (c - bk + ak^2 - k^3) = 0.$$

Choose the shift $k = a/3$.



# Analytical Solution



$$\boxed{y^3 + py + q = 0}$$

Assuming $y = u + v$, we have $y^3 = u^3 + v^3 + 3uv(u + v)$.

$$\begin{aligned} uv &= -p/3 \\ u^3 + v^3 &= -q \\ \text{and hence} \quad (u^3 - v^3)^2 &= q^2 + \frac{4p^3}{27}. \end{aligned}$$

Solution:

$$u^3, v^3 = -\frac{q}{2} \pm \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} \;=\; A, B \quad \text{(say)}.$$

$u = A_1, A_1\omega, A_1\omega^2$, and $v = B_1, B_1\omega, B_1\omega^2$

$$y_1 = A_1 + B_1,\; y_2 = A_1\omega + B_1\omega^2 \text{ and } y_3 = A_1\omega^2 + B_1\omega.$$

At least one of the roots is real!!



# Analytical Solution



**Quartic equations (Ferrari)**

$$x^4 + ax^3 + bx^2 + cx + d = 0 \;\Rightarrow\; \left(x^2 + \frac{a}{2}x\right)^2 = \left(\frac{a^2}{4} - b\right)x^2 - cx - d$$

For a perfect square,

$$\left(x^2 + \frac{a}{2}x + \frac{y}{2}\right)^2 = \left(\frac{a^2}{4} - b + y\right)x^2 + \left(\frac{ay}{2} - c\right)x + \left(\frac{y^2}{4} - d\right)$$

*Under what condition*, the new RHS will be a perfect square?

$$\left(\frac{ay}{2} - c\right)^2 - 4\left(\frac{a^2}{4} - b + y\right)\left(\frac{y^2}{4} - d\right) = 0$$

Resolvent of a quartic:

$$y^3 - by^2 + (ac - 4d)y + (4bd - a^2d - c^2) = 0$$

# Analytical Solution



**Procedure**

- Frame the cubic resolvent.
- Solve this cubic equation.
- Pick up one solution as $y$.
- Insert this $y$ to form

$$\left(x^2+\frac{a}{2}x+\frac{y}{2}\right)^2=(ex+f)^2.$$

- Split it into two quadratic equations as

$$x^2+\frac{a}{2}x+\frac{y}{2}=\pm(ex+f).$$

- Solve each of the two quadratic equations to obtain a total of four solutions of the original quartic equation.



# General Polynomial Equations



Analytical solution of the general quintic equation?
Galois: group theory:

> *A general quintic, or higher degree, equation is not* solvable *by radicals.*

**General polynomial equations:** iterative algorithms

- Methods for nonlinear equations
- Methods specific to *polynomial equations*

**Solution through the companion matrix**

> *Roots of a polynomial equation are the same as the eigenvalues of its companion matrix.*

$$\text{Companion matrix: }\begin{bmatrix} 0 & 0 & \cdots & 0 & -a_n \\ 1 & 0 & \cdots & 0 & -a_{n-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -a_2 \\ 0 & 0 & \cdots & 1 & -a_1 \end{bmatrix}$$



# General Polynomial Equations



**Bairstow's method**

> *to separate out factors of small degree.*

Attempt to separate real linear factors?

$$\boxed{\text{Real quadratic factors}}$$

Synthetic division with a guess factor $x^2+q_1x+q_2$:

> remainder $r_1x+r_2$

$\mathbf{r}=[r_1 \;\; r_2]^T$ is a vector function of $\mathbf{q}=[q_1 \;\; q_2]^T$.

Iterate over $(q_1,q_2)$ to make $(r_1,r_2)$ zero.

Newton-Raphson (Jacobian based) iteration: see exercise.



# Two Simultaneous Equations



$$\begin{aligned} p_1x^2+q_1xy+r_1y^2+u_1x+v_1y+w_1 &= 0 \\ p_2x^2+q_2xy+r_2y^2+u_2x+v_2y+w_2 &= 0 \end{aligned}$$

Rearranging,

$$\begin{aligned} a_1x^2+b_1x+c_1 &= 0 \\ a_2x^2+b_2x+c_2 &= 0 \end{aligned}$$

Cramer's rule:

$$\frac{x^2}{b_1c_2-b_2c_1}=\frac{-x}{a_1c_2-a_2c_1}=\frac{1}{a_1b_2-a_2b_1}$$

$$\Rightarrow\; x=-\frac{b_1c_2-b_2c_1}{a_1c_2-a_2c_1}=-\frac{a_1c_2-a_2c_1}{a_1b_2-a_2b_1}$$

Consistency condition:

$$(a_1b_2-a_2b_1)(b_1c_2-b_2c_1)-(a_1c_2-a_2c_1)^2=0$$

$$\boxed{\text{A 4th degree equation in } y}$$

## Elimination Methods*



*The method operates similarly even if the degrees of the original equations in $y$ are higher.*

What about the degree of the eliminant equation?

*Two equations in $x$ and $y$ of degrees $n_1$ and $n_2$:*
*$x$-eliminant is an equation of degree $n_1 n_2$ in $y$*

Maximum number of solutions:

*Bezout number* $= n_1 n_2$

Note: *Deficient systems* may have less number of solutions.

Classical methods of elimination

- Sylvester's dialytic method
- Bezout's method



## Advanced Techniques*



Three or more independent equations in as many unknowns?

- Cascaded elimination? Objections!
- Exploitation of special structures through *clever heuristics*

  *(mechanisms kinematics literature)*
- Gröbner basis representation

  *(algebraic geometry)*
- Continuation or homotopy method by Morgan

  *For solving the system* $\mathbf{f}(\mathbf{x}) = \mathbf{0}$, *identify another structurally similar system* $\mathbf{g}(\mathbf{x}) = \mathbf{0}$ *with known solutions and construct the parametrized system*

  $$\mathbf{h}(\mathbf{x}) = t\mathbf{f}(\mathbf{x}) + (1-t)\mathbf{g}(\mathbf{x}) = \mathbf{0} \quad \text{for} \quad t \in [0,1].$$

  *Track each solution from $t = 0$ to $t = 1$.*



## Points to note



- Roots of cubic and quartic polynomials by the methods of Cardano and Ferrari
- For higher degree polynomials,
  - Bairstow's method: a clever implementation of Newton-Raphson method for polynomials
  - Eigenvalue problem of a companion matrix
- Reduction of a system of polynomial equations in two unknowns by elimination

Necessary Exercises: **1,3,4,6**



## Outline

# Methods for Nonlinear Equations

Algebraic and transcendental equations in the form

$$f(x) = 0$$

Practical problem: to find *one* real root (zero) of $f(x)$

Example of $f(x)$: $x^3 - 2x + 5$, $x^3 \ln x - \sin x + 2$, etc.

If $f(x)$ is continuous, then

Bracketing: $f(x_0)f(x_1) < 0 \Rightarrow$ there must be a root of $f(x)$ between $x_0$ and $x_1$.

Bisection: Check the sign of $f(\frac{x_0+x_1}{2})$. Replace either $x_0$ or $x_1$ with $\frac{x_0+x_1}{2}$.



# Methods for Nonlinear Equations

**Fixed point iteration**

Rearrange $f(x) = 0$ in the form $x = g(x)$.

Example:
For $f(x) = \tan x - x^3 - 2$, possible rearrangements:
$g_1(x) = \tan^{-1}(x^3 + 2)$
$g_2(x) = (\tan x - 2)^{1/3}$
$g_3(x) = \frac{\tan x - 2}{x^2}$

Iteration: $x_{k+1} = g(x_k)$

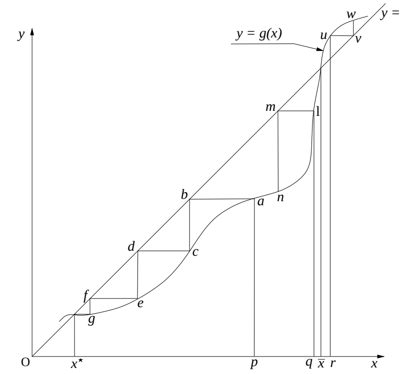



Figure: Fixed point iteration

*If $x^*$ is the unique solution in interval $J$ and $|g'(x)| \le h < 1$ in $J$, then any $x_0 \in J$ converges to $x^*$.*



# Methods for Nonlinear Equations

**Newton-Raphson method**

First order Taylor series
$f(x + \delta x) \approx f(x) + f'(x)\delta x$
From $f(x_k + \delta x) = 0$,
$\delta x = -f(x_k)/f'(x_k)$
Iteration:
$x_{k+1} = x_k - f(x_k)/f'(x_k)$
Convergence criterion:
$|f(x)f''(x)| < |f'(x)|^2$
Draw tangent to $f(x)$.
Take its $x$-intercept.

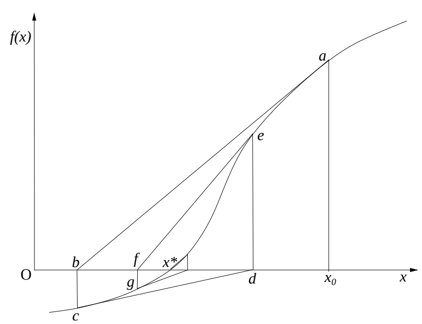



Figure: Newton-Raphson method

Merit: quadratic speed of convergence: $|x_{k+1} - x^*| = c|x_k - x^*|^2$
Demerit: If the starting point is not appropriate,

*haphazard wandering, oscillations or outright divergence!*



# Methods for Nonlinear Equations

**Secant method and method of false position**

In the Newton-Raphson formula,
$f'(x) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}$

$\Rightarrow \quad x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} f(x_k)$

Draw the chord or secant to $f(x)$ through $(x_{k-1}, f(x_{k-1}))$ and $(x_k, f(x_k))$.
Take its $x$-intercept.

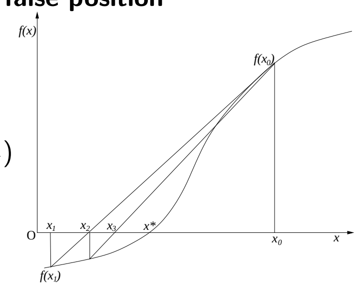



Figure: Method of false position

Special case: Maintain a bracket over the root at every iteration.

*The method of false position or regula falsi*

Convergence is guaranteed!

## Methods for Nonlinear Equations



**Quadratic interpolation method or Muller method**

Evaluate $f(x)$ at three points and model $y = a + bx + cx^2$. Set $y = 0$ and solve for $x$.

**Inverse quadratic interpolation**

Evaluate $f(x)$ at three points and model $x = a + by + cy^2$. Set $y = 0$ to get $x = a$.

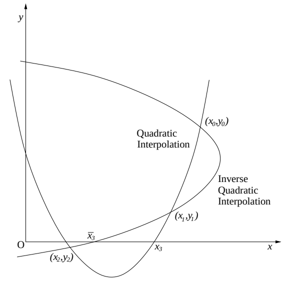



Figure: Interpolation schemes

**Van Wijngaarden-Dekker Brent method**

- maintains the bracket,
- uses inverse quadratic interpolation, and
- accepts outcome if within bounds, else takes a bisection step.

Opportunistic manoeuvring between a fast method and a safe one!



## Systems of Nonlinear Equations



$$
\begin{aligned}
f_1(x_1, x_2, \cdots, x_n) &= 0, \\
f_2(x_1, x_2, \cdots, x_n) &= 0, \\
\cdots \quad \cdots \quad \cdots \quad & \quad \cdots \\
f_n(x_1, x_2, \cdots, x_n) &= 0.
\end{aligned}
$$

$$\boxed{\mathbf{f}(\mathbf{x}) = \mathbf{0}}$$

- Number of variables and number of equations?
- No bracketing!
- Fixed point iteration schemes $\mathbf{x} = \mathbf{g}(\mathbf{x})$?

**Newton's method for systems of equations**

$$\mathbf{f}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{f}(\mathbf{x}) + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x})\right] \delta\mathbf{x} + \cdots \approx \mathbf{f}(\mathbf{x}) + \mathbf{J}(\mathbf{x})\delta\mathbf{x}$$

$$\Rightarrow \ \mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{J}(\mathbf{x}_k)]^{-1}\mathbf{f}(\mathbf{x}_k)$$

with the usual merits and demerits!



## Closure



**Modified Newton's method**

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k[\mathbf{J}(\mathbf{x}_k)]^{-1}\mathbf{f}(\mathbf{x}_k)$$

**Broyden's secant method**

*Jacobian is not* evaluated *at every iteration, but gets* developed *through updates.*

**Optimization-based formulation**

Global minimum of the function

$$\|\mathbf{f}(\mathbf{x})\|^2 = f_1^2 + f_2^2 + \cdots + f_n^2$$

**Levenberg-Marquardt method**



## Points to note



- Iteration schemes for solving $f(x) = 0$
- Newton (or Newton-Raphson) iteration for a system of equations
$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{J}(\mathbf{x}_k)]^{-1}\mathbf{f}(\mathbf{x}_k)$$
- Optimization formulation of a multi-dimensional root finding problem

Necessary Exercises: **1,2,3**

# Outline





# The Methodology of Optimization

- Parameters and variables
- The statement of the optimization problem

$$\begin{array}{ll} \textit{Minimize} & f(\mathbf{x}) \\ \textit{subject to} & \mathbf{g}(\mathbf{x}) \le \mathbf{0}, \\ & \mathbf{h}(\mathbf{x}) = \mathbf{0}. \end{array}$$

- Optimization methods
- Sensitivity analysis
- Optimization problems: unconstrained and constrained
- Optimization problems: linear and nonlinear
- Single-variable and multi-variable problems



# Single-Variable Optimization

For a function $f(x)$, a point $x^*$ is defined as a relative (local) minimum if $\exists\ \epsilon$ such that $f(x) \geq f(x^*)\ \forall\ x \in [x^* - \epsilon, x^* + \epsilon]$.

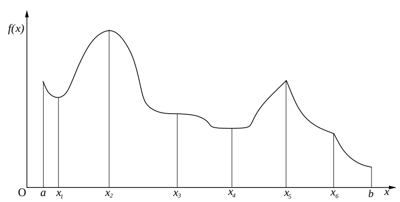



Figure: Schematic of optima of a univariate function

**Optimality criteria**

First order necessary condition: If $x^*$ is a local minimum or maximum point and if $f'(x^*)$ exists, then $f'(x^*) = 0$.

Second order necessary condition: If $x^*$ is a local minimum point and $f''(x^*)$ exists, then $f''(x^*) \geq 0$.

Second order sufficient condition: If $f'(x^*) = 0$ and $f''(x^*) > 0$ then $x^*$ is a local minimum point.



# Single-Variable Optimization

Higher order analysis: From Taylor's series,

$$\begin{aligned} \Delta f &= f(x^* + \delta x) - f(x^*) \\ &= f'(x^*)\delta x + \frac{1}{2!}f''(x^*)\delta x^2 + \frac{1}{3!}f'''(x^*)\delta x^3 + \frac{1}{4!}f^{iv}(x^*)\delta x^4 + \cdots \end{aligned}$$

*For an extremum to occur at point $x^*$, the lowest order derivative with non-zero value should be of even order.*

If $f'(x^*) = 0$, then

- $x^*$ is a *stationary point*, a candidate for an extremum.
- Evaluate higher order derivatives till one of them is found to be non-zero.
  - If its order is odd, then $x^*$ is an inflection point.
  - If its order is even, then $x^*$ is a local minimum or maximum, as the derivative value is positive or negative, respectively.

## Single-Variable Optimization

**Iterative methods of line search**

Methods based on gradient root finding

- Newton's method

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

- Secant method

$$x_{k+1} = x_k - \frac{x_k - x_{k-1}}{f'(x_k) - f'(x_{k-1})} f'(x_k)$$

- Method of cubic estimation

  *point of vanishing gradient of the cubic fit with* $f(x_{k-1})$, $f(x_k)$, $f'(x_{k-1})$ *and* $f'(x_k)$

- Method of quadratic estimation

  *point of vanishing gradient of the quadratic fit through three points*

Disadvantage: treating all stationary points alike!



## Single-Variable Optimization

Bracketing:

$$x_1 < x_2 < x_3 \text{ with } f(x_1) \geq f(x_2) \leq f(x_3)$$

Exhaustive search method or its variants

Direct optimization algorithms

- **Fibonacci search** uses a pre-defined number $N$, of function evaluations, and the Fibonacci sequence

$$F_0 = 1,\ F_1 = 1,\ F_2 = 2,\ \cdots,\ F_j = F_{j-2} + F_{j-1},\ \cdots$$

  to tighten a bracket with economized number of function evaluations.

- **Golden section search** uses a constant ratio

$$\tau = \frac{\sqrt{5}-1}{2} \approx 0.618,$$

  the *golden section ratio*, of interval reduction, that is determined as the limiting case of $N \to \infty$ and the actual number of steps is decided by the accuracy desired.



## Conceptual Background of Multivariate Optimization

Unconstrained minimization problem

$\mathbf{x}^*$ *is called a local minimum of* $f(\mathbf{x})$ *if* $\exists\ \delta$ *such that* $f(\mathbf{x}) \geq f(\mathbf{x}^*)$ *for all* $\mathbf{x}$ *satisfying* $\|\mathbf{x} - \mathbf{x}^*\| < \delta$.

**Optimality criteria**

From Taylor's series,

$$f(\mathbf{x}) - f(\mathbf{x}^*) = [\mathbf{g}(\mathbf{x}^*)]^T \delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T[\mathbf{H}(\mathbf{x}^*)]\delta\mathbf{x} + \cdots.$$

For $\mathbf{x}^*$ to be a local minimum,

necessary condition: $\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$ and $\mathbf{H}(\mathbf{x}^*)$ is positive semi-definite,

sufficient condition: $\mathbf{g}(\mathbf{x}^*) = \mathbf{0}$ and $\mathbf{H}(\mathbf{x}^*)$ is positive definite.

Indefinite Hessian matrix characterizes a *saddle point*.



## Conceptual Background of Multivariate Optimization

**Convexity**

Set $S \subseteq R^n$ is a *convex set* if

$$\forall\ \mathbf{x}_1, \mathbf{x}_2 \in S \ \text{ and } \ \alpha \in (0,1),\ \alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2 \in S.$$

Function $f(x)$ over a convex set $S$: a *convex function* if $\forall\ \mathbf{x}_1, \mathbf{x}_2 \in S$ and $\alpha \in (0,1)$,

$$f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \leq \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2).$$

Chord approximation is an *overestimate* at intermediate points!

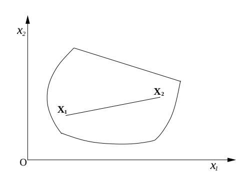



Figure: A convex domain

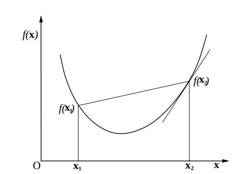



Figure: A convex function

## Conceptual Background of Multivariate Optimization

**First order characterization of convexity**

From $f(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2) \le \alpha f(\mathbf{x}_1) + (1-\alpha)f(\mathbf{x}_2)$,

$$f(\mathbf{x}_1) - f(\mathbf{x}_2) \ge \frac{f(\mathbf{x}_2 + \alpha(\mathbf{x}_1 - \mathbf{x}_2)) - f(\mathbf{x}_2)}{\alpha}.$$

As $\alpha \to 0$, $\quad f(\mathbf{x}_1) \ge f(\mathbf{x}_2) + [\nabla f(\mathbf{x}_2)]^T(\mathbf{x}_1 - \mathbf{x}_2)$.

Tangent approximation is an *underestimate* at intermediate points!

**Second order characterization:** Hessian is positive semi-definite.

**Convex programming problem**: convex function over convex set

*A local minimum is also a global minimum, and all minima are connected in a convex set.*

*Note:* Convexity is a stronger condition than unimodality!



## Conceptual Background of Multivariate Optimization

**Quadratic function**

$$q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x} + c$$

Gradient $\nabla q(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ and Hessian $= \mathbf{A}$ is constant.

- If $\mathbf{A}$ is positive definite, then the unique solution of $\mathbf{A}\mathbf{x} = -\mathbf{b}$ is the only minimum point.
- If $\mathbf{A}$ is positive semi-definite and $-\mathbf{b} \in Range(\mathbf{A})$, then the entire subspace of solutions of $\mathbf{A}\mathbf{x} = -\mathbf{b}$ are global minima.
- If $\mathbf{A}$ is positive semi-definite but $-\mathbf{b} \notin Range(\mathbf{A})$, then the function is unbounded!

**Note:** A quadratic problem (with positive definite Hessian) acts as a benchmark for optimization algorithms.



## Conceptual Background of Multivariate Optimization

**Optimization Algorithms**

From the *current* point, move to *another* point, hopefully *better*.

**Which way to go? How far to go? Which decision is first?**

Strategies and versions of algorithms:

Trust Region: Develop a *local* quadratic model

$$f(\mathbf{x}_k + \delta\mathbf{x}) = f(\mathbf{x}_k) + [\mathbf{g}(\mathbf{x}_k)]^T\delta\mathbf{x} + \frac{1}{2}\delta\mathbf{x}^T\mathbf{F}_k\delta\mathbf{x},$$

and minimize it in a small trust region around $\mathbf{x}_k$. (Define trust region with dummy boundaries.)

Line search: Identify a *descent direction* $\mathbf{d}_k$ and minimize the function along it through the univariate function

$$\phi(\alpha) = f(\mathbf{x}_k + \alpha\mathbf{d}_k).$$

- *Exact* or *accurate* line search
- *Inexact* or *inaccurate* line search
  - Armijo, Goldstein and Wolfe conditions



## Conceptual Background of Multivariate Optimization

**Convergence of algorithms:** notions of *guarantee* and *speed*

Global convergence: the ability of an algorithm to *approach* and converge to an optimal solution for an *arbitrary* problem, starting from an *arbitrary* point

- Practically, a sequence (or even subsequence) of monotonically decreasing errors is enough.

Local convergence: the rate/speed of approach, measured by $p$, where

$$\beta = \lim_{k\to\infty} \frac{\|\mathbf{x}_{k+1} - \mathbf{x}^*\|}{\|\mathbf{x}_k - \mathbf{x}^*\|^p} < \infty$$

- Linear, quadratic and superlinear rates of convergence for $p = 1,\ 2$ and intermediate.
- Comparison among algorithms with linear rates of convergence is by the convergence ratio $\beta$.

## Points to note

- Theory and methods of single-variable optimization
- Optimality criteria in multivariate optimization
- Convexity in optimization
- The quadratic function
- Trust region
- Line search
- Global and local convergence

Necessary Exercises: **1,2,5,7,8**



## Outline





## Direct Methods

Direct search methods using only function values

- Cyclic coordinate search
- Rosenbrock's method
- Hooke-Jeeves pattern search
- Box's complex method
- Nelder and Mead's simplex search
- Powell's conjugate directions method

Useful for functions, for which derivative either does not exist at all points in the domain or is computationally costly to evaluate.

*Note:* When derivatives are easily available, gradient-based algorithms appear as mainstream methods.



## Direct Methods

**Nelder and Mead's simplex method**

Simplex in $n$-dimensional space: polytope formed by $n+1$ vertices

Nelder and Mead's method iterates over simplices that are non-degenerate (i.e. enclosing non-zero hypervolume).

First, $n+1$ suitable points are selected for the starting simplex.

Among vertices of the current simplex, identify the worst point $\mathbf{x}_w$, the best point $\mathbf{x}_b$ and the second worst point $\mathbf{x}_s$.

*Need to replace $\mathbf{x}_w$ with a good point.*

Centre of gravity of the face *not* containing $\mathbf{x}_w$:

$$\mathbf{x}_c = \frac{1}{n}\sum_{i=1, i\neq w}^{n+1} \mathbf{x}_i$$

Reflect $\mathbf{x}_w$ with respect to $\mathbf{x}_c$ as $\mathbf{x}_r = 2\mathbf{x}_c - \mathbf{x}_w$. Consider options.

# Direct Methods



Default $\mathbf{x}_{new} = \mathbf{x}_r$.
Revision possibilities:

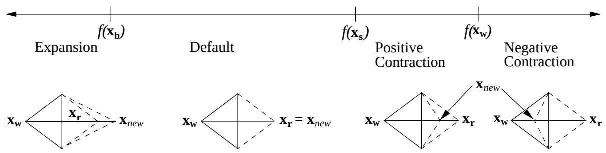



Figure: Nelder and Mead's simplex method

1. For $f(\mathbf{x}_r) < f(\mathbf{x}_b)$, expansion:
   $\mathbf{x}_{new} = \mathbf{x}_c + \alpha(\mathbf{x}_c - \mathbf{x}_w), \ \alpha > 1$.
2. For $f(\mathbf{x}_r) \geq f(\mathbf{x}_w)$, negative contraction:
   $\mathbf{x}_{new} = \mathbf{x}_c - \beta(\mathbf{x}_c - \mathbf{x}_w), \ 0 < \beta < 1$.
3. For $f(\mathbf{x}_s) < f(\mathbf{x}_r) < f(\mathbf{x}_w)$, positive contraction:
   $\mathbf{x}_{new} = \mathbf{x}_c + \beta(\mathbf{x}_c - \mathbf{x}_w)$, with $0 < \beta < 1$.

Replace $\mathbf{x}_w$ with $\mathbf{x}_{new}$. Continue with new simplex.



# Steepest Descent (Cauchy) Method



From a point $\mathbf{x}_k$, a move through $\alpha$ units in direction $\mathbf{d}_k$:

$$f(\mathbf{x}_k + \alpha\mathbf{d}_k) = f(\mathbf{x}_k) + \alpha[\mathbf{g}(\mathbf{x}_k)]^T\mathbf{d}_k + \mathcal{O}(\alpha^2)$$

Descent direction $\mathbf{d}_k$: For $\alpha > 0$, $[\mathbf{g}(\mathbf{x}_k)]^T\mathbf{d}_k < 0$

Direction of *steepest descent*: $\mathbf{d}_k = -\mathbf{g}_k$ [ or $\mathbf{d}_k = -\mathbf{g}_k/\|\mathbf{g}_k\|$]

Minimize

$$\phi(\alpha) = f(\mathbf{x}_k + \alpha\mathbf{d}_k).$$

Exact line search:

$$\phi'(\alpha_k) = [\mathbf{g}(\mathbf{x}_k + \alpha_k\mathbf{d}_k)]^T\mathbf{d}_k = 0$$

Search direction tangential to the contour surface at $(\mathbf{x}_k + \alpha_k\mathbf{d}_k)$.

*Note:* Next direction $\mathbf{d}_{k+1} = -\mathbf{g}(\mathbf{x}_{k+1})$ *orthogonal* to $\mathbf{d}_k$



# Steepest Descent (Cauchy) Method



**Steepest descent algorithm**

1. Select a starting point $\mathbf{x}_0$, set $k = 0$ and several parameters: tolerance $\epsilon_G$ on gradient, absolute tolerance $\epsilon_A$ on reduction in function value, relative tolerance $\epsilon_R$ on reduction in function value and maximum number of iterations $M$.
2. If $\|\mathbf{g}_k\| \leq \epsilon_G$, STOP. Else $\mathbf{d}_k = -\mathbf{g}_k/\|\mathbf{g}_k\|$.
3. Line search: Obtain $\alpha_k$ by minimizing $\phi(\alpha) = f(\mathbf{x}_k + \alpha\mathbf{d}_k)$, $\alpha > 0$. Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.
4. If $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| \leq \epsilon_A + \epsilon_R|f(\mathbf{x}_k)|$,STOP. Else $k \leftarrow k + 1$.
5. If $k > M$, STOP. Else go to step 2.

Very good global convergence.

*But, why so many "STOPS"?*



# Steepest Descent (Cauchy) Method



**Analysis on a quadratic function**

For minimizing $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}$, the error function:

$$E(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T\mathbf{A}(\mathbf{x} - \mathbf{x}^*)$$

Convergence ratio: $\frac{E(\mathbf{x}_{k+1})}{E(\mathbf{x}_k)} \leq \left(\frac{\kappa(\mathbf{A})-1}{\kappa(\mathbf{A})+1}\right)^2$

*Local convergence is poor.*

Importance of steepest descent method

- conceptual understanding
- initial iterations in a completely new problem
- spacer steps in other sophisticated methods

Re-scaling of the problem through change of variables?

# Newton's Method



Second order approximation of a function:

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + [\mathbf{g}(\mathbf{x}_k)]^T(\mathbf{x}-\mathbf{x}_k) + \frac{1}{2}(\mathbf{x}-\mathbf{x}_k)^T\mathbf{H}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k)$$

Vanishing of gradient

$$\mathbf{g}(\mathbf{x}) \approx \mathbf{g}(\mathbf{x}_k) + \mathbf{H}(\mathbf{x}_k)(\mathbf{x}-\mathbf{x}_k)$$

gives the iteration formula

$$\mathbf{x}_{k+1} = \mathbf{x}_k - [\mathbf{H}(\mathbf{x}_k)]^{-1}\mathbf{g}(\mathbf{x}_k).$$

Excellent local convergence property!

$$\frac{\|\mathbf{x}_{k+1}-\mathbf{x}^*\|}{\|\mathbf{x}_k-\mathbf{x}^*\|^2} \leq \beta$$

**Caution:** Does not have global convergence.

*If* $\mathbf{H}(\mathbf{x}_k)$ *is positive definite then* $\mathbf{d}_k = -[\mathbf{H}(\mathbf{x}_k)]^{-1}\mathbf{g}(\mathbf{x}_k)$ *is a* descent direction.



# Newton's Method



*Modified Newton's method*

- Replace the Hessian by $\mathbf{F}_k = \mathbf{H}(\mathbf{x}_k) + \gamma I$.
- Replace full Newton's step by a line search.

**Algorithm**

1. Select $\mathbf{x}_0$, tolerance $\epsilon$ and $\delta > 0$. Set $k = 0$.
2. Evaluate $\mathbf{g}_k = \mathbf{g}(\mathbf{x}_k)$ and $\mathbf{H}(\mathbf{x}_k)$.
   Choose $\gamma$, find $\mathbf{F}_k = \mathbf{H}(\mathbf{x}_k) + \gamma I$, solve $\mathbf{F}_k\mathbf{d}_k = -\mathbf{g}_k$ for $\mathbf{d}_k$.
3. Line search: obtain $\alpha_k$ to minimize $\phi(\alpha) = f(\mathbf{x}_k + \alpha\mathbf{d}_k)$.
   Update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.
4. Check convergence: If $|f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k)| < \epsilon$, STOP.
   Else, $k \leftarrow k+1$ and go to step 2.



# Hybrid (Levenberg-Marquardt) Method



**Methods of deflected gradients**

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k[\mathbf{M}_k]\mathbf{g}_k$$

- identity matrix in place of $\mathbf{M}_k$: steepest descent step
- $\mathbf{M}_k = \mathbf{F}_k^{-1}$: step of modified Newton's method
- $\mathbf{M}_k = [\mathbf{H}(\mathbf{x}_k)]^{-1}$ and $\alpha_k = 1$: pure Newton's step

In $\mathbf{M}_k = [\mathbf{H}(\mathbf{x}_k) + \lambda_k I]^{-1}$, tune parameter $\lambda_k$ over iterations.

- Initial value of $\lambda$: large enough to favour steepest descent trend
- Improvement in an iteration: $\lambda$ reduced by a factor
- Increase in function value: step rejected and $\lambda$ increased

Opportunism systematized!

*Note:* Cost of evaluating the Hessian remains a bottleneck.
Useful for problems where Hessian estimates come cheap!



# Least Square Problems



*Linear* least square problem:

$$y(\theta) = x_1\phi_1(\theta) + x_2\phi_2(\theta) + \cdots + x_n\phi_n(\theta)$$

For measured values $y(\theta_i) = y_i$,

$$e_i = \sum_{k=1}^{n} x_k\phi_k(\theta_i) - y_i = [\Phi(\theta_i)]^T\mathbf{x} - y_i.$$

Error vector: $\mathbf{e} = \mathbf{A}\mathbf{x} - \mathbf{y}$

Last square fit:

*Minimize* $E = \frac{1}{2}\sum_i e_i^2 = \frac{1}{2}\mathbf{e}^T\mathbf{e}$

Pseudoinverse solution and its variants

## Least Square Problems



**Nonlinear least square problem**

For model function in the form

$$y(\theta) = f(\theta, \mathbf{x}) = f(\theta, x_1, x_2, \cdots, x_n),$$

square error function

$$E(\mathbf{x}) = \frac{1}{2}\mathbf{e}^T\mathbf{e} = \frac{1}{2}\sum_i e_i^2 = \frac{1}{2}\sum_i [f(\theta_i, \mathbf{x}) - y_i]^2$$

Gradient: $\mathbf{g}(\mathbf{x}) = \nabla E(\mathbf{x}) = \sum_i [f(\theta_i, \mathbf{x}) - y_i]\nabla f(\theta_i, \mathbf{x}) = \mathbf{J}^T\mathbf{e}$

Hessian: $\mathbf{H}(\mathbf{x}) = \frac{\partial^2}{\partial \mathbf{x}^2}E(\mathbf{x}) = \mathbf{J}^T\mathbf{J} + \sum_i e_i \frac{\partial^2}{\partial \mathbf{x}^2} f(\theta_i, \mathbf{x}) \approx \mathbf{J}^T\mathbf{J}$

Combining a modified form $\lambda\ \mathrm{diag}(\mathbf{J}^T\mathbf{J})\ \delta\mathbf{x} = -\mathbf{g}(\mathbf{x})$ of steepest descent formula with Newton's formula,

| Levenberg-Marquardt step: $[\mathbf{J}^T\mathbf{J} + \lambda\ \mathrm{diag}(\mathbf{J}^T\mathbf{J})]\delta\mathbf{x} = -\mathbf{g}(\mathbf{x})$ |
|---|



## Least Square Problems



**Levenberg-Marquardt algorithm**

1. Select $\mathbf{x}_0$, evaluate $E(\mathbf{x}_0)$. Select tolerance $\epsilon$, initial $\lambda$ and its update factor. Set $k = 0$.
2. Evaluate $\mathbf{g}_k$ and $\bar{\mathbf{H}}_k = \mathbf{J}^T\mathbf{J} + \lambda\ \mathrm{diag}(\mathbf{J}^T\mathbf{J})$.
   Solve $\bar{\mathbf{H}}_k\delta\mathbf{x} = -\mathbf{g}_k$. Evaluate $E(\mathbf{x}_k + \delta\mathbf{x})$.
3. If $|E(\mathbf{x}_k + \delta\mathbf{x}) - E(\mathbf{x}_k)| < \epsilon$, STOP.
4. If $E(\mathbf{x}_k + \delta\mathbf{x}) < E(\mathbf{x}_k)$, then decrease $\lambda$,
   update $\mathbf{x}_{k+1} = \mathbf{x}_k + \delta\mathbf{x}$, $k \leftarrow k + 1$.
   Else increase $\lambda$.
5. Go to step 2.

Professional procedure for nonlinear least square problems and also for solving systems of nonlinear equations in the form $\mathbf{h}(\mathbf{x}) = \mathbf{0}$.



## Points to note



- Simplex method of Nelder and Mead
- Steepest descent method with its global convergence
- Newton's method for fast local convergence
- Levenberg-Marquardt method for equation solving and least squares

Necessary Exercises: **1,2,3,4,5,6**



## Outline

## Conjugate Direction Methods

Conjugacy of directions:

*Two vectors $\mathbf{d}_1$ and $\mathbf{d}_2$ are mutually conjugate with respect to a symmetric matrix $\mathbf{A}$, if $\mathbf{d}_1^T\mathbf{A}\mathbf{d}_2 = 0$.*

Linear independence of conjugate directions:

*Conjugate directions with respect to a positive definite matrix are linearly independent.*

**Expanding subspace property:** In $R^n$, with conjugate vectors $\{\mathbf{d}_0, \mathbf{d}_1, \cdots, \mathbf{d}_{n-1}\}$ with respect to symmetric positive definite $\mathbf{A}$, for any $\mathbf{x}_0 \in R^n$, the sequence $\{\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}$ generated as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k, \quad \text{with} \quad \alpha_k = -\frac{\mathbf{g}_k^T\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k},$$

where $\mathbf{g}_k = \mathbf{A}\mathbf{x}_k + \mathbf{b}$, has the property that

*$\mathbf{x}_k$ minimizes $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}$ on the line $\mathbf{x}_{k-1} + \alpha\mathbf{d}_{k-1}$, as well as on the linear variety $\mathbf{x}_0 + \mathcal{B}_k$, where $\mathcal{B}_k$ is the span of $\mathbf{d}_0$, $\mathbf{d}_1$, $\cdots$, $\mathbf{d}_{k-1}$.*



## Conjugate Direction Methods

**Question:** How to find a set of $n$ conjugate directions?

Gram-Schmidt procedure is a poor option!

**Conjugate gradient method**

Starting from $\mathbf{d}_0 = -\mathbf{g}_0$,

$$\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k\mathbf{d}_k$$

Imposing the condition of conjugacy of $\mathbf{d}_{k+1}$ with $\mathbf{d}_k$,

$$\beta_k = \frac{\mathbf{g}_{k+1}^T\mathbf{A}\mathbf{d}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k} = \frac{\mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k)}{\alpha_k\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}$$

*Resulting $\mathbf{d}_{k+1}$ conjugate to all the earlier directions, for a quadratic problem.*



## Conjugate Direction Methods

Using $k$ in place of $k+1$ in the formula for $\mathbf{d}_{k+1}$,

$$\mathbf{d}_k = -\mathbf{g}_k + \beta_{k-1}\mathbf{d}_{k-1}$$

$$\Rightarrow \ \mathbf{g}_k^T\mathbf{d}_k = -\mathbf{g}_k^T\mathbf{g}_k \ \text{ and } \alpha_k = \frac{\mathbf{g}_k^T\mathbf{g}_k}{\mathbf{d}_k^T\mathbf{A}\mathbf{d}_k}$$

Polak-Ribiere formula:

$$\beta_k = \frac{\mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{g}_k^T\mathbf{g}_k}$$

No need to know $\mathbf{A}$!
Further,

$$\mathbf{g}_{k+1}^T\mathbf{d}_k = 0 \ \Rightarrow \ \mathbf{g}_{k+1}^T\mathbf{g}_k = \beta_{k-1}(\mathbf{g}_k^T + \alpha_k\mathbf{d}_k^T\mathbf{A})\mathbf{d}_{k-1} = 0.$$

Fletcher-Reeves formula:

$$\beta_k = \frac{\mathbf{g}_{k+1}^T\mathbf{g}_{k+1}}{\mathbf{g}_k^T\mathbf{g}_k}$$



## Conjugate Direction Methods

**Extension to general (non-quadratic) functions**

- Varying Hessian $\mathbf{A}$: determine the step size by line search.
- After $n$ steps, minimum not attained.
  But, $\mathbf{g}_k^T\mathbf{d}_k = -\mathbf{g}_k^T\mathbf{g}_k$ implies guaranteed descent.
  Globally convergent, with superlinear rate of convergence.
- *What to do* after $n$ steps? Restart or continue?

**Algorithm**

1. Select $\mathbf{x}_0$ and tolerances $\epsilon_G$, $\epsilon_D$. Evaluate $\mathbf{g}_0 = \nabla f(\mathbf{x}_0)$.
2. Set $k = 0$ and $\mathbf{d}_k = -\mathbf{g}_k$.
3. Line search: find $\alpha_k$; update $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k\mathbf{d}_k$.
4. Evaluate $\mathbf{g}_{k+1} = \nabla f(\mathbf{x}_{k+1})$. If $\|\mathbf{g}_{k+1}\| \leq \epsilon_G$, STOP.
5. Find $\beta_k = \frac{\mathbf{g}_{k+1}^T(\mathbf{g}_{k+1} - \mathbf{g}_k)}{\mathbf{g}_k^T\mathbf{g}_k}$ (Polak-Ribiere)
   or $\beta_k = \frac{\mathbf{g}_{k+1}^T\mathbf{g}_{k+1}}{\mathbf{g}_k^T\mathbf{g}_k}$ (Fletcher-Reeves).
   Obtain $\mathbf{d}_{k+1} = -\mathbf{g}_{k+1} + \beta_k\mathbf{d}_k$.
6. If $1 - \left|\frac{\mathbf{d}_k^T\mathbf{d}_{k+1}}{\|\mathbf{d}_k\| \ \|\mathbf{d}_{k+1}\|}\right| < \epsilon_D$, reset $\mathbf{g}_0 = \mathbf{g}_{k+1}$and go to step 2.
   Else, $k \leftarrow k+1$ and go to step 3.

## Conjugate Direction Methods

**Powell's conjugate direction method**

For $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}$, suppose

$$\mathbf{x}_1 = \mathbf{x}_A + \alpha_1\mathbf{d} \text{ such that } \mathbf{d}^T\mathbf{g}_1 = 0 \text{ and}$$
$$\mathbf{x}_2 = \mathbf{x}_B + \alpha_2\mathbf{d} \text{ such that } \mathbf{d}^T\mathbf{g}_2 = 0.$$

Then, $\mathbf{d}^T\mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) = \mathbf{d}^T(\mathbf{g}_2 - \mathbf{g}_1) = 0$.

**Parallel subspace property:** *In $R^n$, consider two parallel linear varieties $\mathcal{S}_1 = \mathbf{v}_1 + \mathcal{B}_k$ and $\mathcal{S}_2 = \mathbf{v}_2 + \mathcal{B}_k$, with $\mathcal{B}_k = \{\mathbf{d}_1, \mathbf{d}_2, \cdots, \mathbf{d}_k\}$, $k < n$.*
*If $\mathbf{x}_1$ and $\mathbf{x}_2$ minimize $q(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}^T\mathbf{x}$ on $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively,*

*then $\mathbf{x}_2 - \mathbf{x}_1$ is conjugate to $\mathbf{d}_1$, $\mathbf{d}_2$, $\cdots$, $\mathbf{d}_k$.*

Assumptions imply $\mathbf{g}_1, \mathbf{g}_2 \perp \mathcal{B}_k$ and hence

$$(\mathbf{g}_2 - \mathbf{g}_1) \perp \mathcal{B}_k \Rightarrow \mathbf{d}_i^T\mathbf{A}(\mathbf{x}_2 - \mathbf{x}_1) = \mathbf{d}_i^T(\mathbf{g}_2 - \mathbf{g}_1) = 0 \quad \text{for} \quad i = 1, 2, \cdots, k.$$



## Conjugate Direction Methods

**Algoithm**

1. Select $\mathbf{x}_0$, $\epsilon$ and a set of $n$ linearly independent (preferably normalized) directions $\mathbf{d}_1$, $\mathbf{d}_2$, $\cdots$, $\mathbf{d}_n$; possibly $\mathbf{d}_i = \mathbf{e}_i$.
2. Line search along $\mathbf{d}_n$ and update $\mathbf{x}_1 = \mathbf{x}_0 + \alpha\mathbf{d}_n$; set $k = 1$.
3. Line searches along $\mathbf{d}_1$, $\mathbf{d}_2$, $\cdots$, $\mathbf{d}_n$ in sequence to obtain $\mathbf{z} = \mathbf{x}_k + \sum_{j=1}^{n}\alpha_j\mathbf{d}_j$.
4. New conjugate direction $\mathbf{d} = \mathbf{z} - \mathbf{x}_k$. If $\|\mathbf{d}\| < \epsilon$, STOP.
5. Reassign directions $\mathbf{d}_j \leftarrow \mathbf{d}_{j+1}$ for $j = 1, 2, \cdots, (n-1)$ and $\mathbf{d}_n = \mathbf{d}/\|\mathbf{d}\|$.
   (Old $\mathbf{d}_1$ gets discarded at this step.)
6. Line search and update $\mathbf{x}_{k+1} = \mathbf{z} + \alpha\mathbf{d}_n$; set $k \leftarrow k+1$ and go to step 3.



## Conjugate Direction Methods

- $\mathbf{x}_0$-$\mathbf{x}_1$ and $b$-$\mathbf{z}_1$: $\mathbf{x}_1$-$\mathbf{z}_1$ is conjugate to $b$-$\mathbf{z}_1$.
- $b$-$\mathbf{z}_1$-$\mathbf{x}_2$ and $c$-$d$-$\mathbf{z}_2$: $c$-$d$, $d$-$\mathbf{z}_2$ and $\mathbf{x}_2$-$\mathbf{z}_2$ are mutually conjugate.

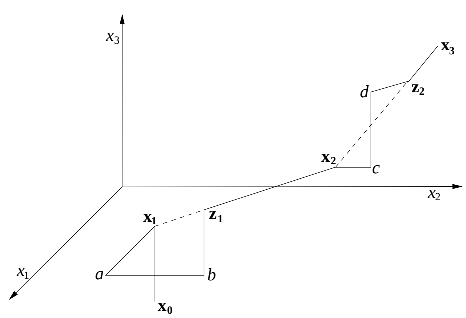



Figure: Schematic of Powell's conjugate direction method

*Performance of Powell's method approaches that of the conjugate gradient method!*



## Quasi-Newton Methods

**Variable metric methods**

*attempt to* construct *the inverse Hessian* $\mathbf{B}_k$.

$$\mathbf{p}_k = \mathbf{x}_{k+1} - \mathbf{x}_k \quad \text{and} \quad \mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k \quad \Rightarrow \quad \mathbf{q}_k \approx \mathbf{H}\mathbf{p}_k$$

With $n$ such steps, $\mathbf{B} = \mathbf{P}\mathbf{Q}^{-1}$: update and construct $\mathbf{B}_k \approx \mathbf{H}^{-1}$.
Rank one correction: $\mathbf{B}_{k+1} = \mathbf{B}_k + a_k\mathbf{z}_k\mathbf{z}_k^T$?
Rank two correction:

$$\mathbf{B}_{k+1} = \mathbf{B}_k + a_k\mathbf{z}_k\mathbf{z}_k^T + b_k\mathbf{w}_k\mathbf{w}_k^T$$

Davidon-Fletcher-Powell (DFP) method

Select $\mathbf{x}_0$, tolerance $\epsilon$ and $\mathbf{B}_0 = \mathbf{I}_n$. For $k = 0, 1, 2, \cdots$,

- $\mathbf{d}_k = -\mathbf{B}_k\mathbf{g}_k$.
- Line search for $\alpha_k$; update $\mathbf{p}_k = \alpha_k\mathbf{d}_k$, $\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{p}_k$, $\mathbf{q}_k = \mathbf{g}_{k+1} - \mathbf{g}_k$.
- If $\|\mathbf{p}_k\| < \epsilon$ or $\|\mathbf{q}_k\| < \epsilon$, STOP.
- Rank two correction: $\mathbf{B}_{k+1}^{DFP} = \mathbf{B}_k + \frac{\mathbf{p}_k\mathbf{p}_k^T}{\mathbf{p}_k^T\mathbf{q}_k} - \frac{\mathbf{B}_k\mathbf{q}_k\mathbf{q}_k^T\mathbf{B}_k}{\mathbf{q}_k^T\mathbf{B}_k\mathbf{q}_k}$.

## Quasi-Newton Methods

Properties of DFP iterations:

1. If $\mathbf{B}_k$ is symmetric and positive definite, then so is $\mathbf{B}_{k+1}$.
2. For quadratic function with positive definite Hessian $\mathbf{H}$,

$$\mathbf{p}_i^T \mathbf{H} \mathbf{p}_j = 0 \quad \text{for} \quad 0 \le i < j \le k,$$
$$\text{and} \quad \mathbf{B}_{k+1} \mathbf{H} \mathbf{p}_i = \mathbf{p}_i \quad \text{for} \quad 0 \le i \le k.$$

Implications:

1. Positive definiteness of inverse Hessian estimate is never lost.
2. Successive search directions are conjugate directions.
3. With $\mathbf{B}_0 = \mathbf{I}$, the algorithm is a conjugate gradient method.
4. For a quadratic problem, the inverse Hessian gets completely constructed after $n$ steps.

*Variants: Broyden-Fletcher-Goldfarb-Shanno (BFGS) method and the Broyden family of methods*



## Closure

**Table 23.1:** Summary of performance of optimization methods

| | **Cauchy** (Steepest Descent) | **Newton** | **Levenberg-Marquardt** (Hybrid) (Deflected Gradient) | **DFP/BFGS** (Quasi-Newton) (Variable Metric) | **FR/PR** (Conjugate Gradient) | **Powell** (Direction Set) |
|---|---|---|---|---|---|---|
| **For Quadratic Problems:** | | | | | | |
| Convergence steps | $N$ Indefinite | 1 | $N$ Unknown | $n$ | $n$ | $n^2$ |
| Evaluations | $Nf$ $Ng$ | $2f$ $2g$ $1H$ | $Nf$ $Ng$ $NH$ | $(n+1)f$ $(n+1)g$ | $(n+1)f$ $(n+1)g$ | $n^2 f$ |
| Equivalent function evaluations | $N(2n+1)$ | $2n^2+2n+1$ | $N(2n^2+1)$ | $2n^2+3n+1$ | $2n^2+3n+1$ | $n^2$ |
| Line searches | $N$ | 0 | $N$ or 0 | $n$ | $n$ | $n^2$ |
| Storage | Vector | Matrix | Matrix | Matrix | Vector | Matrix |
| Performance in general problems | Slow | Risky | Costly | Flexible | Good | Okay |
| Practically good for | Unknown start-up | Good functions | NL Eqn. systems NL least squares | Bad functions | Large problems | Small problems |



## Points to note

- Conjugate directions and the expanding subspace property
- Conjugate gradient method
- Powell-Smith direction set method
- The quasi-Newton concept in professional optimization

Necessary Exercises: **1,2,3**



## Outline

# Constraints



Constrained optimization problem:

$$\begin{array}{rlll} \text{Minimize} & f(\mathbf{x}) & & \\ \text{subject to} & g_i(\mathbf{x}) \le 0 & \text{for } i = 1, 2, \cdots, l, & \text{or } \mathbf{g}(\mathbf{x}) \le \mathbf{0}; \\ \text{and} & h_j(\mathbf{x}) = 0 & \text{for } j = 1, 2, \cdots, m, & \text{or } \mathbf{h}(\mathbf{x}) = \mathbf{0}. \end{array}$$

Conceptually, "minimize $f(\mathbf{x})$, $\mathbf{x} \in \Omega$".
Equality constraints reduce the domain to a surface or a manifold, possessing a **tangent plane** at every point.
Gradient of the vector function $\mathbf{h}(\mathbf{x})$:

$$\nabla \mathbf{h}(\mathbf{x}) \equiv [\nabla h_1(\mathbf{x}) \;\; \nabla h_2(\mathbf{x}) \;\; \cdots \;\; \nabla h_m(\mathbf{x})] \equiv \begin{bmatrix} \frac{\partial \mathbf{h}^T}{\partial x_1} \\ \frac{\partial \mathbf{h}^T}{\partial x_2} \\ \vdots \\ \frac{\partial \mathbf{h}^T}{\partial x_n} \end{bmatrix},$$

related to the usual Jacobian as $\mathbf{J}_h(\mathbf{x}) = \frac{\partial \mathbf{h}}{\partial \mathbf{x}} = [\nabla \mathbf{h}(\mathbf{x})]^T$.



# Constraints



**Constraint qualification**

$\nabla h_1(\mathbf{x})$, $\nabla h_2(\mathbf{x})$ etc are linearly independent, i.e. $\nabla \mathbf{h}(\mathbf{x})$ is full-rank.

If a feasible point $\mathbf{x}_0$, with $\mathbf{h}(\mathbf{x}_0) = \mathbf{0}$, satisfies the constraint qualification condition, we call it a **regular point**.

At a regular feasible point $\mathbf{x}_0$, *tangent plane*

$$\mathcal{M} = \{\mathbf{y} : [\nabla \mathbf{h}(\mathbf{x}_0)]^T \mathbf{y} = \mathbf{0}\}$$

gives the collection of feasible directions.

Equality constraints reduce the *dimension* of the problem.

Variable elimination?



# Constraints



Active inequality constraints $g_i(\mathbf{x}_0) = 0$:

*included among* $h_j(\mathbf{x}_0)$

for the tangent plane.

*Cone* of feasible directions:

$$[\nabla \mathbf{h}(\mathbf{x}_0)]^T \mathbf{d} = \mathbf{0} \quad \text{and} \quad [\nabla g_i(\mathbf{x}_0)]^T \mathbf{d} \le 0 \;\; \text{for } i \in I$$

where $I$ is the set of indices of active inequality constraints.

Handling inequality constraints:

- **Active set strategy** maintains a list of active constraints, keeps checking at every step for a change of scenario and updates the list by inclusions and exclusions.
- **Slack variable strategy** replaces all the inequality constraints by equality constraints as $g_i(\mathbf{x}) + x_{n+i} = 0$ with the inclusion of non-negative slack variables ($x_{n+i}$).



# Optimality Criteria



Suppose $\mathbf{x}^*$ is a regular point with

- active inequality constraints: $\mathbf{g}^{(a)}(\mathbf{x}) \le \mathbf{0}$
- inactive constraints: $\mathbf{g}^{(i)}(\mathbf{x}) \le \mathbf{0}$

Columns of $\nabla \mathbf{h}(\mathbf{x}^*)$ and $\nabla \mathbf{g}^{(a)}(\mathbf{x}^*)$: basis for orthogonal complement of the tangent plane

Basis of the tangent plane: $\mathbf{D} = [\mathbf{d}_1 \quad \mathbf{d}_2 \quad \cdots \quad \mathbf{d}_k]$

Then, $[\mathbf{D} \quad \nabla \mathbf{h}(\mathbf{x}^*) \quad \nabla \mathbf{g}^{(a)}(\mathbf{x}^*)]$: basis of $R^n$

Now, $-\nabla f(\mathbf{x}^*)$ is a vector in $R^n$.

$$-\nabla f(\mathbf{x}^*) = [\mathbf{D} \quad \nabla \mathbf{h}(\mathbf{x}^*) \quad \nabla \mathbf{g}^{(a)}(\mathbf{x}^*)] \begin{bmatrix} \mathbf{z} \\ \boldsymbol{\lambda} \\ \boldsymbol{\mu}^{(a)} \end{bmatrix}$$

with unique $\mathbf{z}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}^{(a)}$ for a given $\nabla f(\mathbf{x}^*)$.

What can you say if $\mathbf{x}^*$ is a solution to the NLP problem?

## Optimality Criteria



**Components of $\nabla f(\mathbf{x}^*)$ in the tangent plane must be zero.**

$$\mathbf{z} = \mathbf{0} \quad \Rightarrow \qquad -\nabla f(\mathbf{x}^*) = [\nabla \mathbf{h}(\mathbf{x}^*)]\boldsymbol{\lambda} + [\nabla \mathbf{g}^{(a)}(\mathbf{x}^*)]\boldsymbol{\mu}^{(a)}$$

For inactive constraints, insisting on $\boldsymbol{\mu}^{(i)} = \mathbf{0}$,

$$-\nabla f(\mathbf{x}^*) = [\nabla \mathbf{h}(\mathbf{x}^*)]\boldsymbol{\lambda} + [\nabla \mathbf{g}^{(a)}(\mathbf{x}^*) \quad \nabla \mathbf{g}^{(i)}(\mathbf{x}^*)] \begin{bmatrix} \boldsymbol{\mu}^{(a)} \\ \boldsymbol{\mu}^{(i)} \end{bmatrix},$$

or

$$\boxed{\nabla f(\mathbf{x}^*) + [\nabla \mathbf{h}(\mathbf{x}^*)]\boldsymbol{\lambda} + [\nabla \mathbf{g}(\mathbf{x}^*)]\boldsymbol{\mu} = \mathbf{0}}$$

where $\mathbf{g}(\mathbf{x}) = \begin{bmatrix} \mathbf{g}^{(a)}(\mathbf{x}) \\ \mathbf{g}^{(i)}(\mathbf{x}) \end{bmatrix}$ and $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}^{(a)} \\ \boldsymbol{\mu}^{(i)} \end{bmatrix}$.

Notice: $\mathbf{g}^{(a)}(\mathbf{x}^*) = \mathbf{0}$ and $\boldsymbol{\mu}^{(i)} = \mathbf{0} \quad \Rightarrow \quad \mu_i g_i(\mathbf{x}^*) = 0 \;\; \forall\, i$, or

$$\boxed{\boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}^*) = 0.}$$

Now, components in $\mathbf{g}(\mathbf{x})$ are free to appear in any order.



## Optimality Criteria



Finally, what about the *feasible* directions in the cone?

**Answer:** Negative gradient $-\nabla f(\mathbf{x}^*)$ can have no component towards *decreasing* $g_i^{(a)}(\mathbf{x})$, i.e. $\mu_i^{(a)} \geq 0, \; \forall\, i$.

Combining it with $\mu_i^{(i)} = 0$, $\quad \boxed{\boldsymbol{\mu} \geq \mathbf{0}.}$

**First order necessary conditions** or **Karusch-Kuhn-Tucker (KKT) conditions:** If $\mathbf{x}^*$ is a regular point of the constraints and a solution to the NLP problem, then there exist Lagrange multiplier vectors, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, such that

$$\begin{array}{lr} \text{Optimality:} & \nabla f(\mathbf{x}^*) + [\nabla \mathbf{h}(\mathbf{x}^*)]\boldsymbol{\lambda} + [\nabla \mathbf{g}(\mathbf{x}^*)]\boldsymbol{\mu} = \mathbf{0}, \quad \boldsymbol{\mu} \geq \mathbf{0}; \\ \text{Feasibility:} & \mathbf{h}(\mathbf{x}^*) = \mathbf{0}, \quad \mathbf{g}(\mathbf{x}^*) \leq \mathbf{0}; \\ \text{Complementarity:} & \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x}^*) = \mathbf{0}. \end{array}$$

**Convex programming problem:** Convex objective function $f(\mathbf{x})$ and convex domain (convex $g_i(\mathbf{x})$ and linear $h_j(\mathbf{x})$):

KKT conditions are sufficient as well!



## Optimality Criteria



**Lagrangian** function:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x})$$

Necessary conditions for a *stationary point* of the Lagrangian:

$$\nabla_x L = \mathbf{0}, \qquad \nabla_\lambda L = \mathbf{0}$$

**Second order conditions**

Consider curve $\mathbf{z}(t)$ in the tangent plane with $\mathbf{z}(0) = \mathbf{x}^*$.

$$\begin{aligned} \left.\frac{d^2}{dt^2} f(\mathbf{z}(t))\right|_{t=0} &= \left.\frac{d}{dt}[\nabla f(\mathbf{z}(t))^T \dot{\mathbf{z}}(t)]\right|_{t=0} \\ &= \dot{\mathbf{z}}(0)^T \mathbf{H}(\mathbf{x}^*)\dot{\mathbf{z}}(0) + [\nabla f(\mathbf{x}^*)]^T \ddot{\mathbf{z}}(0) \geq 0 \end{aligned}$$

Similarly, from $h_j(\mathbf{z}(t)) = 0$,

$$\dot{\mathbf{z}}(0)^T \mathbf{H}_{h_j}(\mathbf{x}^*)\dot{\mathbf{z}}(0) + [\nabla h_j(\mathbf{x}^*)]^T \ddot{\mathbf{z}}(0) = 0.$$



## Optimality Criteria



Including contributions from all *active* constraints,

$$\left.\frac{d^2}{dt^2} f(\mathbf{z}(t))\right|_{t=0} = \dot{\mathbf{z}}(0)^T \mathbf{H}_L(\mathbf{x}^*)\dot{\mathbf{z}}(0) + [\nabla_x L(\mathbf{x}^*, \boldsymbol{\lambda}, \boldsymbol{\mu})]^T \ddot{\mathbf{z}}(0) \geq 0,$$

where $\mathbf{H}_L(\mathbf{x}) = \frac{\partial^2 L}{\partial \mathbf{x}^2} = \mathbf{H}(\mathbf{x}) + \sum_j \lambda_j \mathbf{H}_{h_j}(\mathbf{x}) + \sum_i \mu_i \mathbf{H}_{g_i}(\mathbf{x})$.

First order necessary condition makes the second term vanish!

Second order necessary condition:

> *The Hessian matrix of the Lagrangian function is positive semi-definite on the tangent plane $\mathcal{M}$.*

Sufficient condition: $\nabla_x L = \mathbf{0}$ and $\mathbf{H}_L(\mathbf{x})$ positive definite on $\mathcal{M}$.

**Restriction** of the mapping $\mathbf{H}_L(\mathbf{x}^*) : R^n \to R^n$ on subspace $\mathcal{M}$?

## Optimality Criteria



Take $\mathbf{y} \in \mathcal{M}$, operate $\mathbf{H}_L(\mathbf{x}^*)$ on it, project the image back to $\mathcal{M}$.

*Restricted mapping* $\mathbf{L}_M : \mathcal{M} \to \mathcal{M}$

**Question:** Matrix representation for $\mathbf{L}_M$ of size $(n-m) \times (n-m)$?

Select local orthonormal basis $\mathbf{D} \in R^{n \times (n-m)}$ for $\mathcal{M}$.

For arbitrary $\mathbf{z} \in R^{n-m}$, map $\mathbf{y} = \mathbf{Dz} \in R^n$ as $\mathbf{H}_L \mathbf{y} = \mathbf{H}_L \mathbf{Dz}$.

Its component along $\mathbf{d}_i$: $\mathbf{d}_i^T \mathbf{H}_L \mathbf{Dz}$

Hence, projection back on $\mathcal{M}$:

$$\mathbf{L}_M \mathbf{z} = \mathbf{D}^T \mathbf{H}_L \mathbf{Dz},$$

The $(n-m) \times (n-m)$ matrix $\mathbf{L}_M = \mathbf{D}^T \mathbf{H}_L \mathbf{D}$: the restriction!

Second order necessary/sufficient condition: $\mathbf{L}_M$ p.s.d./p.d.



## Sensitivity



Suppose original objective and constraint functions as

$f(\mathbf{x}, \mathbf{p})$, $\mathbf{g}(\mathbf{x}, \mathbf{p})$ *and* $\mathbf{h}(\mathbf{x}, \mathbf{p})$

By choosing parameters $(\mathbf{p})$, we arrive at $\mathbf{x}^*$. Call it $\mathbf{x}^*(\mathbf{p})$.

**Question:** How does $f(\mathbf{x}^*(\mathbf{p}), \mathbf{p})$ depend on $\mathbf{p}$?

Total gradients

$$\begin{aligned} \bar{\nabla}_p f(\mathbf{x}^*(\mathbf{p}), \mathbf{p}) &= \nabla_p \mathbf{x}^*(\mathbf{p}) \nabla_x f(\mathbf{x}^*, \mathbf{p}) + \nabla_p f(\mathbf{x}^*, \mathbf{p}), \\ \bar{\nabla}_p \mathbf{h}(\mathbf{x}^*(\mathbf{p}), \mathbf{p}) &= \nabla_p \mathbf{x}^*(\mathbf{p}) \nabla_x \mathbf{h}(\mathbf{x}^*, \mathbf{p}) + \nabla_p \mathbf{h}(\mathbf{x}^*, \mathbf{p}) = \mathbf{0}, \end{aligned}$$

and similarly for $\mathbf{g}(\mathbf{x}^*(\mathbf{p}), \mathbf{p})$.

In view of $\nabla_x L = 0$, from KKT conditions,

$$\bar{\nabla}_p f(\mathbf{x}^*(\mathbf{p}), \mathbf{p}) = \nabla_p f(\mathbf{x}^*, \mathbf{p}) + [\nabla_p \mathbf{h}(\mathbf{x}^*, \mathbf{p})]\boldsymbol{\lambda} + [\nabla_p \mathbf{g}(\mathbf{x}^*, \mathbf{p})]\boldsymbol{\mu}$$



## Sensitivity



**Sensitivity to constraints**

In particular, in a revised problem, with $\mathbf{h}(\mathbf{x}) = \mathbf{c}$ and $\mathbf{g}(\mathbf{x}) \leq \mathbf{d}$, using $\mathbf{p} = \mathbf{c}$,

$$\nabla_p f(\mathbf{x}^*, \mathbf{p}) = \mathbf{0}, \; \nabla_p \mathbf{h}(\mathbf{x}^*, \mathbf{p}) = -\mathbf{I} \text{ and } \nabla_p \mathbf{g}(\mathbf{x}^*, \mathbf{p}) = \mathbf{0}.$$

$$\boxed{\bar{\nabla}_c f(\mathbf{x}^*(\mathbf{p}), \mathbf{p}) = -\boldsymbol{\lambda}}$$

Similarly, using $\mathbf{p} = \mathbf{d}$, we get $\boxed{\bar{\nabla}_d f(\mathbf{x}^*(\mathbf{p}), \mathbf{p}) = -\boldsymbol{\mu}.}$

Lagrange multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ signify costs of *pulling* the minimum point in order to satisfy the constraints!

- Equality constraint: both sides infeasible, sign of $\lambda_j$ identifies one side or the other of the hypersurface.
- Inequality constraint: one side is feasible, no cost of pulling from that side, so $\mu_i \geq 0$.



## Duality*



**Dual problem:**

Reformulation of a problem in terms of the Lagrange multipliers.
Suppose $\mathbf{x}^*$ as a local minimum for the problem

*Minimize* $f(\mathbf{x})$ *subject to* $\mathbf{h}(\mathbf{x}) = \mathbf{0}$,

with Lagrange multiplier (vector) $\boldsymbol{\lambda}^*$.

$$\nabla f(\mathbf{x}^*) + [\nabla \mathbf{h}(\mathbf{x}^*)]\boldsymbol{\lambda}^* = \mathbf{0}$$

If $\mathbf{H}_L(\mathbf{x}^*)$ is positive definite (assumption of local duality), then $\mathbf{x}^*$ is also a local minimum of

$$\bar{f}(\mathbf{x}) = f(\mathbf{x}) + \boldsymbol{\lambda}^{*T} \mathbf{h}(\mathbf{x}).$$

If we vary $\boldsymbol{\lambda}$ around $\boldsymbol{\lambda}^*$, the minimizer of

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})$$

varies continuously with $\boldsymbol{\lambda}$.

# Duality*

In the neighbourhood of $\boldsymbol{\lambda}^*$, define the dual function

$$\Phi(\boldsymbol{\lambda}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}) = \min_{\mathbf{x}} [f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x})].$$

*For a pair* $\{\mathbf{x}, \boldsymbol{\lambda}\}$, *the dual solution is* feasible *if and only if the primal solution is optimal.*

Define $\mathbf{x}(\boldsymbol{\lambda})$ as the local minimizer of $L(\mathbf{x}, \boldsymbol{\lambda})$.

$$\Phi(\boldsymbol{\lambda}) = L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = f(\mathbf{x}(\boldsymbol{\lambda})) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$$

First derivative:

$$\nabla \Phi(\boldsymbol{\lambda}) = \nabla_{\lambda} \mathbf{x}(\boldsymbol{\lambda}) \nabla_x L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) + \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda})) = \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$$

*For a pair* $\{\mathbf{x}, \boldsymbol{\lambda}\}$, *the dual solution is* optimal *if and only if the primal solution is feasible.*



# Duality*

Hessian of the dual function:

$$\mathbf{H}_{\phi}(\boldsymbol{\lambda}) = \nabla_{\lambda} \mathbf{x}(\boldsymbol{\lambda}) \nabla_x \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))$$

Differentiating $\nabla_x L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) = \mathbf{0}$, we have

$$\nabla_{\lambda} \mathbf{x}(\boldsymbol{\lambda}) \mathbf{H}_L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda}) + [\nabla_x \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))]^T = \mathbf{0}.$$

Solving for $\nabla_{\lambda} \mathbf{x}(\boldsymbol{\lambda})$ and substituting,

$$\mathbf{H}_{\phi}(\boldsymbol{\lambda}) = -[\nabla_x \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda}))]^T [\mathbf{H}_L(\mathbf{x}(\boldsymbol{\lambda}), \boldsymbol{\lambda})]^{-1} \nabla_x \mathbf{h}(\mathbf{x}(\boldsymbol{\lambda})),$$

negative definite!

At $\boldsymbol{\lambda}^*$, $\mathbf{x}(\boldsymbol{\lambda}^*) = \mathbf{x}^*$, $\nabla \Phi(\boldsymbol{\lambda}^*) = \mathbf{h}(\mathbf{x}^*) = \mathbf{0}$, $\mathbf{H}_{\phi}(\boldsymbol{\lambda}^*)$ is negative definite and the dual function is maximized.

$$\boxed{\Phi(\boldsymbol{\lambda}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = f(\mathbf{x}^*)}$$



# Duality*

Consolidation (including *all* constraints)

- Assuming local convexity, the dual function:

$$\Phi(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x}} [f(\mathbf{x}) + \boldsymbol{\lambda}^T \mathbf{h}(\mathbf{x}) + \boldsymbol{\mu}^T \mathbf{g}(\mathbf{x})].$$

- Constraints on the dual: $\nabla_x L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \mathbf{0}$, optimality of the primal.
- Corresponding to inequality constraints of the primal problem, non-negative variables $\boldsymbol{\mu}$ in the dual problem.
- First order necessary conditons for the dual optimality: equivalent to the feasibility of the primal problem.
- The dual function is concave *globally*!
- Under suitable conditions, $\Phi(\boldsymbol{\lambda}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = f(\mathbf{x}^*)$.
- The Lagrangian $L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$ has a *saddle point* in the combined space of primal and dual variables: positive curvature along $\mathbf{x}$ directions and negative curvature along $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ directions.



# Structure of Methods: An Overview*

For a problem of $n$ variables, with $m$ active constraints,
nature and dimension of working spaces

**Penalty methods ($R^n$):** Minimize the *penalized function*

$$q(c, \mathbf{x}) = f(\mathbf{x}) + cP(\mathbf{x}).$$

Example: $P(\mathbf{x}) = \frac{1}{2}\|\mathbf{h}(\mathbf{x})\|^2 + \frac{1}{2}[\max(\mathbf{0}, \mathbf{g}(\mathbf{x}))]^2$.

**Primal methods ($R^{n-m}$):** Work only in feasible domain, restricting steps to the tangent plane.
Example: Gradient projection method.

**Dual methods ($R^m$):** Transform the problem to the space of Lagrange multipliers and maximize the dual.
Example: Augmented Lagrangian method.

**Lagrange methods ($R^{m+n}$):** Solve equations appearing in the KKT conditions directly.
Example: Sequential quadratic programming.

# Points to note

- Constraint qualification
- KKT conditions
- Second order conditions
- Basic ideas for solution strategy

Necessary Exercises: **1,2,3,4,5,6**



# Outline





# Linear Programming

**Standard form** of an LP problem:

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}, \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b},\ \mathbf{x} \geq \mathbf{0}; \qquad \text{with} \quad \mathbf{b} \geq \mathbf{0}. \end{array}$$

**Preprocessing** to cast a problem to the standard form

- Maximization: Minimize the negative function.
- Variables of unrestricted sign: Use two variables.
- Inequality constraints: Use slack/surplus variables.
- Negative RHS: Multiply with $-1$.

**Geometry** of an LP problem

- Infinite domain: does a minimum *exist*?
- Finite convex polytope: *existence* guaranteed
- Operating with vertices sufficient as a strategy
- Extension with slack/surplus variables: original solution space a *subspace* in the extented space, $\mathbf{x} \geq \mathbf{0}$ marking the domain
- Essence of the non-negativity condition of variables



# Linear Programming

**The simplex method**

Suppose $\mathbf{x} \in R^N$, $\mathbf{b} \in R^M$ and $\mathbf{A} \in R^{M \times N}$ full-rank, with $M < N$.

$$\mathbf{I}_M \mathbf{x}_B + \mathbf{A}' \mathbf{x}_{NB} = \mathbf{b}'$$

Basic and non-basic variables: $\mathbf{x}_B \in R^M$ and $\mathbf{x}_{NB} \in R^{N-M}$

Basic feasible solution: $\mathbf{x}_B = \mathbf{b}' \geq \mathbf{0}$ and $\mathbf{x}_{NB} = \mathbf{0}$

At every iteration,

- selection of a non-basic variable to enter the basis
  - edge of travel selected based on maximum rate of descent
  - no qualifier: current vertex is optimal
- selection of a basic variable to leave the basis
  - based on the first constraint becoming active along the edge
  - no constraint ahead: function is unbounded
- elementary row operations: new basic feasible solution

Two-phase method: Inclusion of a pre-processing phase with artificial variables to develop a *basic feasible solution*

# Linear Programming

**General perspective**

LP problem:

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x},\mathbf{y}) = \mathbf{c}_1^T\mathbf{x} + \mathbf{c}_2^T\mathbf{y}; \\ \text{subject to} & \mathbf{A}_{11}\mathbf{x} + \mathbf{A}_{12}\mathbf{y} = \mathbf{b}_1, \quad \mathbf{A}_{21}\mathbf{x} + \mathbf{A}_{22}\mathbf{y} \leq \mathbf{b}_2, \quad \mathbf{y} \geq \mathbf{0}. \end{array}$$

Lagrangian:

$$L(\mathbf{x},\mathbf{y},\boldsymbol{\lambda},\boldsymbol{\mu},\boldsymbol{\nu}) = \mathbf{c}_1^T\mathbf{x} + \mathbf{c}_2^T\mathbf{y} + \boldsymbol{\lambda}^T(\mathbf{A}_{11}\mathbf{x} + \mathbf{A}_{12}\mathbf{y} - \mathbf{b}_1) + \boldsymbol{\mu}^T(\mathbf{A}_{21}\mathbf{x} + \mathbf{A}_{22}\mathbf{y} - \mathbf{b}_2) - \boldsymbol{\nu}^T\mathbf{y}$$

Optimality conditions:

$$\mathbf{c}_1 + \mathbf{A}_{11}^T\boldsymbol{\lambda} + \mathbf{A}_{21}^T\boldsymbol{\mu} = \mathbf{0} \quad \text{and} \quad \boldsymbol{\nu} = \mathbf{c}_2 + \mathbf{A}_{12}^T\boldsymbol{\lambda} + \mathbf{A}_{22}^T\boldsymbol{\mu} \geq \mathbf{0}$$

Substituting back, optimal function value: $f^* = -\boldsymbol{\lambda}^T\mathbf{b}_1 - \boldsymbol{\mu}^T\mathbf{b}_2$

Sensitivity to the constraints: $\frac{\partial f^*}{\partial \mathbf{b}_1} = -\lambda$ and $\frac{\partial f^*}{\partial \mathbf{b}_2} = -\boldsymbol{\mu}$

Dual problem:

$$\begin{array}{ll} \text{maximize} & \Phi(\boldsymbol{\lambda},\boldsymbol{\mu}) = -\mathbf{b}_1^T\boldsymbol{\lambda} - \mathbf{b}_2^T\boldsymbol{\mu}; \\ \text{subject to} & \mathbf{A}_{11}^T\boldsymbol{\lambda} + \mathbf{A}_{21}^T\boldsymbol{\mu} = -\mathbf{c}_1, \quad \mathbf{A}_{12}^T\boldsymbol{\lambda} + \mathbf{A}_{22}^T\boldsymbol{\mu} \geq -\mathbf{c}_2, \quad \boldsymbol{\mu} \geq \mathbf{0}. \end{array}$$

Notice the symmetry between the primal and dual problems.



# Quadratic Programming

A quadratic objective function and linear constraints define

*a QP problem.*

Equations from the KKT conditions: *linear*!

*Lagrange methods are the natural choice!*

With equality constraints only,

$$\text{Minimize} \quad f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}, \qquad \text{subject to} \quad \mathbf{A}\mathbf{x} = \mathbf{b}.$$

First order necessary conditions:

$$\begin{bmatrix} \mathbf{Q} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x}^* \\ \boldsymbol{\lambda} \end{bmatrix} = \begin{bmatrix} -\mathbf{c} \\ \mathbf{b} \end{bmatrix}$$

Solution of this linear system yields the complete result!

**Caution:** This coefficient matrix is *indefinite*.



# Quadratic Programming

**Active set method**

$$\begin{array}{ll} \text{Minimize} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}; \\ \text{subject to} & \mathbf{A}_1\mathbf{x} = \mathbf{b}_1, \\ & \mathbf{A}_2\mathbf{x} \leq \mathbf{b}_2. \end{array}$$

Start the iterative process from a feasible point.

- Construct active set of constraints as $\mathbf{A}\mathbf{x} = \mathbf{b}$.
- From the current point $\mathbf{x}_k$, with $\mathbf{x} = \mathbf{x}_k + \mathbf{d}_k$,

$$\begin{aligned} f(\mathbf{x}) &= \frac{1}{2}(\mathbf{x}_k + \mathbf{d}_k)^T\mathbf{Q}(\mathbf{x}_k + \mathbf{d}_k) + \mathbf{c}^T(\mathbf{x}_k + \mathbf{d}_k) \\ &= \frac{1}{2}\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k + (\mathbf{c} + \mathbf{Q}\mathbf{x}_k)^T\mathbf{d}_k + f(\mathbf{x}_k). \end{aligned}$$

- Since $\mathbf{g}_k \equiv \nabla f(\mathbf{x}_k) = \mathbf{c} + \mathbf{Q}\mathbf{x}_k$, subsidiary quadratic program:

  *minimize* $\frac{1}{2}\mathbf{d}_k^T\mathbf{Q}\mathbf{d}_k + \mathbf{g}_k^T\mathbf{d}_k$ *subject to* $\mathbf{A}\mathbf{d}_k = \mathbf{0}$.

- Examining solution $\mathbf{d}_k$ and Lagrange multipliers, decide to terminate, proceed or revise the active set.



# Quadratic Programming

**Linear complementary problem (LCP)**

*Slack variable strategy with inequality constraints*

$$\text{Minimize} \quad \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{c}^T\mathbf{x}, \qquad \text{subject to} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq \mathbf{0}.$$

KKT conditions: With $\mathbf{x},\mathbf{y},\boldsymbol{\mu},\boldsymbol{\nu} \geq \mathbf{0}$,

$$\begin{aligned} \mathbf{Q}\mathbf{x} + \mathbf{c} + \mathbf{A}^T\boldsymbol{\mu} - \boldsymbol{\nu} &= \mathbf{0}, \\ \mathbf{A}\mathbf{x} + \mathbf{y} &= \mathbf{b}, \\ \mathbf{x}^T\boldsymbol{\nu} = \boldsymbol{\mu}^T\mathbf{y} &= 0. \end{aligned}$$

Denoting

$$\mathbf{z} = \begin{bmatrix} \mathbf{x} \\ \boldsymbol{\mu} \end{bmatrix}, \mathbf{w} = \begin{bmatrix} \boldsymbol{\nu} \\ \mathbf{y} \end{bmatrix}, \mathbf{q} = \begin{bmatrix} \mathbf{c} \\ \mathbf{b} \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} \mathbf{Q} & \mathbf{A}^T \\ -\mathbf{A} & \mathbf{0} \end{bmatrix},$$

$$\mathbf{w} - \mathbf{M}\mathbf{z} = \mathbf{q}, \quad \mathbf{w}^T\mathbf{z} = \mathbf{0}.$$

Find mutually complementary *non-negative* $\mathbf{w}$ and $\mathbf{z}$.

## Quadratic Programming

If $\mathbf{q} \geq \mathbf{0}$, then $\mathbf{w} = \mathbf{q}$, $\mathbf{z} = \mathbf{0}$ is a solution!

**Lemke's method**: artificial variable $z_0$ with $\mathbf{e} = [1 \;\; 1 \;\; 1 \;\; \cdots \;\; 1]^T$:

$$\mathbf{Iw} - \mathbf{Mz} - \mathbf{e}z_0 = \mathbf{q}$$

With $z_0 = \max(-q_i)$,

$\mathbf{w} = \mathbf{q} + \mathbf{e}z_0 \geq \mathbf{0}$ *and* $\mathbf{z} = \mathbf{0}$*: basic feasible solution*

- Evolution of the basis similar to the simplex method.
- Out of a pair of $w$ and $z$ variables, only one can be there in any basis.
- At every step, one variable is driven out of the basis and its partner called in.
- The step driving out $z_0$ flags termination.

Handling of *equality constraints*? Very clumsy!!



## Points to note

- Fundamental issues and general perspective of the linear programming problem
- The simplex method
- Quadratic programming
  - The active set method
  - Lemke's method via the linear complementary problem

Necessary Exercises: **1,2,3,4,5**



## Outline





## Polynomial Interpolation

**Problem:** To develop an analytical representation of a function from information at discrete data points.

**Purpose**

- Evaluation at arbitrary points
- Differentiation and/or integration
- Drawing conclusion regarding the trends or *nature*

**Interpolation:** *one of the ways* of function representation

- sampled data are *exactly* satisfied

**Polynomial:** a convenient class of basis functions

For $y_i = f(x_i)$ for $i = 0, 1, 2, \cdots, n$ with $x_0 < x_1 < x_2 < \cdots < x_n$,

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n.$$

Find the coefficients such that $p(x_i) = f(x_i)$ for $i = 0, 1, 2, \cdots, n$.

*Values of* $p(x)$ *for* $x \in [x_0, x_n]$ **interpolate** $n+1$ *values of* $f(x)$*, an outside estimate is* **extrapolation**.

## Polynomial Interpolation



To determine $p(x)$, solve the linear system

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \cdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ f(x_2) \\ \cdots \\ f(x_n) \end{bmatrix} ?$$

*Vandermonde matrix*: invertible, but typically ill-conditioned!

Invertibility means existence and uniqueness of polynomial $p(x)$.

Two polynomials $p_1(x)$ and $p_2(x)$ matching the function $f(x)$ at $x_0, x_1, x_2, \cdots, x_n$ imply

*$n$-th degree polynomial $\Delta p(x) = p_1(x) - p_2(x)$ with $n+1$ roots!*

$\Delta p \equiv 0 \;\Rightarrow\; p_1(x) = p_2(x)$: $p(x)$ is unique.



## Polynomial Interpolation



**Lagrange interpolation**

Basis functions:

$$L_k(x) = \frac{\prod_{j=0,j\neq k}^{n}(x-x_j)}{\prod_{j=0,j\neq k}^{n}(x_k-x_j)} = \frac{(x-x_0)(x-x_1)\cdots(x-x_{k-1})(x-x_{k+1})\cdots(x-x_n)}{(x_k-x_0)(x_k-x_1)\cdots(x_k-x_{k-1})(x_k-x_{k+1})\cdots(x_k-x_n)}$$

Interpolating polynomial:

$$p(x) = \alpha_0 L_0(x) + \alpha_1 L_1(x) + \alpha_2 L_2(x) + \cdots + \alpha_n L_n(x)$$

At the data points, $L_k(x_i) = \delta_{ik}$.

*Coefficient matrix identity and $\alpha_i = f(x_i)$.*

Lagrange interpolation formula:

$$p(x) = \sum_{k=0}^{n} f(x_k)L_k(x) = L_0(x)f(x_0) + L_1(x)f(x_1) + \cdots + L_n(x)f(x_n)$$

Existence of $p(x)$ is a trivial consequence!



## Polynomial Interpolation



Two interpolation formulae

- one costly to determine, but easy to process
- the other trivial to determine, costly to process

**Newton interpolation** for an intermediate trade-off:
$p(x) = c_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) + \cdots + c_n \prod_{i=0}^{n-1}(x-x_i)$

**Hermite interpolation**

*uses derivatives as well as function values.*

Data: $f(x_i)$, $f'(x_i)$, $\cdots$, $f^{(n_i-1)}(x_i)$ at $x = x_i$, for $i = 0, 1, \cdots, m$:

- At $(m+1)$ points, a total of $n+1 = \sum_{i=0}^{m} n_i$ conditions

**Limitations of single-polynomial interpolation**

With large number of data points, polynomial degree is high.

- Computational cost and numerical imprecision
- Lack of representative nature due to *oscillations*



## Piecewise Polynomial Interpolation



**Piecewise linear interpolation**

$$f(x) = f(x_{i-1}) + \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}}(x - x_{i-1}) \quad \text{for} \quad x \in [x_{i-1}, x_i]$$

Handy for many uses with dense data. *But, not differentiable.*

**Piecewise cubic interpolation**

With function values and derivatives at $(n+1)$ points,

*$n$ cubic Hermite segments*

Data for the $j$-th segment:

$$f(x_{j-1}) = f_{j-1},\; f(x_j) = f_j,\; f'(x_{j-1}) = f'_{j-1} \text{ and } f'(x_j) = f'_j$$

Interpolating polynomial:

$$p_j(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$$

Coefficients $a_0$, $a_1$, $a_2$, $a_3$: linear combinations of $f_{j-1}$, $f_j$, $f'_{j-1}$, $f'_j$

*Composite function $\mathcal{C}^1$ continuous at knot points.*

# Piecewise Polynomial Interpolation



**General formulation** through normalization of intervals

$$x = x_{j-1} + t(x_j - x_{j-1}), \;\; t \in [0,1]$$

With $g(t) = f(x(t)),\; g'(t) = (x_j - x_{j-1})f'(x(t))$;

$$g_0 = f_{j-1},\; g_1 = f_j,\; g_0' = (x_j - x_{j-1})f_{j-1}' \text{ and } g_1' = (x_j - x_{j-1})f_j'.$$

Cubic polynomial for the $j$-th segment:

$$q_j(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3$$

Modular expression:

$$q_j(t) = [\alpha_0 \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3] \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \end{bmatrix} = [g_0 \;\; g_1 \;\; g_0' \;\; g_1'] \, \mathbf{W} \begin{bmatrix} 1 \\ t \\ t^2 \\ t^3 \end{bmatrix} = \mathbf{G}_j \mathbf{W} \mathbf{T}$$

Packaging data, interpolation type and variable terms separately!

**Question:** *How to supply derivatives? And,* ***why****?*



# Piecewise Polynomial Interpolation

Polynomial Interpolation
Piecewise Polynomial Interpolation
Interpolation of Multivariate Functions
A Note on Approximation of Functions
Modelling of Curves and Surfaces*

**Spline interpolation**

| Spline: a drafting tool to draw a smooth curve through key points. |
|---|

Data: $f_i = f(x_i)$, for $x_0 < x_1 < x_2 < \cdots < x_n$.

If $k_j = f'(x_j)$, then

*$p_j(x)$ can be determined in terms of $f_{j-1}, f_j, k_{j-1}, k_j$ and $p_{j+1}(x)$ in terms of $f_j, f_{j+1}, k_j, k_{j+1}$.*

Then, $p_j''(x_j) = p_{j+1}''(x_j)$: a linear equation in $k_{j-1}$, $k_j$ and $k_{j+1}$

From $n-1$ interior knot points,

*$n-1$ linear equations in derivative values $k_0, k_1, \cdots, k_n$.*

Prescribing $k_0$ and $k_n$, a **diagonally dominant tridiagonal** system!

*A spline is a* **smooth** *interpolation, with $\mathcal{C}^2$ continuity.*



# Interpolation of Multivariate Functions

Polynomial Interpolation
Piecewise Polynomial Interpolation
Interpolation of Multivariate Functions
A Note on Approximation of Functions
Modelling of Curves and Surfaces*

**Piecewise bilinear interpolation**

Data: $f(x,y)$ over a dense rectangular grid

$$x = x_0, x_1, x_2, \cdots, x_m \text{ and } y = y_0, y_1, y_2, \cdots, y_n$$

Rectangular domain: $\{(x,y) : x_0 \le x \le x_m, y_0 \le y \le y_n\}$

For $x_{i-1} \le x \le x_i$ and $y_{j-1} \le y \le y_j$,

$$f(x,y) = a_{0,0} + a_{1,0}x + a_{0,1}y + a_{1,1}xy = [1 \;\; x] \begin{bmatrix} a_{0,0} & a_{0,1} \\ a_{1,0} & a_{1,1} \end{bmatrix} \begin{bmatrix} 1 \\ y \end{bmatrix}$$

With data at four corner points, coefficient matrix determined from

$$\begin{bmatrix} 1 & x_{i-1} \\ 1 & x_i \end{bmatrix} \begin{bmatrix} a_{0,0} & a_{0,1} \\ a_{1,0} & a_{1,1} \end{bmatrix} \begin{bmatrix} 1 & 1 \\ y_{j-1} & y_j \end{bmatrix} = \begin{bmatrix} f_{i-1,j-1} & f_{i-1,j} \\ f_{i,j-1} & f_{i,j} \end{bmatrix}.$$

*Approximation only $\mathcal{C}^0$ continuous.*



# Interpolation of Multivariate Functions

Polynomial Interpolation
Piecewise Polynomial Interpolation
Interpolation of Multivariate Functions
A Note on Approximation of Functions
Modelling of Curves and Surfaces*

**Alternative local formula through reparametrization**

With $u = \frac{x - x_{i-1}}{x_i - x_{i-1}}$ and $v = \frac{y - y_{j-1}}{y_j - y_{j-1}}$, denoting

$$f_{i-1,j-1} = g_{0,0},\; f_{i,j-1} = g_{1,0},\; f_{i-1,j} = g_{0,1} \text{ and } f_{i,j} = g_{1,1};$$

bilinear interpolation:

$$g(u,v) = [1 \;\; u] \begin{bmatrix} \alpha_{0,0} & \alpha_{0,1} \\ \alpha_{1,0} & \alpha_{1,1} \end{bmatrix} \begin{bmatrix} 1 \\ v \end{bmatrix} \quad \text{for } u, v \in [0,1].$$

Values at four corner points fix the coefficient matrix as

$$\begin{bmatrix} \alpha_{0,0} & \alpha_{0,1} \\ \alpha_{1,0} & \alpha_{1,1} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} g_{0,0} & g_{0,1} \\ g_{1,0} & g_{1,1} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

Concisely, $\boxed{g(u,v) = \mathbf{U}^T \mathbf{W}^T \mathbf{G}_{i,j} \mathbf{W} \mathbf{V}}$ in which

$$\mathbf{U} = \begin{bmatrix} 1 \\ u \end{bmatrix}, \; \mathbf{V} = \begin{bmatrix} 1 \\ v \end{bmatrix}, \; \mathbf{W} = \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}, \; \mathbf{G}_{i,j} = \begin{bmatrix} f_{i-1,j-1} & f_{i-1,j} \\ f_{i,j-1} & f_{i,j} \end{bmatrix}.$$

## Interpolation of Multivariate Functions



**Piecewise bicubic interpolation**

Data: $f$, $\frac{\partial f}{\partial x}$, $\frac{\partial f}{\partial y}$ and $\frac{\partial^2 f}{\partial x \partial y}$ over grid points

With normalizing parameters $u$ and $v$,

$$\frac{\partial g}{\partial u} = (x_i - x_{i-1})\frac{\partial f}{\partial x}, \quad \frac{\partial g}{\partial v} = (y_j - y_{j-1})\frac{\partial f}{\partial y}, \quad \text{and}$$

$$\frac{\partial^2 g}{\partial u \partial v} = (x_i - x_{i-1})(y_j - y_{j-1})\frac{\partial^2 f}{\partial x \partial y}$$

In $\{(x,y) : x_{i-1} \le x \le x_i, y_{j-1} \le y \le y_j\}$ or $\{(u,v) : u, v \in [0,1]\}$,

$$g(u,v) = \mathbf{U}^T \mathbf{W}^T \mathbf{G}_{i,j} \mathbf{W} \mathbf{V},$$

with $\mathbf{U} = [1 \;\; u \;\; u^2 \;\; u^3]^T$, $\mathbf{V} = [1 \;\; v \;\; v^2 \;\; v^3]^T$, and

$$\mathbf{G}_{i,j} = \begin{bmatrix} g(0,0) & g(0,1) & g_v(0,0) & g_v(0,1) \\ g(1,0) & g(1,1) & g_v(1,0) & g_v(1,1) \\ g_u(0,0) & g_u(0,1) & g_{uv}(0,0) & g_{uv}(0,1) \\ g_u(1,0) & g_u(1,1) & g_{uv}(1,0) & g_{uv}(1,1) \end{bmatrix}.$$



## A Note on Approximation of Functions



A common strategy of function approximation is to

- express a function as a linear combination of a set of basis functions (*which*?), and
- determine coefficients based on some criteria (*what*?).

**Criteria:**

Interpolatory approximation: Exact agreement with sampled data

Least square approximation: Minimization of a sum (or integral) of square errors over sampled data

Minimax approximation: Limiting the largest deviation

Basis functions:

*polynomials, sinusoids, orthogonal eigenfunctions or field-specific heuristic choice*



## Points to note



- Lagrange, Newton and Hermite interpolations
- Piecewise polynomial functions and splines
- Bilinear and bicubic interpolation of bivariate functions

*Direct extension to vector functions: curves and surfaces!*

Necessary Exercises: **1,2,4,6**



## Outline

## Newton-Cotes Integration Formulae

$$J = \int_a^b f(x)dx$$

Divide $[a,b]$ into $n$ sub-intervals with

$$a = x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n = b,$$

where $x_i - x_{i-1} = h = \frac{b-a}{n}$.

$$\bar{J} = \sum_{i=1}^{n} hf(x_i^*) = h[f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)]$$

Taking $x_i^* \in [x_{i-1}, x_i]$ as $x_{i-1}$ and $x_i$, we get summations $J_1$ and $J_2$.

*As $n \to \infty$ (i.e. $h \to 0$), if $J_1$ and $J_2$ approach the same limit, then function $f(x)$ is* integrable *over interval $[a,b]$.*

A rectangular rule or a one-point rule

**Question:** Which point to take as $x_i^*$?



## Newton-Cotes Integration Formulae

**Mid-point rule**

Selecting $x_i^*$ as $\bar{x}_i = \frac{x_{i-1}+x_i}{2}$,

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx hf(\bar{x}_i) \quad \text{and} \quad \int_a^b f(x)dx \approx h\sum_{i=1}^{n} f(\bar{x}_i).$$

Error analysis: From Taylor's series of $f(x)$ about $\bar{x}_i$,

$$\int_{x_{i-1}}^{x_i} f(x)dx = \int_{x_{i-1}}^{x_i} \left[f(\bar{x}_i) + f'(\bar{x}_i)(x-\bar{x}_i) + f''(\bar{x}_i)\frac{(x-\bar{x}_i)^2}{2} + \cdots\right] dx$$

$$= hf(\bar{x}_i) + \frac{h^3}{24}f''(\bar{x}_i) + \frac{h^5}{1920}f^{iv}(\bar{x}_i) + \cdots,$$

*third* order accurate!

Over the entire domain $[a,b]$,

$$\int_a^b f(x)dx \approx h\sum_{i=1}^{n} f(\bar{x}_i) + \frac{h^3}{24}\sum_{i=1}^{n} f''(\bar{x}_i) = h\sum_{i=1}^{n} f(\bar{x}_i) + \frac{h^2}{24}(b-a)f''(\xi),$$

for $\xi \in [a,b]$ (from mean value theorem): second order accurate.



## Newton-Cotes Integration Formulae

**Trapezoidal rule**

Approximating function $f(x)$ with a linear interpolation,

$$\int_{x_{i-1}}^{x_i} f(x)dx \approx \frac{h}{2}[f(x_{i-1}) + f(x_i)]$$

and

$$\int_a^b f(x)dx \approx h\left[\frac{1}{2}f(x_0) + \sum_{i=1}^{n-1} f(x_i) + \frac{1}{2}f(x_n)\right].$$

Taylor series expansions about the mid-point:

$$f(x_{i-1}) = f(\bar{x}_i) - \frac{h}{2}f'(\bar{x}_i) + \frac{h^2}{8}f''(\bar{x}_i) - \frac{h^3}{48}f'''(\bar{x}_i) + \frac{h^4}{384}f^{iv}(\bar{x}_i) - \cdots$$

$$f(x_i) = f(\bar{x}_i) + \frac{h}{2}f'(\bar{x}_i) + \frac{h^2}{8}f''(\bar{x}_i) + \frac{h^3}{48}f'''(\bar{x}_i) + \frac{h^4}{384}f^{iv}(\bar{x}_i) + \cdots$$

$$\Rightarrow \frac{h}{2}[f(x_{i-1}) + f(x_i)] = hf(\bar{x}_i) + \frac{h^3}{8}f''(\bar{x}_i) + \frac{h^5}{384}f^{iv}(\bar{x}_i) + \cdots$$

Recall $\int_{x_{i-1}}^{x_i} f(x)dx = hf(\bar{x}_i) + \frac{h^3}{24}f''(\bar{x}_i) + \frac{h^5}{1920}f^{iv}(\bar{x}_i) + \cdots$.



## Newton-Cotes Integration Formulae

**Error estimate of trapezoidal rule**

$$\int_{x_{i-1}}^{x_i} f(x)dx = \frac{h}{2}[f(x_{i-1}) + f(x_i)] - \frac{h^3}{12}f''(\bar{x}_i) - \frac{h^5}{480}f^{iv}(\bar{x}_i) + \cdots$$

Over an extended domain,

$$\int_a^b f(x)dx = h\left[\frac{1}{2}\{f(x_0) + f(x_n)\} + \sum_{i=1}^{n-1} f(x_i)\right] - \frac{h^2}{12}(b-a)f''(\xi) + \cdots.$$

The same order of accuracy as the mid-point rule!

Different sources of merit

- **Mid-point rule:** Use of mid-point leads to symmetric error-cancellation.
- **Trapezoidal rule:** Use of end-points allows double utilization of boundary points in adjacent intervals.

How to use **both the merits**?

## Newton-Cotes Integration Formulae

**Simpson's rules**

Divide $[a, b]$ into an even number ($n = 2m$) of intervals.
Fit a quadratic polynomial over a panel of two intervals.
For this panel of length $2h$, two estimates:

$$M(f) = 2hf(x_i) \quad \text{and} \quad T(f) = h[f(x_{i-1}) + f(x_{i+1})]$$

$$J = M(f) + \frac{h^3}{3}f''(x_i) + \frac{h^5}{60}f^{iv}(x_i) + \cdots$$
$$J = T(f) - \frac{2h^3}{3}f''(x_i) - \frac{h^5}{15}f^{iv}(x_i) + \cdots$$

Simpson's one-third rule (with error estimate):

$$\int_{x_{i-1}}^{x_{i+1}} f(x)dx = \frac{h}{3}[f(x_{i-1}) + 4f(x_i) + f(x_{i+1})] - \frac{h^5}{90}f^{iv}(x_i)$$

Fifth (not fourth) order accurate!

A four-point rule: *Simpson's three-eighth rule*

*Still higher order rules* **NOT** *advisable!*



## Richardson Extrapolation and Romberg Integration

To determine quantity $F$

- using a step size $h$, estimate $F(h)$
- error terms: $h^p$, $h^q$, $h^r$ etc ($p < q < r$)
- $F = \lim_{\delta \to 0} F(\delta)$?
- plot $F(h)$, $F(\alpha h)$, $F(\alpha^2 h)$ (with $\alpha < 1$) and extrapolate?

$$\boxed{1} \quad F(h) = F + ch^p + \mathcal{O}(h^q)$$
$$\boxed{2} \quad F(\alpha h) = F + c(\alpha h)^p + \mathcal{O}(h^q)$$
$$\boxed{4} \quad F(\alpha^2 h) = F + c(\alpha^2 h)^p + \mathcal{O}(h^q)$$

Eliminate $c$ and determine (better estimates of) $F$:

$$\boxed{3} \quad F_1(h) = \frac{F(\alpha h) - \alpha^p F(h)}{1 - \alpha^p} = F + c_1 h^q + \mathcal{O}(h^r)$$
$$\boxed{5} \quad F_1(\alpha h) = \frac{F(\alpha^2 h) - \alpha^p F(\alpha h)}{1 - \alpha^p} = F + c_1(\alpha h)^q + \mathcal{O}(h^r)$$

Still better estimate: $\boxed{6} \quad F_2(h) = \frac{F_1(\alpha h) - \alpha^q F_1(h)}{1 - \alpha^q} = F + \mathcal{O}(h^r)$

Richardson extrapolation



## Richardson Extrapolation and Romberg Integration

Trapezoidal rule for $J = \int_a^b f(x)dx$: $p = 2$, $q = 4$, $r = 6$ etc

$$T(f) = J + ch^2 + dh^4 + eh^6 + \cdots$$

*With $\alpha = \frac{1}{2}$, half the sum available for successive levels.*

**Romberg integration**

- Trapezoidal rule with $h = H$: find $J_{11}$.
- With $h = H/2$, find $J_{12}$.

$$J_{22} = \frac{J_{12} - \left(\frac{1}{2}\right)^2 J_{11}}{1 - \left(\frac{1}{2}\right)^2} = \frac{4J_{12} - J_{11}}{3}.$$

- If $|J_{22} - J_{12}|$ is within tolerance, STOP. Accept $J \approx J_{22}$.
- With $h = H/4$, find $J_{13}$.

$$J_{23} = \frac{4J_{13} - J_{12}}{3} \quad \text{and} \quad J_{33} = \frac{J_{23} - \left(\frac{1}{2}\right)^4 J_{22}}{1 - \left(\frac{1}{2}\right)^4} = \frac{16J_{23} - J_{22}}{15}.$$

- If $|J_{33} - J_{23}|$ is within tolerance, STOP with $J \approx J_{33}$.



## Further Issues

Featured functions: *adaptive quadrature*

- With prescribed tolerance $\epsilon$, assign quota $\epsilon_i = \frac{\epsilon(x_i - x_{i-1})}{b-a}$ of error to every interval $[x_{i-1}, x_i]$.
- For each interval, find *two* estimates of the integral and estimate the error.
- If error estimate is not within quota, then subdivide.

Function as tabulated data

- Only trapezoidal rule applicable?
- Fit a spline over data points and integrate the segments?

Improper integral: Newton-Cotes *closed formulae* not applicable!

- Open Newton-Cotes formulae
- Gaussian quadrature

# Points to note

- Definition of an integral and *integrability*
- Closed Newton-Cotes formulae and their error estimates
- Richardson extrapolation as a general technique
- Romberg integration
- Adaptive quadrature

Necessary Exercises: **1,2,3,4**



# Outline





# Gaussian Quadrature

A typical quadrature formula: a weighted sum $\sum_{i=0}^{n} w_i f_i$

- $f_i$: function value at $i$-th sampled point
- $w_i$: corresponding weight

Newton-Cotes formulae:

- Abscissas ($x_i$'s) of sampling prescribed
- Coefficients or weight values determined to eliminate dominant error terms

Gaussian quadrature rules:

- no prescription of quadrature points
- only the 'number' of quadrature points prescribed
- locations as well as weights contribute to the accuracy criteria
- with $n$ integration points, $2n$ degrees of freedom
- can be made exact for polynomials of degree up to $2n-1$
- best locations: interior points
- open quadrature rules: can handle integrable singularities



# Gaussian Quadrature

**Gauss-Legendre quadrature**

$$\int_{-1}^{1} f(x)dx = w_1 f(x_1) + w_2 f(x_2)$$

Four variables: Insist that it is exact for 1, $x$, $x^2$ and $x^3$.

$$\begin{aligned} w_1 + w_2 &= \int_{-1}^{1} dx = 2, \\ w_1 x_1 + w_2 x_2 &= \int_{-1}^{1} x dx = 0, \\ w_1 x_1^2 + w_2 x_2^2 &= \int_{-1}^{1} x^2 dx = \frac{2}{3} \\ \text{and} \quad w_1 x_1^3 + w_2 x_2^3 &= \int_{-1}^{1} x^3 dx = 0. \end{aligned}$$

$$x_1 = -x_2,\ w_1 = w_2 \Rightarrow \boxed{w_1 = w_2 = 1,\ x_1 = -\frac{1}{\sqrt{3}},\ x_2 = \frac{1}{\sqrt{3}}}$$

## Gaussian Quadrature

**Two-point Gauss-Legendre quadrature formula**

$$\int_{-1}^{1} f(x)dx = f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}})$$

Exact for any cubic polynomial: parallels Simpson's rule!
Three-point quadrature rule along similar lines:

$$\int_{-1}^{1} f(x)dx = \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + \frac{5}{9}f\left(\sqrt{\frac{3}{5}}\right)$$

A large number of formulae: Consult mathematical handbooks.
For domain of integration $[a, b]$,

$$x = \frac{a+b}{2} + \frac{b-a}{2}t \quad \text{and} \quad dx = \frac{b-a}{2}dt$$

With scaling and relocation,

$$\int_a^b f(x)dx = \frac{b-a}{2}\int_{-1}^{1} f[x(t)]dt$$



## Gaussian Quadrature

**General Framework** for $n$-point formula

- $f(x)$: a polynomial of degree $2n-1$
- $p(x)$: Lagrange polynomial through the $n$ quadrature points
- $f(x) - p(x)$: a $(2n-1)$-degree polynomial having $n$ of its roots at the quadrature points

Then, with $\phi(x) = (x - x_1)(x - x_2)\cdots(x - x_n)$,

$$f(x) - p(x) = \phi(x)q(x).$$

Quotient polynomial: $q(x) = \sum_{i=0}^{n-1} \alpha_i x^i$
Direct integration:

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} p(x)dx + \int_{-1}^{1}\left[\phi(x)\sum_{i=0}^{n-1}\alpha_i x^i\right]dx$$

How to make the second term vanish?



## Gaussian Quadrature

Choose quadrature points $x_1, x_2, \cdots, x_n$ so that $\phi(x)$ is orthogonal to all polynomials of degree less than $n$.

*Legendre polynomial*

Gauss-Legendre quadrature

1. Choose $P_n(x)$, Legendre polynomial of degree $n$, as $\phi(x)$.
2. Take its roots $x_1, x_2, \cdots, x_n$ as the quadrature points.
3. Fit Lagrange polynomial of $f(x)$, using these $n$ points.

$$p(x) = L_1(x)f(x_1) + L_2(x)f(x_2) + \cdots + L_n(x)f(x_n)$$

4. 

$$\int_{-1}^{1} f(x)dx = \int_{-1}^{1} p(x)dx = \sum_{j=1}^{n} f(x_j)\int_{-1}^{1} L_j(x)dx$$

Weight values: $w_j = \int_{-1}^{1} L_j(x)dx, \quad \text{for } j = 1, 2, \cdots, n$



## Gaussian Quadrature

**Weight functions in Gaussian quadrature**

What is so great about exact integration of polynomials?

Demand something else: generalization

*Exact integration of polynomials times function $W(x)$*

Given weight function $W(x)$ and number $(n)$ of quadrature points,

*work out the locations ($x_j$'s) of the $n$ points and the corresponding weights ($w_j$'s), so that integral*

$$\int_a^b W(x)f(x)dx = \sum_{j=1}^{n} w_j f(x_j)$$

*is exact for an arbitrary polynomial $f(x)$ of degree up to $(2n-1)$.*

## Gaussian Quadrature

A family of orthogonal polynomials with increasing degree:

*quadrature points: roots of n-th member of the family.*

For different kinds of functions and different domains,

- Gauss-Chebyshev quadrature
- Gauss-Laguerre quadrature
- Gauss-Hermite quadrature
- ... ... ...

Several singular functions and infinite domains can be handled.

A very special case:

*For $W(x) = 1$, Gauss-Legendre quadrature!*



## Multiple Integrals

$$S = \int_a^b \int_{g_1(x)}^{g_2(x)} f(x,y)\ dy\ dx$$

$$\Rightarrow\ F(x) = \int_{g_1(x)}^{g_2(x)} f(x,y)\ dy \ \text{ and }\ S = \int_a^b F(x)dx$$

with complete flexibility of individual quadrature methods.

**Double integral on rectangular domain**

Two-dimensional version of Simpson's one-third rule:

$$\begin{aligned}&\int_{-1}^1 \int_{-1}^1 f(x,y)dxdy\\ &= w_0 f(0,0) + w_1[f(-1,0) + f(1,0) + f(0,-1) + f(0,1)]\\ &\quad + w_2[f(-1,-1) + f(-1,1) + f(1,-1) + f(1,1)]\end{aligned}$$

Exact for bicubic functions: $w_0 = 16/9$, $w_1 = 4/9$ and $w_2 = 1/9$.



## Multiple Integrals

**Monte Carlo integration**

$$I = \int_\Omega f(\mathbf{x})dV$$

Requirements:

- a simple volume $V$ enclosing the domain $\Omega$
- a point classification scheme

Generating random points in $V$,

$$F(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } \mathbf{x} \in \Omega, \\ 0 & \text{otherwise} . \end{cases}$$

$$I \approx \frac{V}{N} \sum_{i=1}^N F(\mathbf{x}_i)$$

Estimate of $I$ (usually) improves with increasing N.



## Points to note

- Basic strategy of Gauss-Legendre quadrature
- Formulation of a double integral from fundamental principle
- Monte Carlo integration

Necessary Exercises: **2,5,6**

# Outline







# Single-Step Methods



**Initial value problem (IVP)** of a first order ODE:

$$\frac{dy}{dx} = f(x, y), \quad y(x_0) = y_0$$

To determine: $y(x)$ for $x \in [a, b]$ with $x_0 = a$.

Numerical solution: Start from the point $(x_0, y_0)$.

- $y_1 = y(x_1) = y(x_0 + h) = ?$
- Found $(x_1, y_1)$. Repeat up to $x = b$.

Information at how many points are used at every step?

- **Single-step method:** Only the current value
- **Multi-step method:** History of several recent steps



# Single-Step Methods



**Euler's method**

- At $(x_n, y_n)$, evaluate slope $\frac{dy}{dx} = f(x_n, y_n)$.
- For a small step $h$,

$$y_{n+1} = y_n + hf(x_n, y_n)$$

Repitition of such steps constructs $y(x)$.

First order truncated Taylor's series:

*Expected error:* $\mathcal{O}(h^2)$

Accumulation over steps

*Total error:* $\mathcal{O}(h)$

Euler's method is a first order method.

**Question:** Total error = Sum of errors over the steps?
**Answer:** No, in general.



# Single-Step Methods



*Initial* slope for the entire step: is it a good idea?

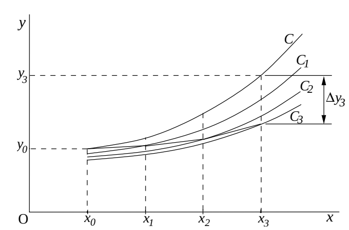



Figure: Euler's method

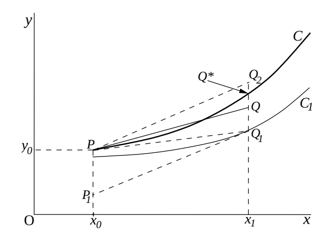



Figure: Improved Euler's method

**Improved Euler's method or Heun's method**

$$\bar{y}_{n+1} = y_n + hf(x_n, y_n)$$
$$y_{n+1} = y_n + \frac{h}{2}[f(x_n, y_n) + f(x_{n+1}, \bar{y}_{n+1})]$$

The order of Heun's method is two.

## Single-Step Methods

**Runge-Kutta methods**
Second order method:

$$k_1 = hf(x_n, y_n), \quad k_2 = hf(x_n + \alpha h, y_n + \beta k_1)$$
$$k = w_1 k_1 + w_2 k_2,$$
$$\text{and} \quad x_{n+1} = x_n + h, \quad y_{n+1} = y_n + k$$

Force agreement up to the second order.

$$y_{n+1}$$
$$= y_n + w_1 hf(x_n, y_n) + w_2 h[f(x_n, y_n) + \alpha h f_x(x_n, y_n) + \beta k_1 f_y(x_n, y_n) + \cdots$$
$$= y_n + (w_1 + w_2) hf(x_n, y_n) + h^2 w_2 [\alpha f_x(x_n, y_n) + \beta f(x_n, y_n) f_y(x_n, y_n)] +$$

From Taylor's series, using $y' = f(x, y)$ and $y'' = f_x + f f_y$,

$$y(x_{n+1}) = y_n + hf(x_n, y_n) + \frac{h^2}{2}[f_x(x_n, y_n) + f(x_n, y_n) f_y(x_n, y_n)] + \cdots$$

$$w_1 + w_2 = 1, \;\; \alpha w_2 = \beta w_2 = \tfrac{1}{2} \;\; \Rightarrow \;\; \boxed{\alpha = \beta = \tfrac{1}{2w_2}, \;\; w_1 = 1 - w_2}$$



## Single-Step Methods

With continuous choice of $w_2$,

*a family of second order Runge Kutta (RK2) formulae*

Popular form of RK2: with choice $w_2 = 1$,

$$k_1 = hf(x_n, y_n), \quad k_2 = hf(x_n + \tfrac{h}{2}, y_n + \tfrac{k_1}{2})$$
$$x_{n+1} = x_n + h, \quad y_{n+1} = y_n + k_2$$

**Fourth order Runge-Kutta method** (RK4):

$$k_1 = hf(x_n, y_n)$$
$$k_2 = hf(x_n + \tfrac{h}{2}, y_n + \tfrac{k_1}{2})$$
$$k_3 = hf(x_n + \tfrac{h}{2}, y_n + \tfrac{k_2}{2})$$
$$k_4 = hf(x_n + h, y_n + k_3)$$

$$k = \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

$$x_{n+1} = x_n + h, \quad y_{n+1} = y_n + k$$



## Practical Implementation of Single-Step Methods

**Question:** How to decide whether the error is within tolerance?
Additional estimates:

- handle to monitor the error
- further efficient algorithms

**Runge-Kutta method with adaptive step size**
In an interval $[x_n, x_n + h]$,

$$y^{(1)}_{n+1} = y_{n+1} + ch^5 + \text{higher order terms}$$

Over two steps of size $\frac{h}{2}$,

$$y^{(2)}_{n+1} = y_{n+1} + 2c\left(\frac{h}{2}\right)^5 + \text{higher order terms}$$

Difference of two estimates:

$$\Delta = y^{(1)}_{n+1} - y^{(2)}_{n+1} \approx \frac{15}{16} ch^5$$

Best available value: $y^*_{n+1} = y^{(2)}_{n+1} - \frac{\Delta}{15} = \frac{16 y^{(2)}_{n+1} - y^{(1)}_{n+1}}{15}$



## Practical Implementation of Single-Step Methods

Evaluation of a step:

$\Delta > \epsilon$: Step size is too large for accuracy. Subdivide the interval.
$\Delta << \epsilon$: Step size is inefficient!

Start with a large step size.
Keep subdividing intervals whenever $\Delta > \epsilon$.

*Fast marching over smooth segments and small steps in zones featured with rapid changes in $y(x)$.*

**Runge-Kutta-Fehlberg method**

With six function values,

*An RK4 formula embedded in an RK5 formula*

- *two independent estimates and an error estimate!*

**RKF45** in professional implementations

## Systems of ODE's

Methods for a single first order ODE

*directly applicable to a first order* vector *ODE*

A typical IVP with an ODE system:

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

An $n$-th order ODE: convert into a system of first order ODE's
Defining *state vector* $\mathbf{z}(x) = [y(x) \quad y'(x) \quad \cdots \quad y^{(n-1)}(x)]^T$,

*work out* $\frac{d\mathbf{z}}{dx}$ *to form the* state space equation.

Initial condition: $\mathbf{z}(x_0) = [y(x_0) \quad y'(x_0) \quad \cdots \quad y^{(n-1)}(x_0)]^T$

A system of higher order ODE's with the highest order derivatives of orders $n_1, n_2, n_3, \cdots, n_k$

- Cast into the *state space form* with the state vector of dimension $n = n_1 + n_2 + n_3 + \cdots + n_k$



## Systems of ODE's

**State space formulation** is directly applicable when

*the highest order derivatives can be solved explicitly.*

The resulting form of the ODE's: *normal system* of ODE's

**Example:**

$$\begin{aligned} y\frac{d^2x}{dt^2} - 3\left(\frac{dy}{dt}\right)\left(\frac{dx}{dt}\right)^2 + 2x\left(\frac{dx}{dt}\right)\sqrt{\frac{d^2y}{dt^2}} + 4 &= 0 \\ e^{xy}\frac{d^3y}{dt^3} - y\left(\frac{d^2y}{dt^2}\right)^{3/2} + 2x + 1 &= e^{-t} \end{aligned}$$

State vector: $\mathbf{z}(t) = \left[x \quad \frac{dx}{dt} \quad y \quad \frac{dy}{dt} \quad \frac{d^2y}{dt^2}\right]^T$
With three trivial derivatives $z_1'(t) = z_2$, $z_3'(t) = z_4$ and $z_4'(t) = z_5$ and the other two obtained from the given ODE's,

*we get the state space equations as* $\frac{d\mathbf{z}}{dt} = \mathbf{f}(t, \mathbf{z})$.



## Multi-Step Methods*

Single-step methods: every step a brand new IVP!

Why not try to capture the trend?

**A typical multi-step formula:**

$$\begin{aligned} y_{n+1} &= y_n + h[c_0 f(x_{n+1}, y_{n+1}) + c_1 f(x_n, y_n) \\ &\quad + c_2 f(x_{n-1}, y_{n-1}) + c_3 f(x_{n-2}, y_{n-2}) + \cdots] \end{aligned}$$

Determine coefficients by demanding the exactness for leading polynomial terms.

Explicit methods: $c_0 = 0$, evaluation easy, but involves extrapolation.

Implicit methods: $c_0 \neq 0$, difficult to evaluate, but better stability.

**Predictor-corrector methods**

*Example: Adams-Bashforth-Moulton method*



## Points to note

- Euler's and Runge-Kutta methods
- Step size adaptation
- State space formulation of dynamic systems

Necessary Exercises: **1,2,5,6**

# Outline





# Stability Analysis

Adaptive RK4 is an extremely successful method.

*But, its* scope *has a limitation.*

Focus of explicit methods (such as RK) is accuracy and efficiency.

*The issue of* stabilty *is handled* indirectly.

**Stabilty of explicit methods**

For the ODE system $\mathbf{y}' = \mathbf{f}(x, \mathbf{y})$, Euler's method gives

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{f}(x_n, \mathbf{y}_n)h + \mathcal{O}(h^2).$$

Taylor's series of the actual solution:

$$\mathbf{y}(x_{n+1}) = \mathbf{y}(x_n) + \mathbf{f}(x_n, \mathbf{y}(x_n))h + \mathcal{O}(h^2)$$

Discrepancy or error:

$$\begin{aligned} \Delta_{n+1} &= \mathbf{y}_{n+1} - \mathbf{y}(x_{n+1}) \\ &= [\mathbf{y}_n - \mathbf{y}(x_n)] + [\mathbf{f}(x_n, \mathbf{y}_n) - \mathbf{f}(x_n, \mathbf{y}(x_n))]h + \mathcal{O}(h^2) \\ &= \Delta_n + \left[\frac{\partial \mathbf{f}}{\partial \mathbf{y}}(x_n, \bar{\mathbf{y}}_n)\Delta_n\right] h + \mathcal{O}(h^2) \approx (\mathbf{I} + h\mathbf{J})\Delta_n \end{aligned}$$



# Stability Analysis

Euler's step magnifies the error by a factor $(\mathbf{I} + h\mathbf{J})$.

Using $\mathbf{J}$ loosely as the representative Jacobian,

$$\Delta_{n+1} \approx (\mathbf{I} + h\mathbf{J})^n \Delta_1.$$

For stability, $\Delta_{n+1} \to 0$ as $n \to \infty$.

*Eigenvalues of* $(\mathbf{I} + h\mathbf{J})$ *must fall within the unit circle* $|z| = 1$. *By shift theorem, eigenvalues of* $h\mathbf{J}$ *must fall inside the unit circle with the centre at* $z_0 = -1$.

$$|1 + h\lambda| < 1 \;\Rightarrow\; h < \frac{-2\mathrm{Re}\,(\lambda)}{|\lambda|^2}$$

**Note:** Same result for single ODE $w' = \lambda w$, with complex $\lambda$.

For second order Runge-Kutta method,

$$\Delta_{n+1} = \left[1 + h\lambda + \frac{h^2\lambda^2}{2}\right]\Delta_n$$

Region of stability in the plane of $z = h\lambda$: $\left|1 + z + \frac{z^2}{2}\right| < 1$



# Stability Analysis

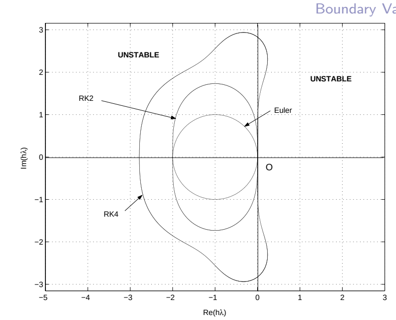



Figure: Stability regions of explicit methods

**Question:** What do these stability regions mean with reference to the system eigenvalues?

**Question:** How does the step size adaptation of RK4 operate on a system with eigenvalues on the left half of complex plane?

Step size adaptation tackles instability by its symptom!

# Implicit Methods

**Backward Euler's method**

$$\mathbf{y}_{n+1} = \mathbf{y}_n + \mathbf{f}(x_{n+1}, \mathbf{y}_{n+1})h$$

Solve it? Is it worth solving?

$$\begin{aligned}\Delta_{n+1} &\approx \mathbf{y}_{n+1} - \mathbf{y}(x_{n+1}) \\ &= [\mathbf{y}_n - \mathbf{y}(x_n)] + h[\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}) - \mathbf{f}(x_{n+1}, \mathbf{y}(x_{n+1}))] \\ &= \Delta_n + h\mathbf{J}(x_{n+1}, \bar{\mathbf{y}}_{n+1})\Delta_{n+1}\end{aligned}$$

Notice the flip in the form of this equation.

$$\Delta_{n+1} \approx (\mathbf{I} - h\mathbf{J})^{-1}\Delta_n$$

Stability: eigenvalues of $(\mathbf{I} - h\mathbf{J})$ *outside* the unit circle $|z| = 1$

$$|h\lambda - 1| > 1 \;\Rightarrow\; h > \frac{2\text{Re}\,(\lambda)}{|\lambda|^2}$$

**Absolute stability** for a stable ODE, i.e. one with Re $(\lambda) < 0$



# Implicit Methods

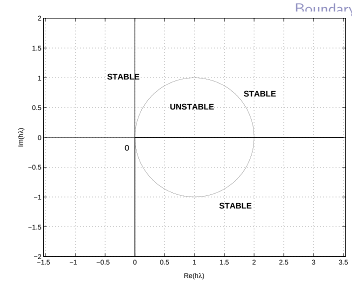



Figure: Stability region of backward Euler's method

How to solve $\mathbf{g}(\mathbf{y}_{n+1}) = \mathbf{y}_n + h\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}) - \mathbf{y}_{n+1} = \mathbf{0}$ for $\mathbf{y}_{n+1}$?
Typical Newton's iteration:

$$\mathbf{y}_{n+1}^{(k+1)} = \mathbf{y}_{n+1}^{(k)} + (\mathbf{I} - h\mathbf{J})^{-1}\left[\mathbf{y}_n - \mathbf{y}_{n+1}^{(k)} + h\mathbf{f}\left(x_{n+1}, \mathbf{y}_{n+1}^{(k)}\right)\right]$$

**Semi-implicit Euler's method** for local solution:

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h(\mathbf{I} - h\mathbf{J})^{-1}\mathbf{f}(x_{n+1}, \mathbf{y}_n)$$

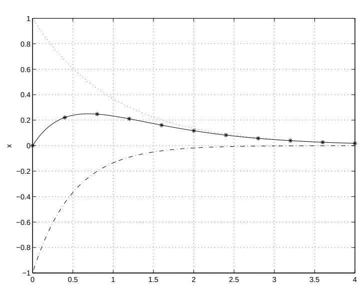


# Stiff Differential Equations

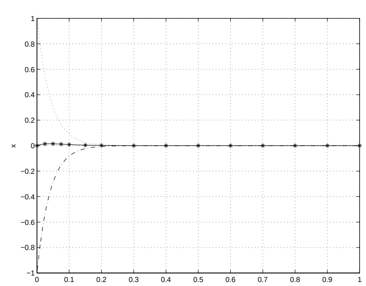
**Example:** IVP of a mass-spring-damper system:

$$\ddot{x} + c\dot{x} + kx = 0, \quad x(0) = 0, \; \dot{x}(0) = 1$$

(a) $c = 3, \; k = 2$: $\quad x = e^{-t} - e^{-2t}$
(b) $c = 49, \; k = 600$: $\quad x = e^{-24t} - e^{-25t}$



(a) Case of $c = 3, \; k = 2$



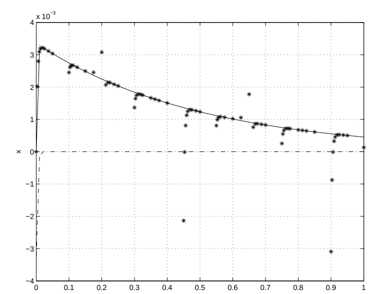
(b) Case of $c = 49, \; k = 600$

Figure: Solutions of a mass-spring-damper system: ordinary situations

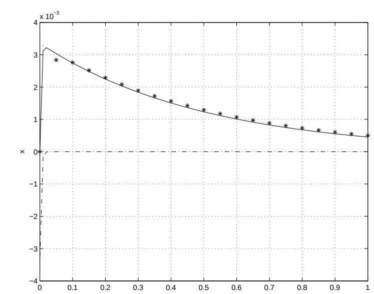


# Stiff Differential Equations

(c) $c = 302, \; k = 600$: $\quad x = \frac{e^{-2t} - e^{-300t}}{298}$



(c) With RK4



(d) With implicit Euler

Figure: Solutions of a mass-spring-damper system: stiff situation

To solve stiff ODE systems,

*use* **implicit method**, *preferably with explicit Jacobian.*

## Boundary Value Problems



A paradigm shift from the initial value problems

- A ball is thrown with a particular velocity. What trajectory does the ball follow?
- How to throw a ball such that it hits a particular window at a neighbouring house after 15 seconds?

**Two-point BVP in ODE's:**

*boundary conditions at two values of the independent variable*

Methods of solution

- Shooting method
- Finite difference (relaxation) method
- Finite element method



## Boundary Value Problems



**Shooting method**

*follows the strategy to adjust trials to hit a target.*

Consider the 2-point BVP

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{g}_1(\mathbf{y}(a)) = \mathbf{0}, \quad \mathbf{g}_2(\mathbf{y}(b)) = \mathbf{0},$$

where $\mathbf{g}_1 \in R^{n_1}$, $\mathbf{g}_2 \in R^{n_2}$ and $n_1 + n_2 = n$.

- Parametrize initial state: $\mathbf{y}(a) = \mathbf{h}(\mathbf{p})$ with $\mathbf{p} \in R^{n_2}$.
- Guess $n_2$ values of $\mathbf{p}$ to define IVP

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(a) = \mathbf{h}(\mathbf{p}).$$

- Solve this IVP for $[a, b]$ and evaluate $\mathbf{y}(b)$.
- Define error vector $\mathbf{E}(\mathbf{p}) = \mathbf{g}_2(\mathbf{y}(b))$.



## Boundary Value Problems



**Objective:** To solve $\mathbf{E}(\mathbf{p}) = \mathbf{0}$

From current vector $\mathbf{p}$, $n_2$ perturbations as $\mathbf{p} + \mathbf{e}_i\delta$: Jacobian $\frac{\partial \mathbf{E}}{\partial \mathbf{p}}$

*Each Newton's step: solution of $n_2 + 1$ initial value problems!*

- Computational cost
- Convergence not guaranteed (initial guess important)

Merits of shooting method

- Very few parameters to start
- In many cases, it is found quite efficient.



## Boundary Value Problems



**Finite difference (relaxation) method**

*adopts a* global *perspective.*

1. Discretize domain $[a, b]$: grid of points $a = x_0 < x_1 < x_2 < \cdots < x_{N-1} < x_N = b$. Function values $\mathbf{y}(x_i)$: $n(N+1)$ unknowns
2. Replace the ODE over intervals by *finite difference equations*. Considering mid-points, a typical (vector) FDE:
$$\mathbf{y}_i - \mathbf{y}_{i-1} - h\mathbf{f}\left(\frac{x_i + x_{i-1}}{2}, \frac{\mathbf{y}_i + \mathbf{y}_{i-1}}{2}\right) = \mathbf{0}, \quad \text{for } i = 1, 2, 3, \cdots, N$$
$nN$ (scalar) equations
3. Assemble additional $n$ equations from boundary conditions.
4. Starting from a guess solution over the grid, solve this system. (Sparse Jacobian is an advantage.)

Iterative schemes for solution of systems of linear equations.

## Points to note



- Numerical stability of ODE solution methods
- Computational cost versus better stability of implicit methods
- Multiscale responses leading to stiffness: failure of explicit methods
- Implicit methods for stiff systems
- Shooting method for two-point boundary value problems
- Relaxation method for boundary value problems

Necessary Exercises: **1,2,3,4,5**



## Outline







## Well-Posedness of Initial Value Problems



**Pierre Simon de Laplace (1749-1827):**

> *"We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes."*



## Well-Posedness of Initial Value Problems



**Initial value problem**

$$y' = f(x,y), \quad y(x_0) = y_0$$

From $(x,y)$, the trajectory develops according to $y' = f(x,y)$.

*The new point:* $(x + \delta x, y + f(x,y)\delta x)$
*The slope now:* $f(x + \delta x, y + f(x,y)\delta x)$

**Question:** Was the old direction of approach valid?

With $\delta x \to 0$, directions appropriate, if

$$\lim_{x \to \bar{x}} f(x,y) = f(\bar{x}, y(\bar{x})),$$

i.e. if $f(x,y)$ is **continuous**.

If $f(x,y) = \infty$, then $y' = \infty$ and trajectory is vertical.

*For the same value of $x$, several values of $y$!*

$y(x)$ **not** a function, unless $f(x,y) \neq \infty$, i.e. $f(x,y)$ is **bounded**.

## Well-Posedness of Initial Value Problems

**Peano's theorem:** If $f(x,y)$ is continuous and bounded in a rectangle $R = \{(x,y) : |x - x_0| < h, |y - y_0| < k\}$, with $|f(x,y)| \le M < \infty$, then the IVP $y' = f(x,y), \quad y(x_0) = y_0$ has a solution $y(x)$ defined in a neighbourhood of $x_0$.

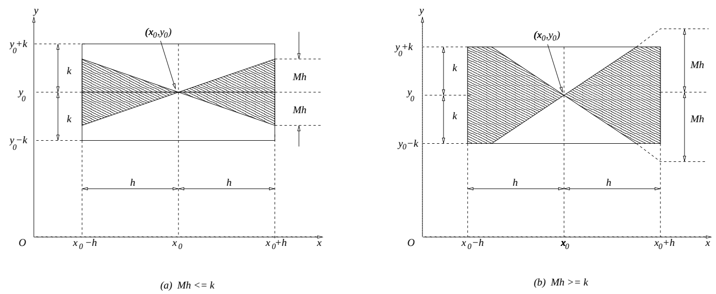



Figure: Regions containing the trajectories

Guaranteed neighbourhood:

$[x_0 - \delta, x_0 + \delta]$, *where* $\delta = \min(h, \frac{k}{M}) > 0$



## Well-Posedness of Initial Value Problems

**Example:**

$$y' = \frac{y-1}{x}, \quad y(0) = 1$$

Function $f(x,y) = \frac{y-1}{x}$ undefined at $(0,1)$.

*Premises of existence theorem not satisfied.*

But, premises here are **sufficient**, not *necessary*!

*Result inconclusive.*

The IVP has solutions: $y(x) = 1 + cx$ for all values of $c$.

*The solution is not unique.*

**Example:** $y'^2 = |y|, \quad y(0) = 0$

*Existence theorem guarantees a solution.*

But, there are **two** solutions:

$y(x) = 0$ *and* $y(x) = \ sgn(x)\ x^2/4$.



## Well-Posedness of Initial Value Problems

Physical system to mathematical model

- Mathematical solution
  - Interpretation about the physical system

Meanings of non-uniqueness of a solution

- Mathematical model admits of extraneous solution(s)?
- Physical system itself can exhibit alternative behaviours?

Indeterminacy of the solution

- Mathematical model of the system is not *complete*.

*The initial value problem is not* well-posed.

After existence, next important question:

Uniqueness of a solution



## Well-Posedness of Initial Value Problems

**Continuous dependence on initial condition**

Suppose that for IVP $y' = f(x,y), \quad y(x_0) = y_0$,

- unique solution: $y_1(x)$.

Applying a small perturbation to the initial condition, the new IVP: $y' = f(x,y), \quad y(x_0) = y_0 + \epsilon$

- unique solution: $y_2(x)$

**Question:** By how much $y_2(x)$ differs from $y_1(x)$ for $x > x_0$?

Large difference: solution *sensitive* to initial condition

- Practically unreliable solution

**Well-posed IVP:**

*An initial value problem is said to be well-posed if there exists a solution to it, the solution is unique and it depends continuously on the initial conditions.*

## Uniqueness Theorems

**Lipschitz condition:**

$$|f(x,y) - f(x,z)| \le L|y - z|$$

$L$: finite positive constant (Lipschitz constant)

> **Theorem:** *If $f(x,y)$ is a continuous function satisfying a Lipschitz condition on a strip $S = \{(x,y) : a < x < b, -\infty < y < \infty\}$, then for any point $(x_0, y_0) \in S$, the initial value problem of $y' = f(x,y), \quad y(x_0) = y_0$ is well-posed.*

Assume $y_1(x)$ and $y_2(x)$: solutions of the ODE $y' = f(x,y)$ with initial conditions $y(x_0) = (y_1)_0$ and $y(x_0) = (y_2)_0$

Consider $E(x) = [y_1(x) - y_2(x)]^2$.

$$E'(x) = 2(y_1 - y_2)(y_1' - y_2') = 2(y_1 - y_2)[f(x,y_1) - f(x,y_2)]$$

Applying Lipschitz condition,

$$|E'(x)| \le 2L(y_1 - y_2)^2 = 2LE(x).$$

Need to consider the case of $E'(x) > 0$ only.



## Uniqueness Theorems

$$\frac{E'(x)}{E(x)} \le 2L \;\Rightarrow\; \int_{x_0}^{x} \frac{E'(x)}{E(x)}\, dx \le 2L(x - x_0)$$

Integrating, $E(x) \le E(x_0)e^{2L(x-x_0)}$.

Hence,

$$|y_1(x) - y_2(x)| \le e^{L(x-x_0)}|(y_1)_0 - (y_2)_0|.$$

Since $x \in [a,b]$, $e^{L(x-x_0)}$ is finite.

$$|(y_1)_0 - (y_2)_0| = \epsilon \;\Rightarrow\; |y_1(x) - y_2(x)| \le e^{L(x-x_0)}\epsilon$$

continuous dependence of the solution on initial condition

In particular, $(y_1)_0 = (y_2)_0 = y_0 \;\Rightarrow\; y_1(x) = y_2(x) \;\forall\; x \in [a,b]$.

**The initial value problem is well-posed.**



## Uniqueness Theorems

A weaker theorem (hypotheses are stronger):

> **Picard's theorem:** *If $f(x,y)$ and $\frac{\partial f}{\partial y}$ are continuous and bounded on a rectangle $R = \{(x,y) : a < x < b, c < y < d\}$, then for every $(x_0, y_0) \in R$, the IVP $y' = f(x,y), \quad y(x_0) = y_0$ has a unique solution in some neighbourhood $|x - x_0| \le h$.*

From the mean value theorem,

$$f(x,y_1) - f(x,y_2) = \frac{\partial f}{\partial y}(\xi)(y_1 - y_2).$$

With Lipschitz constant $L = \sup\left|\frac{\partial f}{\partial y}\right|$,

*Lipschitz condition is satisfied 'lavishly'!*

**Note:** All these theorems give only *sufficient* conditions!

Hypotheses of Picard's theorem $\Rightarrow$ Lipschitz condition $\Rightarrow$ Well-posedness $\Rightarrow$ Existence and uniqueness



## Extension to ODE Systems

For ODE System

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x,\mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

- Lipschitz condition:

$$\|\mathbf{f}(x,\mathbf{y}) - \mathbf{f}(x,\mathbf{z})\| \le L\|\mathbf{y} - \mathbf{z}\|$$

- Scalar function $E(x)$ generalized as

$$E(x) = \|\mathbf{y}_1(x) - \mathbf{y}_2(x)\|^2 = (\mathbf{y}_1 - \mathbf{y}_2)^T(\mathbf{y}_1 - \mathbf{y}_2)$$

- Partial derivative $\frac{\partial f}{\partial y}$ replaced by the Jacobian $\mathbf{A} = \frac{\partial \mathbf{f}}{\partial \mathbf{y}}$
- Boundedness to be inferred from the boundedness of its norm

With these generalizations, the formulations work as usual.

## Extension to ODE Systems

**IVP of linear first order ODE system**

$$\mathbf{y}' = \mathbf{A}(x)\mathbf{y} + \mathbf{g}(x), \quad \mathbf{y}(x_0) = \mathbf{y}_0$$

Rate function: $\mathbf{f}(x, \mathbf{y}) = \mathbf{A}(x)\mathbf{y} + \mathbf{g}(x)$

*Continuity and boundedness of the coefficient functions in* $\mathbf{A}(x)$ *and* $\mathbf{g}(x)$ *are sufficient for well-posedness.*

**An $n$-th order linear ordinary differential equation**

$$y^{(n)}+P_1(x)y^{(n-1)}+P_2(x)y^{(n-2)}+\cdots+P_{n-1}(x)y'+P_n(x)y = R(x)$$

State vector: $\mathbf{z} = [y \quad y' \quad y'' \quad \cdots \quad y^{(n-1)}]^T$
With $z_1' = z_2$, $z_2' = z_3$, $\cdots$, $z_{n-1}' = z_n$ and $z_n'$ from the ODE,

- state space equation in the form $\mathbf{z}' = \mathbf{A}(x)\mathbf{z} + \mathbf{g}(x)$

*Continuity and boundedness of* $P_1(x), P_2(x), \cdots, P_n(x)$ *and* $R(x)$ *guarantees well-posedness.*



## Closure

A practical by-product of existence and uniqueness results:

- important results concerning the solutions

A sizeable segment of current research: *ill-posed* problems

- Dynamics of some nonlinear systems
  - **Chaos:** *sensitive dependence* on initial conditions

For boundary value problems,

*No general criteria for existence and uniqueness*

*Note:* Taking clue from the shooting method, a BVP in ODE's can be visualized as a complicated root-finding problem!

Multiple solutions or non-existence of solution is no surprise.



## Points to note

- For a solution of initial value problems, questions of existence, uniqueness and continuous dependence on initial condition are of crucial importance.
- These issues pertain to aspects of practical relevance regarding a physical system and its dynamic simulation
- Lipschitz condition is the tightest (avaliable) criterion for deciding these questions regarding well-posedness

Necessary Exercises: **1,2**



## Outline

## Formation of Differential Equations and Their Solutions



A differential equation represents a class of functions.

**Example:** $y(x) = cx^k$

With $\frac{dy}{dx} = ckx^{k-1}$ and $\frac{d^2y}{dx^2} = ck(k-1)x^{k-2}$,

$$xy\frac{d^2y}{dx^2} = x\left(\frac{dy}{dx}\right)^2 - y\frac{dy}{dx}$$

A compact 'intrinsic' description.

Important terms

- **Order** and **degree** of differential equations
- Homogeneous and non-homogeneous ODE's

Solution of a differential equation

- general, particular and singular solutions



## Separation of Variables



**ODE form with separable variables:**

$$y' = f(x,y) \;\Rightarrow\; \frac{dy}{dx} = \frac{\phi(x)}{\psi(y)} \;\text{ or }\; \psi(y)dy = \phi(x)dx$$

Solution as quadrature:

$$\int \psi(y)dy = \int \phi(x)dx + c.$$

**Separation of variables through substitution**

Example:

$$y' = g(\alpha x + \beta y + \gamma)$$

Substitute $v = \alpha x + \beta y + \gamma$ to arrive at

$$\frac{dv}{dx} = \alpha + \beta g(v) \;\Rightarrow\; x = \int \frac{dv}{\alpha + \beta g(v)} + c$$



## ODE's with Rational Slope Functions



$$y' = \frac{f_1(x,y)}{f_2(x,y)}$$

If $f_1$ and $f_2$ are homogeneous functions of $n$-th degree, then substitution $y = ux$ separates variables $x$ and $u$.

$$\frac{dy}{dx} = \frac{\phi_1(y/x)}{\phi_2(y/x)} \;\Rightarrow\; u + x\frac{du}{dx} = \frac{\phi_1(u)}{\phi_2(u)} \;\Rightarrow\; \frac{dx}{x} = \frac{\phi_2(u)}{\phi_1(u) - u\phi_2(u)}du$$

For $y' = \frac{a_1x+b_1y+c_1}{a_2x+b_2y+c_2}$, coordinate shift

$$x = X + h, \quad y = Y + k \;\Rightarrow\; y' = \frac{dy}{dx} = \frac{dY}{dX}$$

produces

$$\frac{dY}{dX} = \frac{a_1X + b_1Y + (a_1h + b_1k + c_1)}{a_2X + b_2Y + (a_2h + b_2k + c_2)}.$$

Choose $h$ and $k$ such that

$$a_1h + b_1k + c_1 = 0 = a_2h + b_2k + c_2.$$

If the system is inconsistent, then substitute $u = a_2x + b_2y$.



## Some Special ODE's



**Clairaut's equation**

$$y = xy' + f(y')$$

Substitute $p = y'$ and differentiate:

$$p = p + x\frac{dp}{dx} + f'(p)\frac{dp}{dx} \quad\Rightarrow\quad \frac{dp}{dx}[x + f'(p)] = 0$$

$\frac{dp}{dx} = 0$ means $y' = p = m$ (constant)

- family of straight lines $y = mx + f(m)$ as general solution

Singular solution:

$$x = -f'(p) \quad\text{and}\quad y = f(p) - pf'(p)$$

*Singular solution is the* envelope *of the family of straight lines that constitute the general solution.*

# Some Special ODE's



**Second order ODE's with the function not appearing explicitly**

$$f(x, y', y'') = 0$$

Substitute $y' = p$ and solve $f(x, p, p') = 0$ for $p(x)$.

**Second order ODE's with independent variable not appearing explicitly**

$$f(y, y', y'') = 0$$

Use $y' = p$ and

$$y'' = \frac{dp}{dx} = \frac{dp}{dy}\frac{dy}{dx} = p\frac{dp}{dy} \Rightarrow f(y, p, p\frac{dp}{dy}) = 0.$$

Solve for $p(y)$.
Resulting equation solved through a quadrature as

$$\frac{dy}{dx} = p(y) \Rightarrow x = x_0 + \int \frac{dy}{p(y)}.$$



# Exact Differential Equations and Reduction to the Exact For



$Mdx + Ndy$: an exact differential if

$$M = \frac{\partial \phi}{\partial x} \text{ and } N = \frac{\partial \phi}{\partial y}, \quad \text{or,} \quad \frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$$

$M(x, y)dx + N(x, y)dy = 0$ is an exact ODE if $\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}$
With $M(x, y) = \frac{\partial \phi}{\partial x}$ and $N(x, y) = \frac{\partial \phi}{\partial y}$,

$$\frac{\partial \phi}{\partial x}dx + \frac{\partial \phi}{\partial y}dy = 0 \Rightarrow d\phi = 0.$$

$$\boxed{\text{Solution: } \phi(x, y) = c}$$

Working rule:

$$\phi_1(x, y) = \int M(x, y)dx + g_1(y) \quad \text{and} \quad \phi_2(x, y) = \int N(x, y)dy + g_2(x)$$

Determine $g_1(y)$ and $g_2(x)$ from $\phi_1(x, y) = \phi_2(x, y) = \phi(x, y)$.
If $\frac{\partial M}{\partial y} \neq \frac{\partial N}{\partial x}$, but $\frac{\partial}{\partial y}(FM) = \frac{\partial}{\partial x}(FN)$?

$$\boxed{F\text{: Integrating factor}}$$



# First Order Linear (Leibnitz) ODE and Associated Forms



General first order linear ODE:

$$\frac{dy}{dx} + P(x)y = Q(x)$$

$$\boxed{\text{Leibnitz equation}}$$

For integrating factor $F(x)$,

$$F(x)\frac{dy}{dx} + F(x)P(x)y = \frac{d}{dx}[F(x)y] \Rightarrow \frac{dF}{dx} = F(x)P(x).$$

Separating variables,

$$\int \frac{dF}{F} = \int P(x)dx \Rightarrow \ln F = \int P(x)dx.$$

*Integrating factor:* $F(x) = e^{\int P(x)dx}$

$$ye^{\int P(x)dx} = \int Q(x)e^{\int P(x)dx}dx + C$$



# First Order Linear (Leibnitz) ODE and Associated Forms



**Bernoulli's equation**

$$\frac{dy}{dx} + P(x)y = Q(x)y^k$$

Substitution: $z = y^{1-k}$, $\frac{dz}{dx} = (1-k)y^{-k}\frac{dy}{dx}$ gives

$$\frac{dz}{dx} + (1-k)P(x)z = (1-k)Q(x),$$

in the Leibnitz form.

**Riccati equation**

$$y' = a(x) + b(x)y + c(x)y^2$$

If one solution $y_1(x)$ is known, then propose $y(x) = y_1(x) + z(x)$.

$$y_1'(x) + z'(x) = a(x) + b(x)[y_1(x) + z(x)] + c(x)[y_1(x) + z(x)]^2$$

Since $y_1'(x) = a(x) + b(x)y_1(x) + c(x)[y_1(x)]^2$,

$$z'(x) = [b(x) + 2c(x)y_1(x)]z(x) + c(x)[z(x)]^2,$$

in the form of Bernoulli's equation.

## Orthogonal Trajectories

In $xy$-plane, one-parameter equation $\phi(x, y, c) = 0$:

*a family of curves*

Differential equation of the family of curves:

$$\frac{dy}{dx} = f_1(x, y)$$

Slope of curves orthogonal to $\phi(x, y, c) = 0$:

$$\frac{dy}{dx} = -\frac{1}{f_1(x, y)}$$

Solving this ODE, another family of curves $\psi(x, y, k) = 0$.

Orthogonal trajectories

*If $\phi(x, y, c) = 0$ represents the potential lines (contours), then $\psi(x, y, k) = 0$ will represent the streamlines!*




## Points to note

- Meaning and solution of ODE's
- Separating variables
- Exact ODE's and integrating factors
- Linear (Leibnitz) equations
- Orthogonal families of curves

Necessary Exercises: **1,3,5,7**




## Outline






## Introduction

Second order ODE:

$$f(x, y, y', y'') = 0$$

Special case of a linear (non-homogeneous) ODE:

$$y'' + P(x)y' + Q(x)y = R(x)$$

Non-homogeneous linear ODE with constant coefficients:

$$y'' + ay' + by = R(x)$$

For $R(x) = 0$, linear homogeneous differential equation

$$y'' + P(x)y' + Q(x)y = 0$$

and linear homogeneous ODE with constant coefficients

$$y'' + ay' + by = 0$$

## Homogeneous Equations with Constant Coefficients



$$y'' + ay' + by = 0$$

Assume

$$y = e^{\lambda x} \quad \Rightarrow \quad y' = \lambda e^{\lambda x} \text{ and } y'' = \lambda^2 e^{\lambda x}.$$

Substitution: $(\lambda^2 + a\lambda + b)e^{\lambda x} = 0$

Auxiliary equation:

$$\lambda^2 + a\lambda + b = 0$$

Solve for $\lambda_1$ and $\lambda_2$:

*Solutions:* $e^{\lambda_1 x}$ *and* $e^{\lambda_2 x}$

Three cases

- Real and distinct ($a^2 > 4b$): $\lambda_1 \neq \lambda_2$

$$y(x) = c_1 y_1(x) + c_2 y_2(x) = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}$$



## Homogeneous Equations with Constant Coefficients

Introduction
Homogeneous Equations with Constant Coefficients
Euler-Cauchy Equation
Theory of the Homogeneous Equations
Basis for Solutions

- Real and equal ($a^2 = 4b$): $\lambda_1 = \lambda_2 = \lambda = -\frac{a}{2}$

  *only solution in hand:* $y_1 = e^{\lambda x}$

  Method to *develop* another solution?
  - Verify that $y_2 = xe^{\lambda x}$ is another solution.

  $$y(x) = c_1 y_1(x) + c_2 y_2(x) = (c_1 + c_2 x)e^{\lambda x}$$

- Complex conjugate ($a^2 < 4b$): $\lambda_{1,2} = -\frac{a}{2} \pm i\omega$

$$\begin{aligned} y(x) &= c_1 e^{(-\frac{a}{2}+i\omega)x} + c_2 e^{(-\frac{a}{2}-i\omega)x} \\ &= e^{-\frac{ax}{2}}[c_1(\cos\omega x + i\sin\omega x) + c_2(\cos\omega x - i\sin\omega x)] \\ &= e^{-\frac{ax}{2}}[A\cos\omega x + B\sin\omega x], \end{aligned}$$

with $A = c_1 + c_2$, $B = i(c_1 - c_2)$.
  - A third form: $y(x) = Ce^{-\frac{ax}{2}}\cos(\omega x - \alpha)$



## Euler-Cauchy Equation

Introduction
Homogeneous Equations with Constant Coefficients
Euler-Cauchy Equation
Theory of the Homogeneous Equations
Basis for Solutions

$$x^2 y'' + axy' + by = 0$$

Substituting $y = x^k$, auxiliary (or indicial) equation:

$$k^2 + (a-1)k + b = 0$$

1. Roots real and distinct $[(a-1)^2 > 4b]$: $k_1 \neq k_2$.

   $$y(x) = c_1 x^{k_1} + c_2 x^{k_2}.$$

2. Roots real and equal $[(a-1)^2 = 4b]$: $k_1 = k_2 = k = -\frac{a-1}{2}$.

   $$y(x) = (c_1 + c_2 \ln x)x^k.$$

3. Roots complex conjugate $[(a-1)^2 < 4b]$: $k_{1,2} = -\frac{a-1}{2} \pm i\nu$.

   $$y(x) = x^{-\frac{a-1}{2}}[A\cos(\nu \ln x) + B\sin(\nu \ln x)] = Cx^{-\frac{a-1}{2}}\cos(\nu \ln x - \alpha).$$

Alternative approach: substitution

$$x = e^t \;\Rightarrow\; t = \ln x,\; \frac{dx}{dt} = e^t = x \text{ and } \frac{dt}{dx} = \frac{1}{x}, \text{ etc.}$$



## Theory of the Homogeneous Equations

Introduction
Homogeneous Equations with Constant Coefficients
Euler-Cauchy Equation
Theory of the Homogeneous Equations
Basis for Solutions

$$y'' + P(x)y' + Q(x)y = 0$$

Well-posedness of its IVP:

*The initial value problem of the ODE, with arbitrary initial conditions* $y(x_0) = Y_0$, $y'(x_0) = Y_1$, *has a unique solution, as long as* $P(x)$ *and* $Q(x)$ *are continuous in the interval under question.*

*At least* two linearly independent solutions:

- $y_1(x)$: IVP with initial conditions $y(x_0) = 1$, $y'(x_0) = 0$
- $y_2(x)$: IVP with initial conditions $y(x_0) = 0$, $y'(x_0) = 1$

$$c_1 y_1(x) + c_2 y_2(x) = 0 \;\Rightarrow\; c_1 = c_2 = 0$$

*At most* two linearly independent solutions?

## Theory of the Homogeneous Equations

**Wronskian** of two solutions $y_1(x)$ and $y_2(x)$:

$$W(y_1, y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1 y_2' - y_2 y_1'$$

Solutions $y_1$ and $y_2$ are linearly dependent, if and only if $\exists\ x_0$ such that $W[y_1(x_0), y_2(x_0)] = 0$.

- $W[y_1(x_0), y_2(x_0)] = 0 \ \Rightarrow\ W[y_1(x), y_2(x)] = 0\ \forall x$.
- $W[y_1(x_1), y_2(x_1)] \neq 0 \ \Rightarrow\ W[y_1(x), y_2(x)] \neq 0\ \forall x$, and $y_1(x)$ and $y_2(x)$ are linearly independent solutions.

**Complete solution:**

*If $y_1(x)$ and $y_2(x)$ are two linearly independent solutions, then the general solution is*

$$y(x) = c_1 y_1(x) + c_2 y_2(x).$$

And, the general solution is the complete solution.

No third linearly independent solution. No singular solution.



## Theory of the Homogeneous Equations

If $y_1(x)$ and $y_2(x)$ are linearly dependent, then $y_2 = ky_1$.

$$W(y_1, y_2) = y_1 y_2' - y_2 y_1' = y_1(ky_1') - (ky_1)y_1' = 0$$

In particular, $W[y_1(x_0), y_2(x_0)] = 0$

Conversely, if there is a value $x_0$, where

$$W[y_1(x_0), y_2(x_0)] = \begin{vmatrix} y_1(x_0) & y_2(x_0) \\ y_1'(x_0) & y_2'(x_0) \end{vmatrix} = 0,$$

then for

$$\begin{bmatrix} y_1(x_0) & y_2(x_0) \\ y_1'(x_0) & y_2'(x_0) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{0},$$

coefficient matrix is singular.

Choose non-zero $\begin{bmatrix} c_1 \\ c_2 \end{bmatrix}$ and frame $y(x) = c_1 y_1 + c_2 y_2$, satisfying

$$IVP \quad y'' + Py' + Qy = 0,\ y(x_0) = 0,\ y'(x_0) = 0.$$

Therefore, $y(x) = 0 \ \Rightarrow$ $y_1$ and $y_2$ are linearly dependent.



## Theory of the Homogeneous Equations

Pick a candidate solution $Y(x)$, choose a point $x_0$, evaluate functions $y_1$, $y_2$, $Y$ and their derivatives at that point, frame

$$\begin{bmatrix} y_1(x_0) & y_2(x_0) \\ y_1'(x_0) & y_2'(x_0) \end{bmatrix} \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} Y(x_0) \\ Y'(x_0) \end{bmatrix}$$

and ask for solution $\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}$.

Unique solution for $C_1, C_2$. Hence, particular solution

$$y^*(x) = C_1 y_1(x) + C_2 y_2(x)$$

is the "unique" solution of the IVP

$$y'' + Py' + Qy = 0,\ y(x_0) = Y(x_0),\ y'(x_0) = Y'(x_0).$$

But, that is the candidate function $Y(x)$! Hence, $Y(x) = y^*(x)$.



## Basis for Solutions

For completely describing the solutions, we need

**two** *linearly independent solutions.*

No guaranteed procedure to identify two basis members!

If one solution $y_1(x)$ is available, then to find another?

**Reduction of order**

Assume the second solution as

$$y_2(x) = u(x) y_1(x)$$

and determine $u(x)$ such that $y_2(x)$ satisfies the ODE.

$$u'' y_1 + 2u' y_1' + u y_1'' + P(u' y_1 + u y_1') + Q u y_1 = 0$$

$$\Rightarrow\ u'' y_1 + 2u' y_1' + P u' y_1 + u(y_1'' + P y_1' + Q y_1) = 0.$$

Since $y_1'' + P y_1' + Q y_1 = 0$, we have $y_1 u'' + (2y_1' + P y_1) u' = 0$

## Basis for Solutions

Denoting $u' = U$, $U' + (2\frac{y_1'}{y_1} + P)U = 0$.

Rearrangement and integration of the reduced equation:

$$\frac{dU}{U} + 2\frac{dy_1}{y_1} + Pdx = 0 \Rightarrow Uy_1^2 e^{\int Pdx} = C = 1 \text{ (choose)}.$$

Then,

$$u' = U = \frac{1}{y_1^2} e^{-\int Pdx},$$

Integrating,

$$u(x) = \int \frac{1}{y_1^2} e^{-\int Pdx} dx,$$

and

$$y_2(x) = y_1(x) \int \frac{1}{y_1^2} e^{-\int Pdx} dx.$$

**Note:** The factor $u(x)$ is never constant!



## Basis for Solutions

**Function space perspective:**

Operator 'D' means differentiation, operates on an *infinite dimensional* function space as a linear transformation.

- It maps all constant functions to zero.
  - It has a one-dimensional null space.

Second derivative or $D^2$ is an operator that has a two-dimensional null space, $c_1 + c_2 x$, with basis $\{1, x\}$.

Examples of composite operators

- $(D + a)$ has a null space $ce^{-ax}$.
- $(xD + a)$ has a null space $cx^{-a}$.

A second order linear operator $D^2 + P(x)D + Q(x)$ possesses a two-dimensional null space.

- Solution of $[D^2 + P(x)D + Q(x)]y = 0$: description of the null space, or a basis for it..
- Analogous to solution of $\mathbf{Ax} = \mathbf{0}$, i.e. development of a basis for $Null(\mathbf{A})$.



## Points to note

- Second order linear homogeneous ODE's
- Wronskian and related results
- Solution basis
- Reduction of order
- Null space of a differential operator

Necessary Exercises: **1,2,3,7,8**



## Outline

# Linear ODE's and Their Solutions

**The Complete Analogy**

Table: Linear systems and mappings: algebraic and differential

| **In ordinary vector space** | **In infinite-dimensional function space** |
|---|---|
| $\mathbf{Ax} = \mathbf{b}$ | $y'' + Py' + Qy = R$ |
| The system is consistent. | $P(x)$, $Q(x)$, $R(x)$ are continuous. |
| A solution $\mathbf{x}^*$ | A solution $y_p(x)$ |
| Alternative solution: $\bar{\mathbf{x}}$ | Alternative solution: $\bar{y}(x)$ |
| $\bar{\mathbf{x}} - \mathbf{x}^*$ satisfies $\mathbf{Ax} = \mathbf{0}$, is in null space of $\mathbf{A}$. | $\bar{y}(x) - y_p(x)$ satisfies $y'' + Py' + Qy = 0$, is in null space of $D^2 + P(x)D + Q(x)$. |
| Complete solution: $\mathbf{x} = \mathbf{x}^* + \sum_i c_i(\mathbf{x}_0)_i$ | Complete solution: $y_p(x) + \sum_i c_i y_i(x)$ |
| Methodology: Find null space of $\mathbf{A}$ i.e. basis members $(\mathbf{x}_0)_i$. Find $\mathbf{x}^*$ and compose. | Methodology: Find null space of $D^2 + P(x)D + Q(x)$ i.e. basis members $y_i(x)$. Find $y_p(x)$ and compose. |



# Linear ODE's and Their Solutions

Procedure to solve $y'' + P(x)y' + Q(x)y = R(x)$

1. First, solve the corresponding homogeneous equation, obtain a basis with two solutions and construct
$$y_h(x) = c_1 y_1(x) + c_2 y_2(x).$$
2. Next, find one particular solution $y_p(x)$ of the NHE and compose the complete solution
$$y(x) = y_h(x) + y_p(x) = c_1 y_1(x) + c_2 y_2(x) + y_p(x).$$
3. If some initial or boundary conditions are known, they can be imposed *now* to determine $c_1$ and $c_2$.

**Caution:** If $y_1$ and $y_2$ are two solutions of the NHE, then **do not expect** $c_1 y_1 + c_2 y_2$ to satisfy the equation.

**Implication of linearity or superposition:**

*With zero initial conditions, if $y_1$ and $y_2$ are responses due to inputs $R_1(x)$ and $R_2(x)$, respectively, then the response due to input $c_1 R_1 + c_2 R_2$ is $c_1 y_1 + c_2 y_2$.*



# Method of Undetermined Coefficients

$$y'' + ay' + by = R(x)$$

- What kind of function to propose as $y_p(x)$ if $R(x) = x^n$?
- And what if $R(x) = e^{\lambda x}$?
- If $R(x) = x^n + e^{\lambda x}$, i.e. in the form $k_1 R_1(x) + k_2 R_2(x)$?

*The principle of superposition (linearity)*

Table: Candidate solutions for linear non-homogeneous ODE's

| **RHS function** $R(x)$ | **Candidate solution** $y_p(x)$ |
|---|---|
| $p_n(x)$ | $q_n(x)$ |
| $e^{\lambda x}$ | $ke^{\lambda x}$ |
| $\cos\omega x$ or $\sin\omega x$ | $k_1\cos\omega x + k_2\sin\omega x$ |
| $e^{\lambda x}\cos\omega x$ or $e^{\lambda x}\sin\omega x$ | $k_1 e^{\lambda x}\cos\omega x + k_2 e^{\lambda x}\sin\omega x$ |
| $p_n(x)e^{\lambda x}$ | $q_n(x)e^{\lambda x}$ |
| $p_n(x)\cos\omega x$ or $p_n(x)\sin\omega x$ | $q_n(x)\cos\omega x + r_n(x)\sin\omega x$ |
| $p_n(x)e^{\lambda x}\cos\omega x$ or $p_n(x)e^{\lambda x}\sin\omega x$ | $q_n(x)e^{\lambda x}\cos\omega x + r_n(x)e^{\lambda x}\sin\omega x$ |



# Method of Undetermined Coefficients

**Example:**

(a) $y'' - 6y' + 5y = e^{3x}$
(b) $y'' - 5y' + 6y = e^{3x}$
(c) $y'' - 6y' + 9y = e^{3x}$

In each case, the first official proposal: $y_p = ke^{3x}$

(a) $y(x) = c_1 e^x + c_2 e^{5x} - e^{3x}/4$
(b) $y(x) = c_1 e^{2x} + c_2 e^{3x} + \; xe^{3x}$
(c) $y(x) = c_1 e^{3x} + c_2 xe^{3x} + \; \frac{1}{2}x^2 e^{3x}$

Modification rule

- If the candidate function ($ke^{\lambda x}$, $k_1\cos\omega x + k_2\sin\omega x$ or $k_1 e^{\lambda x}\cos\omega x + k_2 e^{\lambda x}\sin\omega x$) is a solution of the corresponding HE; with $\lambda$, $\pm i\omega$ or $\lambda \pm i\omega$ (respectively) satisfying the auxiliary equation; then modify it by multiplying with $x$.
- In the case of $\lambda$ being a double root, i.e. both $e^{\lambda x}$ and $xe^{\lambda x}$ being solutions of the HE, choose $y_p = kx^2 e^{\lambda x}$.

## Method of Variation of Parameters



Solution of the HE:

$$y_h(x) = c_1 y_1(x) + c_2 y_2(x),$$

in which $c_1$ and $c_2$ are constant 'parameters'.

For solution of the NHE,

*how about 'variable parameters'?*

Propose

$$y_p(x) = u_1(x)y_1(x) + u_2(x)y_2(x)$$

and force $y_p(x)$ to satisfy the ODE.

*A* single *second order ODE in $u_1(x)$ and $u_2(x)$.*
*We need one more condition to* fix *them.*



## Method of Variation of Parameters



From $y_p = u_1 y_1 + u_2 y_2$,

$$y_p' = u_1' y_1 + u_1 y_1' + u_2' y_2 + u_2 y_2'.$$

Condition $\boxed{u_1' y_1 + u_2' y_2 = 0}$ gives

$$y_p' = u_1 y_1' + u_2 y_2'.$$

Differentiating,

$$y_p'' = u_1' y_1' + u_2' y_2' + u_1 y_1'' + u_2 y_2''.$$

Substitution into the ODE:

$$u_1' y_1' + u_2' y_2' + u_1 y_1'' + u_2 y_2'' + P(x)(u_1 y_1' + u_2 y_2') + Q(x)(u_1 y_1 + u_2 y_2) = R(x)$$

Rearranging,

$$u_1' y_1' + u_2' y_2' + u_1(y_1'' + P(x)y_1' + Q(x)y_1) + u_2(y_2'' + P(x)y_2' + Q(x)y_2) = R(x).$$

As $y_1$ and $y_2$ satisfy the associated HE, $\boxed{u_1' y_1' + u_2' y_2' = R(x)}$



## Method of Variation of Parameters



$$\begin{bmatrix} y_1 & y_2 \\ y_1' & y_2' \end{bmatrix} \begin{bmatrix} u_1' \\ u_2' \end{bmatrix} = \begin{bmatrix} 0 \\ R \end{bmatrix}$$

Since Wronskian is non-zero, this system has unique solution

$$u_1' = -\frac{y_2 R}{W} \quad \text{and} \quad u_2' = \frac{y_1 R}{W}.$$

Direct quadrature:

$$u_1(x) = -\int \frac{y_2(x)R(x)}{W[y_1(x), y_2(x)]}dx \quad \text{and} \quad u_2(x) = \int \frac{y_1(x)R(x)}{W[y_1(x), y_2(x)]}dx$$

*In contrast to the method of undetermined multipliers, variation of parameters is* **general**. *It is applicable for all continuous functions as $P(x)$, $Q(x)$ and $R(x)$.*



## Points to note



- Function space perspective of linear ODE's
- Method of undetermined coefficients
- Method of variation of parameters

Necessary Exercises: **1,3,5,6**

# Outline







# Theory of Linear ODE's



$$y^{(n)}+P_1(x)y^{(n-1)}+P_2(x)y^{(n-2)}+\cdots+P_{n-1}(x)y'+P_n(x)y = R(x)$$

General solution: $y(x) = y_h(x) + y_p(x)$, where

- $y_p(x)$: a particular solution
- $y_h(x)$: general solution of corresponding HE

$$y^{(n)}+P_1(x)y^{(n-1)}+P_2(x)y^{(n-2)}+\cdots+P_{n-1}(x)y'+P_n(x)y = 0$$

For the HE, suppose we have $n$ solutions $y_1(x)$, $y_2(x)$, $\cdots$, $y_n(x)$.
Assemble the state vectors in matrix

$$\mathbf{Y}(x) = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ y_1'' & y_2'' & \cdots & y_n'' \\ \vdots & \vdots & \ddots & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{bmatrix}.$$

**Wronskian:**

$$W(y_1, y_2, \cdots, y_n) = \det[\mathbf{Y}(x)]$$



# Theory of Linear ODE's



- If solutions $y_1(x)$, $y_2(x)$, $\cdots$, $y_n(x)$ of HE are linearly dependent, then for a non-zero $\mathbf{k} \in R^n$,

$$\sum_{i=1}^{n} k_i y_i(x) = 0 \;\Rightarrow\; \sum_{i=1}^{n} k_i y_i^{(j)}(x) = 0 \text{ for } j = 1,2,3,\cdots,(n-1)$$
$$\Rightarrow [\mathbf{Y}(x)]\mathbf{k} = \mathbf{0} \Rightarrow [\mathbf{Y}(x)] \text{ is singular},$$
$$\Rightarrow W[y_1(x), y_2(x), \cdots, y_n(x)] = 0.$$

- If Wronskian is zero at $x = x_0$, then $\mathbf{Y}(x_0)$ is singular and a non-zero $\mathbf{k} \in Null[\mathbf{Y}(x_0)]$ gives $\sum_{i=1}^{n} k_i y_i(x) = 0$, implying $y_1(x)$, $y_2(x)$, $\cdots$, $y_n(x)$ to be linearly dependent.
- Zero Wronskian at some $x = x_0$ implies zero Wronskian everywhere. Non-zero Wronskian at some $x = x_1$ ensures non-zero Wronskian everywhere and the corrseponding solutions as linearly independent.
- With $n$ linearly independent solutions $y_1(x)$, $y_2(x)$, $\cdots$, $y_n(x)$ of the HE, we have its general solution $y_h(x) = \sum_{i=1}^{n} c_i y_i(x)$, acting as the *complementary function* for the NHE.



# Homogeneous Equations with Constant Coefficients



$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_{n-1} y' + a_n y = 0$$

With trial solution $y = e^{\lambda x}$, the auxiliary equation:

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

Construction of the basis:

1. For every simple real root $\lambda = \gamma$, $e^{\gamma x}$ is a solution.
2. For every simple pair of complex roots $\lambda = \mu \pm i\omega$, $e^{\mu x}\cos\omega x$ and $e^{\mu x}\sin\omega x$ are linearly independent solutions.
3. For every real root $\lambda = \gamma$ of multiplicity $r$; $e^{\gamma x}$, $xe^{\gamma x}$, $x^2 e^{\gamma x}$, $\cdots$, $x^{r-1}e^{\gamma x}$ are all linearly independent solutions.
4. For every complex pair of roots $\lambda = \mu \pm i\omega$ of multiplicity $r$; $e^{\mu x}\cos\omega x$, $e^{\mu x}\sin\omega x$, $xe^{\mu x}\cos\omega x$, $xe^{\mu x}\sin\omega x$, $\cdots$, $x^{r-1}e^{\mu x}\cos\omega x$, $x^{r-1}e^{\mu x}\sin\omega x$ are the required solutions.

## Non-Homogeneous Equations

**Method of undetermined coefficients**

$$y^{(n)} + a_1 y^{(n-1)} + a_2 y^{(n-2)} + \cdots + a_{n-1} y' + a_n y = R(x)$$

Extension of the second order case

**Method of variation of parameters**

$$y_p(x) = \sum_{i=1}^{n} u_i(x) y_i(x)$$

| **Imposed condition** | | **Derivative** |
|---|---|---|
| $\sum_{i=1}^n u_i'(x) y_i(x) = 0$ | $\Rightarrow$ | $y_p'(x) = \sum_{i=1}^n u_i(x) y_i'(x)$ |
| $\sum_{i=1}^n u_i'(x) y_i'(x) = 0$ | $\Rightarrow$ | $y_p''(x) = \sum_{i=1}^n u_i(x) y_i''(x)$ |
| $\cdots \quad \cdots \quad \cdots$ | $\Rightarrow$ | $\cdots \quad \cdots \quad \cdots$ |
| $\sum_{i=1}^n u_i'(x) y_i^{(n-2)}(x) = 0$ | $\Rightarrow$ | $y_p^{(n-1)}(x) = \sum_{i=1}^n u_i(x) y_i^{(n-1)}(x)$ |

Finally, $y_p^{(n)}(x) = \sum_{i=1}^n u_i'(x) y_i^{(n-1)}(x) + \sum_{i=1}^n u_i(x) y_i^{(n)}(x)$

$$\Rightarrow \sum_{i=1}^{n} u_i'(x) y_i^{(n-1)}(x) + \sum_{i=1}^{n} u_i(x) \left[ y_i^{(n)} + P_1 y_i^{(n-1)} + \cdots + P_n y_i \right] = R(x)$$



## Non-Homogeneous Equations

Since each $y_i(x)$ is a solution of the HE,

$$\sum_{i=1}^{n} u_i'(x) y_i^{(n-1)}(x) = R(x).$$

Assembling all conditions on $\mathbf{u}'(x)$ together,

$$[\mathbf{Y}(x)]\mathbf{u}'(x) = \mathbf{e}_n R(x).$$

Since $\mathbf{Y}^{-1} = \frac{\mathrm{adj}\ \mathbf{Y}}{\det(\mathbf{Y})}$,

$$\mathbf{u}'(x) = \frac{1}{\det[\mathbf{Y}(x)]}[\mathrm{adj}\ \mathbf{Y}(x)]\mathbf{e}_n R(x) = \frac{R(x)}{W(x)}[\text{last column of } \mathrm{adj}\ \mathbf{Y}(x)].$$

Using cofactors of elements from last row only,

$$u_i'(x) = \frac{W_i(x)}{W(x)} R(x),$$

with $W_i(x)$ = Wronskian evaluated with $\mathbf{e}_n$ in place of $i$-th column.

$$\boxed{u_i(x) = \int \frac{W_i(x) R(x)}{W(x)} dx}$$



## Points to note

- Wronskian for a higher order ODE
- General theory of linear ODE's
  - Variation for parameters for $n$-th order ODE

Necessary Exercises: **1,3,4**



## Outline

## Introduction



Classical perspective

- Entire differential equation is known in advance.
- Go for a complete solution first.
- Afterwards, use the initial (or other) conditions.

A practical situation

- You have a plant
  - intrinsic dynamic model as well as the starting conditions.
- You may drive the plant with different kinds of inputs on different occasions.

Implication

- Left-hand side of the ODE and the initial conditions are known *a priori*.
- Right-hand side, $R(x)$, changes from task to task.



## Introduction



Another question: What if $R(x)$ is *not* continuous?

- When power is switched on or off, what happens?
- If there is a sudden voltage fluctuation, what happens to the equipment connected to the power line?

Or, does "anything" happen in the immediate future?
"Something" certainly happens. The IVP has a solution!

> *Laplace transforms provide a tool to* find *the solution, in spite of the discontinuity of* $R(x)$.

**Integral transform:**

$$T[f(t)](s) = \int_a^b K(s,t)f(t)dt$$

$s$: frequency variable
$K(s,t)$: kernel of the transform

**Note:** $T[f(t)]$ is a function of $s$, not $t$.



## Introduction



With kernel function $K(s,t) = e^{-st}$, and limits $a = 0$, $b = \infty$,
**Laplace transform**

$$F(s) = L\{f(t)\} = \int_0^\infty e^{-st} f(t)dt = \lim_{b\to\infty} \int_0^b e^{-st} f(t)dt$$

When this integral exists, $f(t)$ has its Laplace transform.

Sufficient condition:

- $f(t)$ is piecewise continuous, and
- it is of exponential order, i.e. $|f(t)| < Me^{ct}$ for some (finite) $M$ and $c$.

Inverse Laplace transform:

$$f(t) = L^{-1}\{F(s)\}$$



## Basic Properties and Results



Linearity:

$$L\{af(t) + bg(t)\} = aL\{f(t)\} + bL\{g(t)\}$$

First shifting property or the frequency shifting rule:

$$L\{e^{at}f(t)\} = F(s-a)$$

Laplace transforms of some elementary functions:

$$\begin{aligned}
L(1) &= \int_0^\infty e^{-st}dt = \left[\frac{e^{-st}}{-s}\right]_0^\infty = \frac{1}{s}, \\
L(t) &= \int_0^\infty e^{-st}tdt = \left[t\frac{e^{-st}}{-s}\right]_0^\infty + \frac{1}{s}\int_0^\infty e^{-st}dt = \frac{1}{s^2}, \\
L(t^n) &= \frac{n!}{s^{n+1}} \quad \text{(for positive integer } n\text{)}, \\
L(t^a) &= \frac{\Gamma(a+1)}{s^{a+1}} \quad \text{(for } a \in R^+\text{)} \\
\text{and } L(e^{at}) &= \frac{1}{s-a}.
\end{aligned}$$

## Basic Properties and Results

$$L(\cos\omega t) = \frac{s}{s^2+\omega^2}, \qquad L(\sin\omega t) = \frac{\omega}{s^2+\omega^2};$$
$$L(\cosh at) = \frac{s}{s^2-a^2}, \qquad L(\sinh at) = \frac{a}{s^2-a^2};$$
$$L(e^{\mu t}\cos\omega t) = \frac{s-\mu}{(s-\mu)^2+\omega^2}, \quad L(e^{\mu t}\sin\omega t) = \frac{\omega}{(s-\mu)^2+\omega^2}.$$

Laplace transform of derivative:

$$\begin{aligned} L\{f'(t)\} &= \int_0^\infty e^{-st} f'(t)dt \\ &= \left[e^{-st}f(t)\right]_0^\infty + s\int_0^\infty e^{-st}f(t)dt \;=\; sL\{f(t)\} - f(0) \end{aligned}$$

Using this process recursively,

$$L\{f^{(n)}(t)\} = s^n L\{f(t)\} - s^{(n-1)}f(0) - s^{(n-2)}f'(0) - \cdots - f^{(n-1)}(0).$$

For integral $g(t) = \int_0^t f(t)dt$, $g(0) = 0$, and
$L\{g'(t)\} = sL\{g(t)\} - g(0) = sL\{g(t)\} \;\Rightarrow\; L\{g(t)\} = \frac{1}{s}L\{f(t)\}$.



## Application to Differential Equations

**Example:**
Initial value problem of a linear constant coefficient ODE

$$y'' + ay' + by = r(t), \quad y(0) = K_0, \; y'(0) = K_1$$

Laplace transforms of both sides of the ODE:

$$\begin{aligned} s^2Y(s) - sy(0) - y'(0) \; + \; a[sY(s) - y(0)] + bY(s) &= R(s) \\ \Rightarrow \; (s^2+as+b)Y(s) &= (s+a)K_0 + K_1 + R(s) \end{aligned}$$

*A differential equation in $y(t)$ has been converted to an algebraic equation in $Y(s)$.*

**Transfer function:** ratio of Laplace transform of output function $y(t)$ to that of input function $r(t)$, with zero initial conditions

$$Q(s) = \frac{Y(s)}{R(s)} = \frac{1}{s^2+as+b} \quad \text{(in this case)}$$
$$Y(s) = [(s+a)K_0 + K_1]Q(s) + Q(s)R(s)$$

Solution of the given IVP: $y(t) = L^{-1}\{Y(s)\}$



## Handling Discontinuities

**Unit step function**

$$u(t-a) = \begin{cases} 0 & \text{if} \quad t < a \\ 1 & \text{if} \quad t > a \end{cases}$$

Its Laplace transform:

$$L\{u(t-a)\} = \int_0^\infty e^{-st}u(t-a)dt = \int_0^a 0\cdot dt + \int_a^\infty e^{-st}dt = \frac{e^{-as}}{s}$$

For input $f(t)$ with a time delay,

$$f(t-a)u(t-a) = \begin{cases} 0 & \text{if} \quad t < a \\ f(t-a) & \text{if} \quad t > a \end{cases}$$

has its Laplace transform as

$$\begin{aligned} L\{f(t-a)u(t-a)\} &= \int_a^\infty e^{-st}f(t-a)dt \\ &= \int_0^\infty e^{-s(a+\tau)}f(\tau)d\tau = e^{-as}L\{f(t)\}. \end{aligned}$$

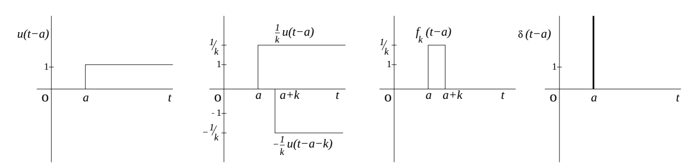
Second shifting property or the time shifting rule



## Handling Discontinuities

Define

$$\begin{aligned} f_k(t-a) &= \begin{cases} 1/k & \text{if} \quad a \le t \le a+k \\ 0 & \text{otherwise} \end{cases} \\ &= \frac{1}{k}u(t-a) - \frac{1}{k}u(t-a-k) \end{aligned}$$




(a) Unit step function (b) Composition (c) Function $f_k$ (d) Dirac's δ − function

Figure: Step and impulse functions

and note that its integral

$$I_k = \int_0^\infty f_k(t-a)dt = \int_a^{a+k}\frac{1}{k}dt = 1.$$

does not depend on $k$.

## Handling Discontinuities

In the limit,

$$\delta(t-a) = \lim_{k\to 0} f_k(t-a)$$

$$\text{or,}\quad \delta(t-a) = \begin{cases} \infty & \text{if} \quad t=a \\ 0 & \text{otherwise} \end{cases} \quad \textbf{and} \quad \int_0^\infty \delta(t-a)dt = 1.$$

*Unit impulse function* or *Dirac's delta function*

$$L\{\delta(t-a)\} = \lim_{k\to 0} \frac{1}{k}[L\{u(t-a)\} - L\{u(t-a-k)\}] = \lim_{k\to 0} \frac{e^{-as} - e^{-(a+k)s}}{ks} = e^{-as}$$

*Through step and impulse functions, Laplace transform method can handle IVP's with discontinuous inputs.*



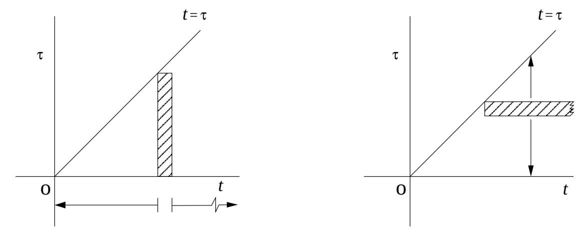
## Convolution

A *generalized product* of two functions

$$h(t) = f(t) * g(t) = \int_0^t f(\tau) g(t-\tau)\, d\tau$$

Laplace transform of the convolution:

$$H(s) = \int_0^\infty e^{-st} \int_0^t f(\tau) g(t-\tau) d\tau\, dt = \int_0^\infty f(\tau) \int_\tau^\infty e^{-st} g(t-\tau) dt\, d\tau$$




(a) Original order (b) Changed order

Figure: Region of integration for $L\{h(t)\}$



## Convolution

Through substitution $t' = t - \tau$,

$$H(s) = \int_0^\infty f(\tau) \int_0^\infty e^{-s(t'+\tau)} g(t')\, dt'\, d\tau = \int_0^\infty f(\tau) e^{-s\tau} \left[ \int_0^\infty e^{-st'} g(t')\, dt' \right] d\tau$$

$$\boxed{H(s) = F(s)G(s)}$$

**Convolution theorem:**

*Laplace transform of the convolution integral of two functions is given by the product of the Laplace transforms of the two functions.*

Utilities:

- To invert $Q(s)R(s)$, one can convolute $y(t) = q(t) * r(t)$.
- In solving some integral equation.



## Points to note

- A paradigm shift in solution of IVP's
- Handling discontinuous input functions
- Extension to ODE systems
- The idea of integral transforms

Necessary Exercises: **1,2,4**

# Outline







# Fundamental Ideas



$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y})$$

Solution: a vector function $\mathbf{y} = \mathbf{h}(t)$

Autonomous system: $\mathbf{y}' = \mathbf{f}(\mathbf{y})$

- Points in $\mathbf{y}$-space where $\mathbf{f}(\mathbf{y}) = 0$:

  *equilibrium points* or *critical points*

System of linear ODE's:

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{g}(t)$$

- *autonomous systems* if $\mathbf{A}$ and $\mathbf{g}$ are constant
- *homogeneous systems* if $\mathbf{g}(t) = 0$
- *homogeneous constant coefficient systems* if $\mathbf{A}$ is constant and $\mathbf{g}(t) = 0$



# Fundamental Ideas



For a homogeneous system,

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$$

- Wronskian: $W(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \cdots, \mathbf{y}_n) = |\mathbf{y}_1 \;\; \mathbf{y}_2 \;\; \mathbf{y}_3 \;\; \cdots \;\; \mathbf{y}_n|$

If Wronskian is non-zero, then

- Fundamental matrix: $\mathcal{Y}(t) = [\mathbf{y}_1 \;\; \mathbf{y}_2 \;\; \mathbf{y}_3 \;\; \cdots \;\; \mathbf{y}_n]$,

giving a basis.

General solution:

$$\mathbf{y}(t) = \sum_{i=1}^{n} c_i \mathbf{y}_i(t) = [\mathcal{Y}(t)]\,\mathbf{c}$$



# Linear Homogeneous Systems with Constant Coefficients



$$\mathbf{y}' = \mathbf{A}\mathbf{y}$$

Non-degenerate case: matrix $\mathbf{A}$ non-singular

- Origin ($\mathbf{y} = \mathbf{0}$) is the unique equilibrium point.

Attempt $\mathbf{y} = \mathbf{x}e^{\lambda t} \;\Rightarrow\; \mathbf{y}' = \lambda \mathbf{x} e^{\lambda t}$.

Substitution: $\mathbf{A}\mathbf{x}e^{\lambda t} = \lambda \mathbf{x} e^{\lambda t} \;\Rightarrow\; \boxed{\mathbf{A}\mathbf{x} = \lambda \mathbf{x}}$

If $\mathbf{A}$ is diagonalizable,

- $n$ linearly independent solutions $\mathbf{y}_i = \mathbf{x}_i e^{\lambda_i t}$ corresponding to $n$ eigenpairs

If $\mathbf{A}$ is *not* diagonalizable?

*All $\mathbf{x}_i e^{\lambda_i t}$ together will not complete the basis.*

Try $\mathbf{y} = \mathbf{x}te^{\mu t}$? Substitution leads to

$$\mathbf{x}e^{\mu t} + \mu \mathbf{x} t e^{\mu t} = \mathbf{A}\mathbf{x}te^{\mu t} \;\Rightarrow\; \mathbf{x}e^{\mu t} = \mathbf{0} \;\Rightarrow\; \mathbf{x} = \mathbf{0}.$$

Absurd!

## Linear Homogeneous Systems with Constant Coefficients



Try a linearly independent solution in the form

$$\mathbf{y} = \mathbf{x}te^{\mu t} + \mathbf{u}e^{\mu t}.$$

> **Linear independence** *here has* **two** *implications: in function space* AND *in ordinary vector space!*

Substitution:

$$\mathbf{x}e^{\mu t} + \mu\mathbf{x}te^{\mu t} + \mu\mathbf{u}e^{\mu t} = \mathbf{A}\mathbf{x}te^{\mu t} + \mathbf{A}\mathbf{u}e^{\mu t} \Rightarrow (\mathbf{A} - \mu\mathbf{I})\mathbf{u} = \mathbf{x}$$

Solve for $\mathbf{u}$, the *generalized eigenvector* of $\mathbf{A}$.
For Jordan blocks of larger sizes,

$$\mathbf{y}_1 = \mathbf{x}e^{\mu t},\ \mathbf{y}_2 = \mathbf{x}te^{\mu t} + \mathbf{u}_1e^{\mu t},\ \mathbf{y}_3 = \frac{1}{2}\mathbf{x}t^2e^{\mu t} + \mathbf{u}_1te^{\mu t} + \mathbf{u}_2e^{\mu t}\ \text{etc.}$$

> *Jordan canonical form (JCF) of* $\mathbf{A}$ *provides a set of* basis functions *to describe the complete solution of the ODE system.*



## Linear Non-Homogeneous Systems



$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}(t)$$

Complementary function:

$$\mathbf{y}_h(t) = \sum_{i=1}^{n} c_i\mathbf{y}_i(t) = [\mathcal{Y}(t)]\mathbf{c}$$

Complete solution:

$$\mathbf{y}(t) = \mathbf{y}_h(t) + \mathbf{y}_p(t)$$

We need to develop one particular solution $\mathbf{y}_p$.

**Method of undetermined coefficients**
Based on $\mathbf{g}(t)$, select candidate function $G_k(t)$ and propose

$$\mathbf{y}_p = \sum_k \mathbf{u}_k G_k(t),$$

*vector* coefficients ($\mathbf{u}_k$) to be determined by substitution.



## Linear Non-Homogeneous Systems



**Method of diagonalization**

*If* $\mathbf{A}$ is a diagonalizable constant matrix, with $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{D}$,

> *changing variables to* $\mathbf{z} = \mathbf{X}^{-1}\mathbf{y}$, *such that* $\mathbf{y} = \mathbf{X}\mathbf{z}$,

$$\mathbf{X}\mathbf{z}' = \mathbf{A}\mathbf{X}\mathbf{z} + \mathbf{g}(t) \Rightarrow \mathbf{z}' = \mathbf{X}^{-1}\mathbf{A}\mathbf{X}\mathbf{z} + \mathbf{X}^{-1}\mathbf{g}(t) = \mathbf{D}\mathbf{z} + \mathbf{h}(t)\ \text{(say)}.$$

Single decoupled Leibnitz equations

$$z_k' = d_kz_k + h_k(t), \quad k = 1, 2, 3, \cdots, n;$$

leading to individual solutions

$$z_k(t) = c_ke^{d_kt} + e^{d_kt}\int e^{-d_kt}h_k(t)dt.$$

After assembling $\mathbf{z}(t)$, we reconstruct $\mathbf{y} = \mathbf{X}\mathbf{z}$.



## Linear Non-Homogeneous Systems



**Method of variation of parameters**
If we can supply a basis $\mathcal{Y}(t)$ of the complementary function $\mathbf{y}_h(t)$, then we propose

$$\mathbf{y}_p(t) = [\mathcal{Y}(t)]\mathbf{u}(t)$$

Substitution leads to

$$\mathcal{Y}'\mathbf{u} + \mathcal{Y}\mathbf{u}' = \mathbf{A}\mathcal{Y}\mathbf{u} + \mathbf{g}.$$

Since $\mathcal{Y}' = \mathbf{A}\mathcal{Y}$,

$$\mathcal{Y}\mathbf{u}' = \mathbf{g}, \quad \text{or, } \mathbf{u}' = [\mathcal{Y}]^{-1}\mathbf{g}.$$

Complete solution:

$$\mathbf{y}(t) = \mathbf{y}_h + \mathbf{y}_p = [\mathcal{Y}]\mathbf{c} + [\mathcal{Y}]\int[\mathcal{Y}]^{-1}\mathbf{g}\,dt$$

*This method is completely general.*

## Points to note



- Theory of ODE's in terms of vector functions
- Methods to find
  - complementary functions in the case of constant coefficients
  - particular solutions for all cases

Necessary Exercises: **1**



## Outline







## Second Order Linear Systems



A system of two first order linear differential equations:

$$\begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2 \end{aligned}$$

or, $\mathbf{y}' = \mathbf{A}\mathbf{y}$

Phase: a pair of values of $y_1$ and $y_2$

Phase plane: plane of $y_1$ and $y_2$

Trajectory: a curve showing the evolution of the system for a particular initial value problem

Phase portrait: all trajectories together showing the complete picture of the behaviour of the dynamic system

Allowing only *isolated equilibrium points*,

- matrix **A** is non-singular: origin is the only equilibrium point.

Eigenvalues of **A**:

$$\lambda^2 - (a_{11} + a_{22})\lambda + (a_{11}a_{22} - a_{12}a_{21}) = 0$$



## Second Order Linear Systems



Characteristic equation:

$$\lambda^2 - p\lambda + q = 0,$$

with $p = (a_{11} + a_{22}) = \lambda_1 + \lambda_2$ and $q = a_{11}a_{22} - a_{12}a_{21} = \lambda_1\lambda_2$

Discriminant $D = p^2 - 4q$ and

$$\lambda_{1,2} = \frac{p}{2} \pm \sqrt{\left(\frac{p}{2}\right)^2 - q} = \frac{p}{2} \pm \frac{\sqrt{D}}{2}.$$

Solution (for diagonalizable **A**):

$$\mathbf{y} = c_1\mathbf{x}_1 e^{\lambda_1 t} + c_2\mathbf{x}_2 e^{\lambda_2 t}$$

Solution for deficient **A**:

$$\begin{aligned} \mathbf{y} &= c_1\mathbf{x}_1 e^{\lambda t} + c_2(t\mathbf{x}_1 + \mathbf{u})e^{\lambda t} \\ \Rightarrow \mathbf{y}' &= c_1\lambda\mathbf{x}_1 e^{\lambda t} + c_2(\mathbf{x}_1 + \lambda\mathbf{u})e^{\lambda t} + \lambda t c_2\mathbf{x}_1 e^{\lambda t} \end{aligned}$$

# Second Order Linear Systems

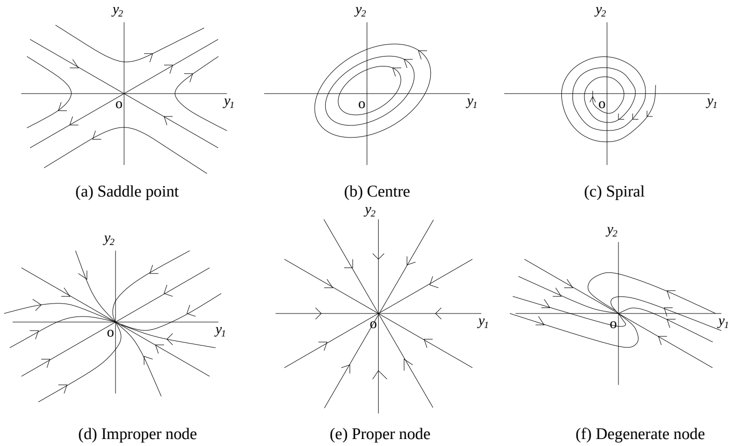



Figure: Neighbourhood of critical points



# Second Order Linear Systems

Table: Critical points of linear systems

| **Type** | **Sub-type** | **Eigenvalues** | **Position in $p$-$q$ chart** | **Stability** |
|---|---|---|---|---|
| Saddle pt | | real, opposite signs | $q < 0$ | unstable |
| Centre | | pure imaginary | $q > 0,\ p = 0$ | stable |
| Spiral | | complex, both non-zero components | $q > 0,\ p \neq 0$ $D = p^2 - 4q < 0$ | stable if $p < 0$, |
| Node | | real, same sign | $q > 0,\ p \neq 0,\ D \geq 0$ | unstable if $p > 0$ |
| | improper | unequal in magnitude | $D > 0$ | |
| | proper | equal, diagonalizable | $D = 0$ | |
| | degenerate | equal, deficient | $D = 0$ | |

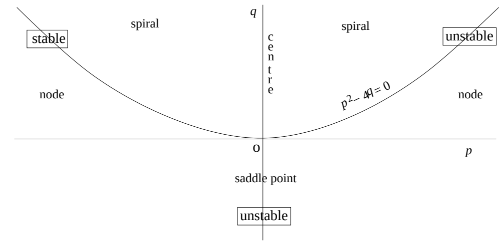



Figure: Zones of critical points in $p$-$q$ chart



# Nonlinear Dynamic Systems

**Phase plane analysis**

- Determine all the critical points.
- Linearize the ODE system around each of them as

$$\mathbf{y}' = \mathbf{J}(\mathbf{y}_0)(\mathbf{y} - \mathbf{y}_0).$$

- With $\mathbf{z} = \mathbf{y} - \mathbf{y}_0$, analyze each neighbourhood from $\mathbf{z}' = \mathbf{J}\mathbf{z}$.
- Assemble outcomes of local phase plane analyses.

*'Features' of a dynamic system are typically captured by its critical points and their neighbourhoods.*

**Limit cycles**

- isolated closed trajectories (only in nonlinear systems)

**Systems with arbitrary dimension of state space**?



# Lyapunov Stability Analysis

**Important terms**

Stability: If $\mathbf{y}_0$ is a critical point of the dynamic system $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ and for every $\epsilon > 0$, $\exists\, \delta > 0$ such that

$$\|\mathbf{y}(t_0) - \mathbf{y}_0\| < \delta \;\Rightarrow\; \|\mathbf{y}(t) - \mathbf{y}_0\| < \epsilon \;\; \forall\, t > t_0,$$

then $\mathbf{y}_0$ is a *stable* critical point. If, further, $\mathbf{y}(t) \to \mathbf{y}_0$ as $t \to \infty$, then $\mathbf{y}_0$ is said to be *asymptotically stable*.

Positive definite function: A function $V(\mathbf{y})$, with $V(\mathbf{0}) = 0$, is called positive definite if

$$V(\mathbf{y}) > 0 \;\forall \mathbf{y} \neq \mathbf{0}.$$

Lyapunov function: A positive definite function $V(\mathbf{y})$, having continuous $\frac{\partial V}{\partial y_i}$, with a negative semi-definite rate of change

$$V' = [\nabla V(\mathbf{y})]^T \mathbf{f}(\mathbf{y}).$$

## Lyapunov Stability Analysis

Lyapunov's stability criteria:

**Theorem:** *For a system* $\mathbf{y}' = \mathbf{f}(\mathbf{y})$ *with the origin as a critical point, if there exists a Lyapunov function* $V(\mathbf{y})$*, then the system is stable at the origin, i.e. the origin is a stable critical point.*
*Further, if* $V'(\mathbf{y})$ *is* negative definite, *then it is* asymptotically stable.

A generalization of the notion of total energy: negativity of its rate correspond to trajectories tending to decrease this 'energy'.

**Note:** Lyapunov's method becomes particularly important when a linearized model allows no analysis or when its results are suspect.

**Caution:** It is a one-way criterion only!



## Points to note

- Analysis of second order systems
- Classification of critical points
- Nonlinear systems and local linearization
- Phase plane analysis

  *Examples in physics, engineering, economics, biological and social systems*

- Lyapunov's method of stability analysis

Necessary Exercises: **1,2,3,4,5**



## Outline





## Power Series Method

Methods to solve an ODE in terms of elementary functions:

- restricted in scope

  **Theory** *allows study of the* properties *of solutions!*

When elementary methods fail,

- gain knowledge about solutions *through* properties, and
- for actual evaluation develop infinite series.

**Power series:**

$$y(x) = \sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + a_4 x^4 + a_5 x^5 + \cdots$$

or in powers of $(x - x_0)$.

**A simple exercise:**

*Try developing power series solutions in the above form and study their properties for differential equations*

$$y'' + y = 0 \quad \text{and} \quad 4x^2 y'' = y.$$

## Power Series Method



$$y'' + P(x)y' + Q(x)y = 0$$

If $P(x)$ and $Q(x)$ are analytic at a point $x = x_0$,

*i.e. if they possess convergent series expansions in powers of $(x - x_0)$ with some radius of convergence $R$,*

then the solution is analytic at $x_0$, and a power series solution

$$y(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^3 + \cdots$$

is convergent at least for $|x - x_0| < R$.

For $x_0 = 0$ (without loss of generality), suppose

$$P(x) = \sum_{n=0}^{\infty} p_n x^n = p_0 + p_1 x + p_2 x^2 + p_3 x^3 + \cdots,$$

$$Q(x) = \sum_{n=0}^{\infty} q_n x^n = q_0 + q_1 x + q_2 x^2 + q_3 x^3 + \cdots,$$

and assume $y(x) = \sum_{n=0}^{\infty} a_n x^n$.



## Power Series Method



Differentiation of $y(x) = \sum_{n=0}^{\infty} a_n x^n$ as

$$y'(x) = \sum_{n=0}^{\infty}(n+1)a_{n+1}x^n \quad \text{and} \quad y''(x) = \sum_{n=0}^{\infty}(n+2)(n+1)a_{n+2}x^n$$

leads to

$$P(x)y' = \sum_{n=0}^{\infty} p_n x^n \left[\sum_{n=0}^{\infty}(n+1)a_{n+1}x^n\right] = \sum_{n=0}^{\infty}\sum_{k=0}^{n} p_{n-k}(k+1)a_{k+1}x^n$$

$$Q(x)y = \sum_{n=0}^{\infty} q_n x^n \left[\sum_{n=0}^{\infty} a_n x^n\right] = \sum_{n=0}^{\infty}\sum_{k=0}^{n} q_{n-k}a_k x^n$$

$$\Rightarrow \sum_{n=0}^{\infty}\left[(n+2)(n+1)a_{n+2} + \sum_{k=0}^{n} p_{n-k}(k+1)a_{k+1} + \sum_{k=0}^{n} q_{n-k}a_k\right]x^n = 0$$

**Recursion formula:**

$$a_{n+2} = -\frac{1}{(n+2)(n+1)}\sum_{k=0}^{n}[(k+1)p_{n-k}a_{k+1} + q_{n-k}a_k]$$



## Frobenius' Method



For the ODE $y'' + P(x)y' + Q(x)y = 0$, a point $x = x_0$ is

ordinary point if $P(x)$ and $Q(x)$ are analytic at $x = x_0$: power series solution is analytic

singular point if any of the two is non-analytic (singular) at $x = x_0$

- regular singularity: $(x - x_0)P(x)$ and $(x - x_0)^2 Q(x)$ are analytic at the point
- irregular singularity

The case of **regular singularity**

For $x_0 = 0$, with $P(x) = \frac{b(x)}{x}$ and $Q(x) = \frac{c(x)}{x^2}$,

$$x^2 y'' + xb(x)y' + c(x)y = 0$$

in which $b(x)$ and $c(x)$ are analytic at the origin.



## Frobenius' Method



Working steps:

1. Assume the solution in the form $y(x) = x^r \sum_{n=0}^{\infty} a_n x^n$.
2. Differentiate to get the series expansions for $y'(x)$ and $y''(x)$.
3. Substitute these series for $y(x)$, $y'(x)$ and $y''(x)$ into the given ODE and collect coefficients of $x^r$, $x^{r+1}$, $x^{r+2}$ etc.
4. Equate the coefficient of $x^r$ to zero to obtain an equation in the index $r$, called the *indicial equation* as

$$r(r-1) + b_0 r + c_0 = 0;$$

   allowing $a_0$ to become arbitrary.

5. For each solution $r$, equate other coefficients to obtain $a_1$, $a_2$, $a_3$ etc in terms of $a_0$.

**Note:** The need is to develop *two* solutions.

# Special Functions Defined as Integrals

Gamma function: $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$, convergent for $n > 0$.

Recurrence relation $\Gamma(1) = 1$, $\Gamma(n+1) = n\Gamma(n)$ allows extension of the definition for the entire real line except for zero and negative integers.

$\Gamma(n+1) = n!$ for non-negative integers.

(A generalization of the factorial function.)

Beta function: $B(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1} dx = 2\int_0^{\pi/2} \sin^{2m-1}\theta \cos^{2n-1}\theta \; d\theta$; $m, n > 0$.

$B(m,n) = B(n,m)$; $B(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$

Error function: $\text{erf}\,(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$.

(Area under the normal or Gaussian distribution)

Sine integral function: $\text{Si}\,(x) = \int_0^x \frac{\sin t}{t} dt$.



# Special Functions Arising as Solutions of ODE's

In the study of some important problems in physics,

*some variable-coefficient ODE's appear recurrently,*

defying analytical solution!

Series solutions $\Rightarrow$ properties and connections

$\Rightarrow$ further problems $\Rightarrow$ further solutions $\Rightarrow \cdots$

Table: Special functions of mathematical physics

| Name of the ODE | Form of the ODE | Resulting functions |
|---|---|---|
| Legendre's equation | $(1-x^2)y'' - 2xy' + k(k+1)y = 0$ | Legendre functions<br>Legendre polynomials |
| Airy's equation | $y'' \pm k^2 xy = 0$ | Airy functions |
| Chebyshev's equation | $(1-x^2)y'' - xy' + k^2 y = 0$ | Chebyshev polynomials |
| Hermite's equation | $y'' - 2xy' + 2ky = 0$ | Hermite functions<br>Hermite polynomials |
| Bessel's equation | $x^2y'' + xy' + (x^2 - k^2)y = 0$ | Bessel functions<br>Neumann functions<br>Hankel functions |
| Gauss's hypergeometric equation | $x(1-x)y'' + [c - (a+b+1)x]y' - aby = 0$ | Hypergeometric function |
| Laguerre's equation | $xy'' + (1-x)y' + ky = 0$ | Laguerre polynomials |



# Special Functions Arising as Solutions of ODE's

**Legendre's equation**

$$(1-x^2)y'' - 2xy' + k(k+1)y = 0$$

$P(x) = -\frac{2x}{1-x^2}$ and $Q(x) = \frac{k(k+1)}{1-x^2}$ are analytic at $x = 0$ with radius of convergence $R = 1$.

*$x = 0$ is an ordinary point and a power series solution $y(x) = \sum_{n=0}^\infty a_n x^n$ is convergent at least for $|x| < 1$.*

Apply power series method:

$$\begin{aligned} a_2 &= -\frac{k(k+1)}{2!}a_0, \\ a_3 &= -\frac{(k+2)(k-1)}{3!}a_1 \\ \text{and} \quad a_{n+2} &= -\frac{(k-n)(k+n+1)}{(n+2)(n+1)}a_n \quad \text{for } n \geq 2. \end{aligned}$$

Solution: $y(x) = a_0 y_1(x) + a_1 y_2(x)$



# Special Functions Arising as Solutions of ODE's

**Legendre functions**

$$\begin{aligned} y_1(x) &= 1 - \frac{k(k+1)}{2!}x^2 + \frac{k(k-2)(k+1)(k+3)}{4!}x^4 - \cdots \\ y_2(x) &= x - \frac{(k-1)(k+2)}{3!}x^3 + \frac{(k-1)(k-3)(k+2)(k+4)}{5!}x^5 - \cdots \end{aligned}$$

Special significance: non-negative integral values of $k$

For each $k = 0, 1, 2, 3, \cdots$,

*one of the series terminates at the term containing $x^k$.*

Polynomial solution: valid for the entire real line!

Recurrence relation in reverse:

$$a_{k-2} = -\frac{k(k-1)}{2(2k-1)}a_k$$

## Special Functions Arising as Solutions of ODE's

**Legendre polynomial**

Choosing $a_k = \frac{(2k-1)(2k-3)\cdots 3\cdot 1}{k!}$,

$$P_k(x) = \frac{(2k-1)(2k-3)\cdots 3\cdot 1}{k!} \times \left[x^k - \frac{k(k-1)}{2(2k-1)}x^{k-2} + \frac{k(k-1)(k-2)(k-3)}{2\cdot 4(2k-1)(2k-3)}x^{k-4} - \cdots\right].$$

This choice of $a_k$ ensures $P_k(1) = 1$ and implies $P_k(-1) = (-1)^k$.

Initial Legendre polynomials:

$$P_0(x) = 1,$$
$$P_1(x) = x,$$
$$P_2(x) = \frac{1}{2}(3x^2 - 1),$$
$$P_3(x) = \frac{1}{2}(5x^3 - 3x),$$
$$P_4(x) = \frac{1}{8}(35x^4 - 30x^2 + 3) \text{ etc.}$$



## Special Functions Arising as Solutions of ODE's

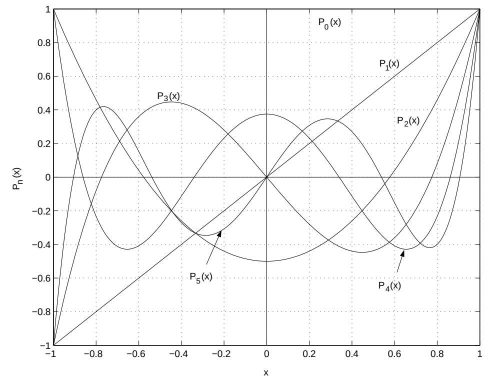



Figure: Legendre polynomials

All roots of a Legendre polynomial are real and they lie in $[-1, 1]$.

**Orthogonality**?



## Special Functions Arising as Solutions of ODE's

**Bessel's equation**

$$x^2y'' + xy' + (x^2 - k^2)y = 0$$

$x = 0$ is a regular singular point.

Frobenius' method: carrying out the early steps,

$$(r^2-k^2)a_0x^r+[(r+1)^2-k^2]a_1x^{r+1}+\sum_{n=2}^{\infty}[a_{n-2}+\{r^2-k^2+n(n+2r)\}a_n]x^{r+n} = 0$$

Indicial equation: $r^2 - k^2 = 0 \Rightarrow \quad r = \pm k$

With $r = k$, $\quad (r+1)^2 - k^2 \neq 0 \;\Rightarrow\; a_1 = 0$ and

$$a_n = -\frac{a_{n-2}}{n(n+2r)} \quad \text{for } n \geq 2.$$

Odd coefficients are zero and

$$a_2 = -\frac{a_0}{2(2k+2)},\; a_4 = \frac{a_0}{2\cdot 4(2k+2)(2k+4)},\text{ etc.}$$



## Special Functions Arising as Solutions of ODE's

**Bessel functions:**

Selecting $a_0 = \frac{1}{2^k\Gamma(k+1)}$ and using $n = 2m$,

$$a_m = \frac{(-1)^m}{2^{k+2m}m!\Gamma(k+m+1)}.$$

Bessel function of the first kind of order $k$:

$$J_k(x) = \sum_{m=0}^{\infty}(-1)^m\frac{x^{k+2m}}{2^{k+2m}m!\Gamma(k+m+1)} = \sum_{m=0}^{\infty}\frac{(-1)^m\left(\frac{x}{2}\right)^{k+2m}}{m!\Gamma(k+m+1)}$$

When $k$ is not an integer, $J_{-k}(x)$ completes the basis.

For integer $k$, $\quad J_{-k}(x) = (-1)^kJ_k(x)$, linearly dependent!

Reduction of order can be used to find another solution.

Bessel function of the second kind or Neumann function

## Points to note

- Solution in power series
- Ordinary points and singularities
- Definition of special functions
- Legendre polynomials
- Bessel functions

Necessary Exercises: **2,3,4,5**



## Outline





## Preliminary Ideas

A simple boundary value problem:

$$y'' + 2y = 0, \quad y(0) = 0, \quad y(\pi) = 0$$

General solution of the ODE:

$$y(x) = a\sin(x\sqrt{2}) + b\cos(x\sqrt{2})$$

Condition $y(0) = 0 \;\Rightarrow\; b = 0$. Hence, $y(x) = a\sin(x\sqrt{2})$.

Then, $y(\pi) = 0 \;\Rightarrow\; a = 0$. Only solution is $y(x) = 0$.

Now, consider the BVP

$$y'' + 4y = 0, \quad y(0) = 0, \quad y(\pi) = 0.$$

The same steps give $y(x) = a\sin(2x)$, with arbitrary value of $a$.

*Infinite number of non-trivial solutions!*



## Preliminary Ideas

**Boundary value problems as eigenvalue problems**

Explore the possible solutions of the BVP

$$y'' + ky = 0, \quad y(0) = 0, \quad y(\pi) = 0.$$

- With $k \le 0$, no hope for a non-trivial solution. Consider $k = \nu^2 > 0$.
- Solutions: $y = a\sin(\nu x)$, only for specific values of $\nu$ (or $k$): $\nu = 0, \pm 1, \pm 2, \pm 3, \cdots$; i.e. $k = 0, 1, 4, 9, \cdots$.

**Question:**

- For what values of $k$ (eigenvalues), does the given BVP possess non-trivial solutions, and
- what are the corresponding solutions (eigenfunctions), up to arbitrary scalar multiples?

Analogous to the *algebraic* eigenvalue problem $\mathbf{Av} = \lambda\mathbf{v}$!

Analogy of a Hermitian matrix: self-adjoint differential operator.

## Preliminary Ideas



Consider the ODE $y'' + P(x)y' + Q(x)y = 0$.

**Question:**

*Is it possible to find functions $F(x)$ and $G(x)$ such that*

$$F(x)y'' + F(x)P(x)y' + F(x)Q(x)y$$

*gets reduced to the derivative of $F(x)y' + G(x)y$?*

Comparing with

$$\frac{d}{dx}[F(x)y' + G(x)y] = F(x)y'' + [F'(x) + G(x)]y' + G'(x)y,$$

$$F'(x) + G(x) = F(x)P(x) \quad \text{and} \quad G'(x) = F(x)Q(x).$$

Elimination of $G(x)$:

$$F''(x) - P(x)F'(x) + [Q(x) - P'(x)]F(x) = 0$$

This is the **adjoint** of the original ODE.



## Preliminary Ideas



**The adjoint ODE**

- The adjoint of the ODE $y'' + P(x)y' + Q(x)y = 0$ is

  $$F'' + P_1F' + Q_1F = 0,$$

  where $P_1 = -P$ and $Q_1 = Q - P'$.
- Then, the adjoint of $F'' + P_1F' + Q_1F = 0$ is

  $$\phi'' + P_2\phi' + Q_2\phi = 0,$$

  where $P_2 = -P_1 = P$ and
  $Q_2 = Q_1 - P_1' = Q - P' - (-P') = Q$.

  *The adjoint of the adjoint of a second order linear homogeneous equation is the original equation itself.*
- When is an ODE its own adjoint?
  - $y'' + P(x)y' + Q(x)y = 0$ is self-adjoint only in the trivial case of $P(x) = 0$.
  - What about $F(x)y'' + F(x)P(x)y' + F(x)Q(x)y = 0$?



## Preliminary Ideas



**Second order self-adjoint ODE**

**Question:** What is the adjoint of $Fy'' + FPy' + FQy = 0$?

**Rephrased question:** What is the ODE that $\phi(x)$ has to satisfy if

$$\phi Fy'' + \phi FPy' + \phi FQy = \frac{d}{dx}\left[\phi Fy' + \xi(x)y\right]?$$

Comparing terms,

$$\frac{d}{dx}(\phi F) + \xi(x) = \phi FP \quad \text{and} \quad \xi'(x) = \phi FQ.$$

Eliminating $\xi(x)$, we have $\frac{d^2}{dx^2}(\phi F) + \phi FQ = \frac{d}{dx}(\phi FP)$.

$$F\phi'' + 2F'\phi' + F''\phi + FQ\phi = FP\phi' + (FP)'\phi$$

$$\Rightarrow F\phi'' + (2F' - FP)\phi' + \left[F'' - (FP)' + FQ\right]\phi = 0$$

This is the same as the original ODE, when $\boxed{F'(x) = F(x)P(x)}$



## Preliminary Ideas



Casting a given ODE into the self-adjoint form:

*Equation $y'' + P(x)y' + Q(x)y = 0$ is converted to the self-adjoint form through the multiplication of $F(x) = e^{\int P(x)dx}$.*

General form of self-adjoint equations:

$$\frac{d}{dx}[F(x)y'] + R(x)y = 0$$

Working rules:

- To determine whether a given ODE is in the self-adjoint form, check whether the coefficient of $y'$ is the derivative of the coefficient of $y''$.
- To convert an ODE into the self-adjoint form, first obtain the equation in normal form by dividing with the coefficient of $y''$. If the coefficient of $y'$ now is $P(x)$, then next multiply the resulting equation with $e^{\int P dx}$.

## Sturm-Liouville Problems



**Sturm-Liouville equation**

$$[r(x)y']' + [q(x) + \lambda p(x)]y = 0,$$

where $p$, $q$, $r$ and $r'$ are continuous on $[a, b]$, with $p(x) > 0$ on $[a, b]$ and $r(x) > 0$ on $(a, b)$.

With different boundary conditions,

Regular S-L problem:
$a_1 y(a) + a_2 y'(a) = 0$ and $b_1 y(b) + b_2 y'(b) = 0$,
vectors $[a_1 \;\; a_2]^T$ and $[b_1 \;\; b_2]^T$ being non-zero.

Periodic S-L problem: With $r(a) = r(b)$,
$y(a) = y(b)$ and $y'(a) = y'(b)$.

Singular S-L problem: If $r(a) = 0$, no boundary condition is needed at $x = a$. If $r(b) = 0$, no boundary condition is needed at $x = b$.
(We just look for bounded solutions over $[a, b]$.)



## Sturm-Liouville Problems



**Orthogonality of eigenfunctions**

**Theorem:** *If $y_m(x)$ and $y_n(x)$ are eigenfunctions (solutions) of a Sturm-Liouville problem corresponding to distinct eigenvalues $\lambda_m$ and $\lambda_n$ respectively, then*

$$(y_m, y_n) \equiv \int_a^b p(x) y_m(x) y_n(x) dx = 0,$$

*i.e. they are orthogonal with respect to the weight function $p(x)$.*

From the hypothesis,

$$(ry_m')' + (q + \lambda_m p)y_m = 0 \quad \Rightarrow \quad (q + \lambda_m p)y_m y_n = -(ry_m')' y_n$$
$$(ry_n')' + (q + \lambda_n p)y_n = 0 \quad \Rightarrow \quad (q + \lambda_n p)y_m y_n = -(ry_n')' y_m$$

Subtracting,

$$\begin{aligned}(\lambda_m - \lambda_n) p y_m y_n &= (ry_n')' y_m + (ry_n') y_m' - (ry_m') y_n' - (ry_m')' y_n \\ &= \left[r(y_m y_n' - y_n y_m')\right]'.\end{aligned}$$



## Sturm-Liouville Problems



Integrating both sides,

$$\begin{aligned}&(\lambda_m - \lambda_n) \int_a^b p(x) y_m(x) y_n(x) dx \\ &\quad = r(b)[y_m(b)y_n'(b) - y_n(b)y_m'(b)] - r(a)[y_m(a)y_n'(a) - y_n(a)y_m'(a)].\end{aligned}$$

- In a regular S-L problem, from the boundary condition at $x = a$, the homogeneous system $\begin{bmatrix} y_m(a) & y_m'(a) \\ y_n(a) & y_n'(a) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ has non-trivial solutions.
  Therefore, $y_m(a)y_n'(a) - y_n(a)y_m'(a) = 0$.
  Similarly, $y_m(b)y_n'(b) - y_n(b)y_m'(b) = 0$.
- In a singular S-L problem, zero value of $r(x)$ at a boundary makes the corresponding term vanish even without a BC.
- In a periodic S-L problem, the two terms cancel out together.

Since $\lambda_m \neq \lambda_n$, in all cases,

$$\int_a^b p(x) y_m(x) y_n(x) dx = 0.$$



## Sturm-Liouville Problems



**Example:** Legendre polynomials over $[-1, 1]$

Legendre's equation

$$\frac{d}{dx}[(1 - x^2)y'] + k(k+1)y = 0$$

is self-adjoint and defines a singular Sturm Liouville problem over $[-1, 1]$ with $p(x) = 1$, $q(x) = 0$, $r(x) = 1 - x^2$ and $\lambda = k(k+1)$.

$$(m-n)(m+n+1) \int_{-1}^{1} P_m(x) P_n(x) dx = [(1-x^2)(P_m P_n' - P_n P_m')]_{-1}^{1} = 0$$

From orthogonal decompositions $1 = P_0(x)$, $x = P_1(x)$,

$$\begin{aligned} x^2 &= \frac{1}{3}(3x^2 - 1) + \frac{1}{3} = \frac{2}{3}P_2(x) + \frac{1}{3}P_0(x), \\ x^3 &= \frac{1}{5}(5x^3 - 3x) + \frac{3}{5}x = \frac{2}{5}P_3(x) + \frac{3}{5}P_1(x), \\ x^4 &= \frac{8}{35}P_4(x) + \frac{4}{7}P_2(x) + \frac{1}{5}P_0(x) \;\; \text{etc;} \end{aligned}$$

$P_k(x)$ is orthogonal to all polynomials of degree less than $k$.

# Sturm-Liouville Problems



**Real eigenvalues**

*Eigenvalues of a Sturm-Liouville problem are real.*

Let eigenvalue $\lambda = \mu + i\nu$ and eigenfunction $y(x) = u(x) + iv(x)$.
Substitution leads to

$$[r(u' + iv')]' + [q + (\mu + i\nu)p](u + iv) = 0.$$

Separation of real and imaginary parts:

$$[ru']' + (q + \mu p)u - \nu pv = 0 \quad \Rightarrow \quad \nu pv^2 = [ru']'v + (q + \mu p)uv$$
$$[rv']' + (q + \mu p)v + \nu pu = 0 \quad \Rightarrow \quad \nu pu^2 = -[rv']'u - (q + \mu p)uv$$

Adding together,

$$\nu p(u^2 + v^2) = [ru']'v + [ru']v' - [rv']u' - [rv']'u = -\left[r(uv' - vu')\right]'$$

Integration and application of boundary conditions leads to

$$\nu \int_a^b p(x)[u^2(x) + v^2(x)]dx = 0.$$

$$\boxed{\nu = 0 \text{ and } \lambda = \mu}$$



# Eigenfunction Expansions



Eigenfunctions of Sturm-Liouville problems:

*convenient and powerful instruments to represent and manipulate fairly general classes of functions*

$\{y_0, y_1, y_2, y_3, \cdots\}$: a family of continuous functions over $[a, b]$, mutually orthogonal with respect to $p(x)$.

Representation of a function $f(x)$ on $[a, b]$:

$$f(x) = \sum_{m=0}^{\infty} a_m y_m(x) = a_0 y_0(x) + a_1 y_1(x) + a_2 y_2(x) + a_3 y_3(x) + \cdots$$

$$\boxed{\text{Generalized Fourier series}}$$

Analogous to the representation of a vector as a linear combination of a set of mutually orthogonal vectors.

**Question:** How to determine the coefficients $(a_n)$?



# Eigenfunction Expansions



Inner product:

$$\begin{aligned}(f, y_n) &= \int_a^b p(x)f(x)y_n(x)dx \\ &= \int_a^b \sum_{m=0}^{\infty}[a_m p(x)y_m(x)y_n(x)]dx = \sum_{m=0}^{\infty} a_m(y_m, y_n) = a_n\|y_n\|^2\end{aligned}$$

where

$$\|y_n\| = \sqrt{(y_n, y_n)} = \sqrt{\int_a^b p(x)y_n^2(x)dx}$$

Fourier coefficients: $a_n = \frac{(f, y_n)}{\|y_n\|^2}$
Normalized eigenfunctions:

$$\phi_m(x) = \frac{y_m(x)}{\|y_m(x)\|}$$

Generalized Fourier series (in orthonormal basis):

$$f(x) = \sum_{m=0}^{\infty} c_m\phi_m(x) = c_0\phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x) + c_3\phi_3(x) + \cdots$$



# Eigenfunction Expansions



In terms of a finite number of members of the family $\{\phi_k(x)\}$,

$$\Phi_N(x) = \sum_{m=0}^{N} \alpha_m\phi_m(x) = \alpha_0\phi_0(x) + \alpha_1\phi_1(x) + \alpha_2\phi_2(x) + \cdots + \alpha_N\phi_N(x).$$

Error

$$E = \|f - \Phi_N\|^2 = \int_a^b p(x)\left[f(x) - \sum_{m=0}^{N}\alpha_m\phi_m(x)\right]^2 dx$$

Error is minimized when

$$\frac{\partial E}{\partial \alpha_n} = \int_a^b 2p(x)\left[f(x) - \sum_{m=0}^{N}\alpha_m\phi_m(x)\right][-\phi_n(x)]dx = 0$$

$$\Rightarrow \int_a^b \alpha_n p(x)\phi_n^2(x)dx = \int_a^b p(x)f(x)\phi_n(x)dx.$$

$$\boxed{\alpha_n = c_n}$$

*best approximation in the mean* or *least square approximation*

## Eigenfunction Expansions



Using the Fourier coefficients, error

$$E = (f,f) - 2\sum_{n=0}^{N} c_n(f,\phi_n) + \sum_{n=0}^{N} c_n^2(\phi_n,\phi_n) = \|f\|^2 - 2\sum_{n=0}^{N} c_n^2 + \sum_{n=0}^{N} c_n^2$$

$$E = \|f\|^2 - \sum_{n=0}^{N} c_n^2 \geq 0.$$

**Bessel's inequality:**

$$\sum_{n=0}^{N} c_n^2 \leq \|f\|^2 = \int_a^b p(x)f^2(x)dx$$

Partial sum

$$s_k(x) = \sum_{m=0}^{k} a_m\phi_m(x)$$

**Question:** Does the sequence of $\{s_k\}$ converge?
**Answer:** The bound in Bessel's inequality ensures convergence.



## Eigenfunction Expansions



**Question:** Does it converge to $f$?

$$\lim_{k\to\infty} \int_a^b p(x)[s_k(x) - f(x)]^2 dx = 0?$$

**Answer:** Depends on the basis used.
**Convergence in the mean** or mean-square convergence:

> *An orthonormal set of functions* $\{\phi_k(x)\}$ *on an interval* $a \leq x \leq b$ *is said to be* complete *in* a class of functions, *or to form a* basis *for it, if the corresponding generalized Fourier series for a function converges in the mean to the function, for every function belonging to that class.*

**Parseval's identity:** $\sum_{n=0}^{\infty} c_n^2 = \|f\|^2$
**Eigenfunction expansion:** generalized Fourier series in terms of eigenfunctions of a Sturm-Liouville problem

- convergent for continuous functions with piecewise continuous derivatives, i.e. they form a basis for this class.



## Points to note



- Eigenvalue problems in ODE's
- Self-adjoint differential operators
- Sturm-Liouville problems
- Orthogonal eigenfunctions
- Eigenfunction expansions

Necessary Exercises: **1,2,4,5**



## Outline

## Basic Theory of Fourier Series



With $q(x) = 0$ and $p(x) = r(x) = 1$, periodic S-L problem:

$$y'' + \lambda y = 0, \quad y(-L) = y(L), \quad y'(-L) = y'(L)$$

Eigenfunctions $1, \cos\frac{\pi x}{L}, \sin\frac{\pi x}{L}, \cos\frac{2\pi x}{L}, \sin\frac{2\pi x}{L}, \cdots$ constitute an orthogonal basis for representing functions.
For a periodic function $f(x)$ of period $2L$, we propose

$$f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}\right)$$

and determine the Fourier coefficients from Euler formulae

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)dx,$$

$$a_m = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{m\pi x}{L}dx \ \text{ and } \ b_m = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{m\pi x}{L}dx.$$

**Question:** Does the series converge?



## Basic Theory of Fourier Series



**Dirichlet's conditions:**

*If $f(x)$ and its derivative are piecewise continuous on $[-L, L]$ and are periodic with a period $2L$, then the series converges to the mean $\frac{f(x+)+f(x-)}{2}$ of one-sided limits, at all points.*

Fourier series

*Note:* The interval of integration can be $[x_0, x_0 + 2L]$ for any $x_0$.

- It is valid to integrate the Fourier series term by term.
- The Fourier series *uniformly* converges to $f(x)$ over an interval on which $f(x)$ is continuous. At a jump discontinuity, convergence to $\frac{f(x+)+f(x-)}{2}$ is not uniform. Mismatch peak shifts with inclusion of more terms (Gibb's phenomenon).
- Term-by-term differentiation of the Fourier series at a point requires $f(x)$ to be smooth at that point.



## Basic Theory of Fourier Series



Multiplying the Fourier series with $f(x)$,

$$f^2(x) = a_0 f(x) + \sum_{n=1}^{\infty}\left[a_n f(x)\cos\frac{n\pi x}{L} + b_n f(x)\sin\frac{n\pi x}{L}\right]$$

**Parseval's identity:**

$$\Rightarrow a_0^2 + \frac{1}{2}\sum_{n=1}^{\infty}(a_n^2 + b_n^2) = \frac{1}{2L}\int_{-L}^{L} f^2(x)dx$$

The Fourier series representation is *complete*.

- A periodic function $f(x)$ is composed of its mean value and several sinusoidal components, or harmonics.
- Fourier coefficients are corresponding amplitudes.
- Parseval's identity is simply a statement on energy balance!

**Bessel's inequality**

$$a_0^2 + \frac{1}{2}\sum_{n=1}^{N}(a_n^2 + b_n^2) \le \frac{1}{2L}\|f(x)\|^2$$



## Extensions in Application



Original spirit of Fouries series

- representation of *periodic* functions over $(-\infty, \infty)$.

**Question:** What about a function $f(x)$ defined only on $[-L, L]$?
**Answer:** Extend the function as

$$F(x) = f(x) \ \text{ for } -L \le x \le L, \quad \text{and} \quad F(x + 2L) = F(x).$$

Fourier series of $F(x)$ acts as the Fourier series representation of $f(x)$ in its own domain.
In Euler formulae, notice that $b_m = 0$ for an even function.

*The Fourier series of an even function is a* **Fourier cosine series**

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi x}{L},$$

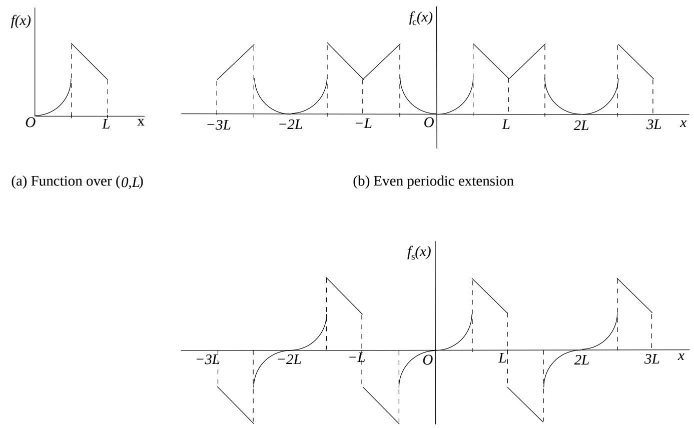
*where* $a_0 = \frac{1}{L}\int_0^L f(x)dx$ *and* $a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}dx.$

Similarly, for an odd function, **Fourier sine series**.

# Extensions in Application



Over $[0, L]$, sometimes we need a series of sine terms only, or cosine terms only!




(a) Function over (0,L)

(b) Even periodic extension

(c) Odd periodic extension

Figure: Periodic extensions for cosine and sine series



# Extensions in Application



**Half-range expansions**

- For Fourier cosine series of a function $f(x)$ over $[0, L]$, *even periodic extension:*

$$f_c(x) = \begin{cases} f(x) & \text{for} \quad 0 \le x \le L, \\ f(-x) & \text{for} \quad -L \le x < 0, \end{cases} \quad \text{and} \quad f_c(x+2L) = f_c(x)$$

- For Fourier sine series of a function $f(x)$ over $[0, L]$, *odd periodic extension:*

$$f_s(x) = \begin{cases} f(x) & \text{for} \quad 0 \le x \le L, \\ -f(-x) & \text{for} \quad -L \le x < 0, \end{cases} \quad \text{and} \quad f_s(x+2L) = f_s(x)$$

To develop the Fourier series of a function, which is available as a set of tabulated values or a black-box library routine,

*integrals in the Euler formulae are evaluated numerically.*

**Important:** Fourier series representation is richer and more powerful compared to interpolatory or least square approximation in many contexts.



# Fourier Integrals



**Question:** How to apply the idea of Fourier series to a non-periodic function over an infinite domain?
**Answer:** Magnify a single period to an infinite length.

Fourier series of function $f_L(x)$ of period $2L$:

$$f_L(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos p_n x + b_n \sin p_n x),$$

where $p_n = \frac{n\pi}{L}$ is the *frequency* of the $n$-th harmonic.

Inserting the expressions for the Fourier coefficients,

$$f_L(x) = \frac{1}{2L}\int_{-L}^{L} f_L(x)dx$$
$$+\frac{1}{\pi}\sum_{n=1}^{\infty}\left[\cos p_n x \int_{-L}^{L} f_L(v)\cos p_n v\, dv + \sin p_n x \int_{-L}^{L} f_L(v)\sin p_n v\, dv\right]\Delta p,$$

where $\Delta p = p_{n+1} - p_n = \frac{\pi}{L}$.



# Fourier Integrals



In the limit (if it exists), as $L \to \infty$, $\Delta p \to 0$,

$$f(x) = \frac{1}{\pi}\int_0^{\infty}\left[\cos px \int_{-\infty}^{\infty} f(v)\cos pv\, dv + \sin px \int_{-\infty}^{\infty} f(v)\sin pv\, dv\right] dp$$

**Fourier integral** of $f(x)$:

$$f(x) = \int_0^{\infty} [A(p)\cos px + B(p)\sin px]dp,$$

where **amplitude functions**

$$A(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos pv\, dv \quad \text{and} \quad B(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\sin pv\, dv$$

are defined for a *continuous* frequency variable $p$.

In phase angle form,

$$f(x) = \frac{1}{\pi}\int_0^{\infty}\int_{-\infty}^{\infty} f(v)\cos p(x-v)dv\, dp.$$

## Fourier Integrals



Using $\cos\theta = \frac{e^{i\theta}+e^{-i\theta}}{2}$ in the phase angle form,

$$f(x) = \frac{1}{2\pi}\int_0^\infty \int_{-\infty}^\infty f(v)[e^{ip(x-v)} + e^{-ip(x-v)}]dv\,dp.$$

With substitution $p = -q$,

$$\int_0^\infty \int_{-\infty}^\infty f(v)e^{-ip(x-v)}dv\,dp = \int_{-\infty}^0 \int_{-\infty}^\infty f(v)e^{iq(x-v)}dv\,dq.$$

**Complex form of Fourier integral**

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^\infty \int_{-\infty}^\infty f(v)e^{ip(x-v)}dv\,dp = \int_{-\infty}^\infty C(p)e^{ipx}dp,$$

in which the complex Fourier integral coefficient is

$$C(p) = \frac{1}{2\pi}\int_{-\infty}^\infty f(v)e^{-ipv}dv.$$



## Points to note



- Fourier series arising out of a Sturm-Liouville problem
- A versatile tool for function representation
- Fourier integral as the limiting case of Fourier series

Necessary Exercises: **1,3,6,8**



## Outline







## Definition and Fundamental Properties



**Complex form of the Fourier integral:**

$$f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty f(v)e^{-iwv}dv\right] e^{iwt}dw$$

*Composition of an infinite number of functions in the form $\frac{e^{iwt}}{\sqrt{2\pi}}$, over a continuous distribution of frequency $w$.*

**Fourier transform:** Amplitude of a frequency component:

$$\mathcal{F}(f) \equiv \hat{f}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty f(t)e^{-iwt}dt$$

Function of the frequency variable.

**Inverse Fourier transform**

$$\mathcal{F}^{-1}(\hat{f}) \equiv f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \hat{f}(w)e^{iwt}dw$$

recovers the original function.

## Definition and Fundamental Properties

**Example:** Fourier transform of $f(t) = 1$?
Let us find out the inverse Fourier transform of $\hat{f}(w) = k\delta(w)$.

$$f(t) = \mathcal{F}^{-1}(\hat{f}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} k\delta(w)e^{iwt}dw = \frac{k}{\sqrt{2\pi}}$$

$$\boxed{\mathcal{F}(1) = \sqrt{2\pi}\delta(w)}$$

Linearity of Fourier transforms:

$$\mathcal{F}\{\alpha f_1(t) + \beta f_2(t)\} = \alpha\hat{f}_1(w) + \beta\hat{f}_2(w)$$

Scaling:

$$\mathcal{F}\{f(at)\} = \frac{1}{|a|}\hat{f}\left(\frac{w}{a}\right) \quad \text{and} \quad \mathcal{F}^{-1}\left\{\hat{f}\left(\frac{w}{a}\right)\right\} = |a|f(at)$$

Shifting rules:

$$\begin{aligned}\mathcal{F}\{f(t-t_0)\} &= e^{-iwt_0}\mathcal{F}\{f(t)\}\\ \mathcal{F}^{-1}\{\hat{f}(w-w_0)\} &= e^{iw_0t}\mathcal{F}^{-1}\{\hat{f}(w)\}\end{aligned}$$



## Important Results on Fourier Transforms

**Fourier transform of the derivative of a function:**

If $f(t)$ is continuous in every interval and $f'(t)$ is piecewise continuous, $\int_{-\infty}^{\infty}|f(t)|dt$ converges and $f(t)$ approaches zero as $t \to \pm\infty$, *then*

$$\begin{aligned}\mathcal{F}\{f'(t)\} &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f'(t)e^{-iwt}dt\\ &= \frac{1}{\sqrt{2\pi}}\left[f(t)e^{-iwt}\right]_{-\infty}^{\infty} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(-iw)f(t)e^{-iwt}dt\\ &= iw\hat{f}(w).\end{aligned}$$

Alternatively, differentiating the inverse Fourier transform,

$$\begin{aligned}\frac{d}{dt}[f(t)] &= \frac{d}{dt}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\hat{f}(w)e^{iwt}dw\right]\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\frac{\partial}{\partial t}\left[\hat{f}(w)e^{iwt}\right]dw = \mathcal{F}^{-1}\{iw\hat{f}(w)\}.\end{aligned}$$



## Important Results on Fourier Transforms

Under *appropriate* premises,

$$\mathcal{F}\{f''(t)\} = (iw)^2\hat{f}(w) = -w^2\hat{f}(w).$$

In general, $\mathcal{F}\{f^{(n)}(t)\} = (iw)^n\hat{f}(w)$.

**Fourier transform of an integral:**

*If $f(t)$ is piecewise continuous on every interval, $\int_{-\infty}^{\infty}|f(t)|dt$ converges and $\hat{f}(0) = 0$, then*

$$\mathcal{F}\left\{\int_{-\infty}^{t} f(\tau)d\tau\right\} = \frac{1}{iw}\hat{f}(w).$$

**Derivative of a Fourier transform** (with respect to the frequency variable):

$$\mathcal{F}\{t^n f(t)\} = i^n\frac{d^n}{dw^n}\hat{f}(w),$$

if $f(t)$ is piecewise continuous and $\int_{-\infty}^{\infty}|t^n f(t)|dt$ converges.



## Important Results on Fourier Transforms

**Convolution** of two functions:

$$h(t) = f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau$$

$$\begin{aligned}\hat{h}(w) &= \mathcal{F}\{h(t)\}\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\tau)g(t-\tau)e^{-iwt}d\tau\, dt\\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(\tau)e^{-iw\tau}\left[\int_{-\infty}^{\infty} g(t-\tau)e^{-iw(t-\tau)}dt\right]d\tau\\ &= \int_{-\infty}^{\infty} f(\tau)e^{-iw\tau}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} g(t')e^{-iwt'}dt'\right]d\tau\end{aligned}$$

Convolution theorem for Fourier transforms:

$$\boxed{\hat{h}(w) = \sqrt{2\pi}\hat{f}(w)\hat{g}(w)}$$

## Important Results on Fourier Transforms



**Conjugate of the Fourier transform:**

$$\hat{f}^*(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(t) e^{iwt} dt$$

Inner product of $\hat{f}(w)$ and $\hat{g}(w)$:

$$\begin{aligned} \int_{-\infty}^{\infty} \hat{f}^*(w)\hat{g}(w)dw &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f^*(t)e^{iwt}dt \; \hat{g}(w)dw \\ &= \int_{-\infty}^{\infty} f^*(t) \left[ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{g}(w)e^{iwt}dw \right] dt \\ &= \int_{-\infty}^{\infty} f^*(t)g(t)dt. \end{aligned}$$

**Parseval's identity:** For $g(t) = f(t)$ in the above,

$$\int_{-\infty}^{\infty} \|\hat{f}(w)\|^2 dw = \int_{-\infty}^{\infty} \|f(t)\|^2 dt,$$

equating the total energy content of the frequency spectrum of a wave or a signal to the total energy flow over time.



## Discrete Fourier Transform



Consider a signal $f(t)$ from actual measurement or *sampling*.
We want to analyze its amplitude spectrum (versus frequency).

For the FT, how to evaluate the integral over $(-\infty, \infty)$?

**Windowing:** Sample the signal $f(t)$ over a finite interval.

A window function:

$$g(t) = \begin{cases} 1 & \text{for } a \le t \le b \\ 0 & \text{otherwise} \end{cases}$$

Actual processing takes place on the windowed function $f(t)g(t)$.

**Next question:** Do we need to evaluate the amplitude for all $w \in (-\infty, \infty)$?

> *Most useful signals are particularly rich only in their own characteristic frequency bands.*

Decide on an *expected* frequency band, say $[-w_c, w_c]$.



## Discrete Fourier Transform



**Time step for sampling?**

With $N$ sampling over $[a, b)$,

$$w_c \Delta \le \pi,$$

data being collected at $t = a, a+\Delta, a+2\Delta, \cdots, a+(N-1)\Delta$, with $N\Delta = b - a$.

Nyquist critical frequency

Note the duality.

- Decision of sampling rate $\Delta$ determines the *band* of frequency content that can be accommodated.
- Decision of the interval $[a, b)$ dictates how *finely* the frequency spectrum can be developed.

**Shannon's sampling theorem**

> *A band-limited signal can be reconstructed from a finite number of samples.*



## Discrete Fourier Transform



With discrete data at $t_k = k\Delta$ for $k = 0, 1, 2, 3, \cdots, N-1$,

$$\hat{\mathbf{f}}(\mathbf{w}) = \frac{\Delta}{\sqrt{2\pi}} \left[ m_j^k \right] \mathbf{f}(\mathbf{t}),$$

where $m_j = e^{-iw_j\Delta}$ and $\left[ m_j^k \right]$ is an $N \times N$ matrix.

A similar discrete version of inverse Fourier transform.

Reconstruction: a trigonometric interpolation of sampled data.

- Structure of Fourier and inverse Fourier transforms reduces the problem with a system of linear equations [$\mathcal{O}(N^3)$ operations] to that of a matrix-vector multiplication [$\mathcal{O}(N^2)$ operations].
- Structure of matrix $\left[ m_j^k \right]$, with patterns of redundancies, opens up a trick to reduce it further to $\mathcal{O}(N \log N)$ operations.

Cooley-Tuckey algorithm:

fast Fourier transform (FFT)

## Discrete Fourier Transform



DFT representation reliable only if the incoming signal is really band-limited in the interval $[-w_c, w_c]$.
Frequencies beyond $[-w_c, w_c]$ distort the spectrum near $w = \pm w_c$ by folding back.

Aliasing

Detection: *a posteriori*
**Bandpass filtering:** If we *expect* a signal having components only in certain frequency bands and want to get rid of unwanted *noise* frequencies,

> *for every band $[w_1, w_2]$ of our interest, we define window function $\hat{\phi}(w)$ with intervals $[-w_2, -w_1]$ and $[w_1, w_2]$.*

Windowed Fourier transform $\hat{\phi}(w)\hat{f}(w)$ filters out frequency components outside this band.
For recovery,

> *convolve raw signal $f(t)$ with IFT $\phi(t)$ of $\hat{\phi}(w)$.*



## Points to note



- Fourier transform as amplitude function in Fourier integral
- Basic operational tools in Fourier and inverse Fourier transforms
- Conceptual notions of discrete Fourier transform (DFT)

Necessary Exercises: **1,3,6**



## Outline







## Approximation with Chebyshev polynomials



**Chebyshev polynomials:**
Polynomial solutions of the singular Sturm-Liouville problem

$$(1-x^2)y'' - xy' + n^2 y = 0 \quad \text{or} \quad \left[\sqrt{1-x^2}\, y'\right]' + \frac{n^2}{\sqrt{1-x^2}} y = 0$$

over $-1 \le x \le 1$, with $T_n(1) = 1$ for all $n$.

Closed-form expressions:

$$T_n(x) = \cos(n \cos^{-1} x),$$

or,

$$T_0(x) = 1,\ T_1(x) = x,\ T_2(x) = 2x^2 - 1,\ T_3(x) = 4x^3 - 3x,\ \cdots;$$

with the three-term recurrence relation

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x).$$

## Approximation with Chebyshev polynomials

Immediate observations

- Coefficients in a Chebyshev polynomial are integers. In particular, the leading coefficient of $T_n(x)$ is $2^{n-1}$.
- For even $n$, $T_n(x)$ is an even function, while for odd $n$ it is an odd function.
- $T_n(1) = 1$, $T_n(-1) = (-1)^n$ and $|T_n(x)| \leq 1$ for $-1 \leq x \leq 1$.
- Zeros of a Chebyshev polynomial $T_n(x)$ are real and lie inside the interval $[-1, 1]$ at locations $x = \cos\frac{(2k-1)\pi}{2n}$ for $k = 1, 2, 3, \cdots, n$.
  These locations are also called *Chebyshev accuracy points*.
  Further, zeros of $T_n(x)$ are interlaced by those of $T_{n+1}(x)$.
- Extrema of $T_n(x)$ are of magnitude equal to unity, alternate in sign and occur at $x = \cos\frac{k\pi}{n}$ for $k = 0, 1, 2, 3, \cdots, n$.
- Orthogonality and norms:

$$\int_{-1}^{1} \frac{T_m(x)T_n(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0 & \text{if} \quad m \neq n, \\ \frac{\pi}{2} & \text{if} \quad m = n \neq 0, \quad \text{and} \\ \pi & \text{if} \quad m = n = 0. \end{cases}$$



## Approximation with Chebyshev polynomials

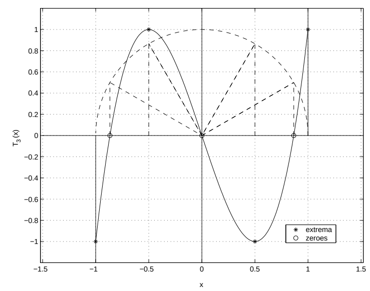



Figure: Extrema and zeros of $T_3(x)$

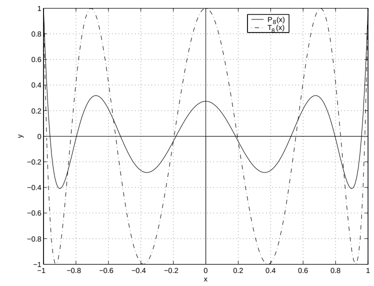


Figure: Contrast: $P_8(x)$ and $T_8(x)$

Being cosines and polynomials at the same time, Chebyshev polynomials possess a wide variety of interesting properties!

Most striking property:

*equal-ripple oscillations, leading to minimax property*



## Approximation with Chebyshev polynomials

**Minimax property**

**Theorem:** *Among all polynomials $p_n(x)$ of degree $n > 0$ with the leading coefficient equal to unity, $2^{1-n}T_n(x)$ deviates least from zero in $[-1, 1]$. That is,*

$$\max_{-1\leq x\leq 1} |p_n(x)| \geq \max_{-1\leq x\leq 1} |2^{1-n}T_n(x)| = 2^{1-n}.$$

If there exists a monic polynomial $p_n(x)$ of degree $n$ such that

$$\max_{-1\leq x\leq 1} |p_n(x)| < 2^{1-n},$$

then at $(n+1)$ locations of alternating extrema of $2^{1-n}T_n(x)$, the polynomial

$$q_n(x) = 2^{1-n}T_n(x) - p_n(x)$$

will have the same sign as $2^{1-n}T_n(x)$.
With alternating signs at $(n+1)$ locations in sequence, $q_n(x)$ will have $n$ intervening zeros, even though it is a polynomial of degree at most $(n-1)$: CONTRADICTION!



## Approximation with Chebyshev polynomials

**Chebyshev series**

$$f(x) = a_0T_0(x) + a_1T_1(x) + a_2T_2(x) + a_3T_3(x) + \cdots$$

with coefficients

$$a_0 = \frac{1}{\pi}\int_{-1}^{1} \frac{f(x)T_0(x)}{\sqrt{1-x^2}}dx \text{ and } a_n = \frac{2}{\pi}\int_{-1}^{1} \frac{f(x)T_n(x)}{\sqrt{1-x^2}}dx \text{ for } n = 1, 2, 3, \cdots$$

A truncated series $\sum_{k=0}^{n} a_kT_k(x)$:

*Chebyshev economization*

Leading error term $a_{n+1}T_{n+1}(x)$ deviates least from zero over $[-1, 1]$ and is *qualitatively similar* to the error function.

**Question:** How to develop a Chebyshev series approximation?
Find out so many Chebyshev polynomials and evaluate coefficients?

# Approximation with Chebyshev polynomials

For approximating $f(t)$ over $[a, b]$, scale the variable as
$t = \frac{a+b}{2} + \frac{b-a}{2}x$, with $x \in [-1, 1]$.

**Remark:** The economized series $\sum_{k=0}^{n} a_k T_k(x)$ gives minimax deviation of the leading error term $a_{n+1} T_{n+1}(x)$.

Assuming $a_{n+1} T_{n+1}(x)$ to *be* the error, at the zeros of $T_{n+1}(x)$, the error will be 'officially' zero, i.e.

$$\sum_{k=0}^{n} a_k T_k(x_j) = f(t(x_j)),$$

where $x_0, x_1, x_2, \cdots, x_n$ are the roots of $T_{n+1}(x)$.

**Recall:** *Values of an n-th degree polynomial at $n+1$ points uniquely fix the entire polynomial.*

Interpolation of these $n+1$ values leads to the same polynomial!

Chebyshev-Lagrange approximation



# Minimax Polynomial Approximation

Situations in which minimax approximation is desirable:

- Develop the approximation once and keep it for use in future.

Requirement: Uniform quality control over the entire domain

**Minimax approximation:**

*deviation limited by the* constant *amplitude of ripple*

**Chebyshev's minimax theorem**

**Theorem:** *Of all polynomials of degree up to n, $p(x)$ is the minimax polynomial approximation of $f(x)$, i.e. it minimizes*

$$\max |f(x) - p(x)|,$$

*if and only if there are $n+2$ points $x_i$ such that*

$$a \leq x_1 < x_2 < x_3 < \cdots < x_{n+2} \leq b,$$

*where the difference $f(x) - p(x)$ takes its extreme values of the same magnitude and alternating signs.*



# Minimax Polynomial Approximation

Utilize any gap to reduce the deviation at the other extrema with values at the bound.

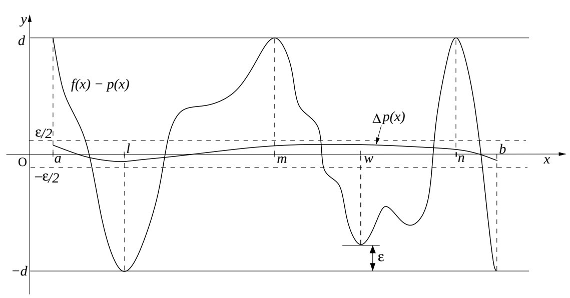



Figure: Schematic of an approximation that is not minimax

**Construction of the minimax polynomial:** Remez algorithm

*Note:* In the light of this theorem and algorithm, examine how $T_{n+1}(x)$ is *qualitatively similar* to the complete error function!



# Points to note

- Unique features of Chebyshev polynomials
- The equal-ripple and minimax properties
- Chebyshev series and Chebyshev-Lagrange approximation
- Fundamental ideas of general minimax approximation

Necessary Exercises: **2,3,4**

# Outline







# Introduction



**Quasi-linear second order PDE's**

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = F(x, y, u, u_x, u_y)$$

hyperbolic if $b^2 - ac > 0$, modelling phenomena which evolve in time perpetually and do not approach a steady state

parabolic if $b^2 - ac = 0$, modelling phenomena which evolve in time in a transient manner, approaching steady state

elliptic if $b^2 - ac < 0$, modelling steady-state configurations, without evolution in time

If $F(x, y, u, u_x, u_y) = 0$,

*second order linear homogeneous differential equation*

Principle of superposition: A linear combination of different solutions is also a solution.

Solutions are often in the form of infinite series.

- Solution techniques in PDE's typically attack the boundary value problem directly.



# Introduction



**Initial and boundary conditions**

Time and space variables are *qualitatively* different.

- Conditions in time: typically initial conditions.
  For second order PDE's, $u$ and $u_t$ over the *entire* space domain: Cauchy conditions
  - Time is a single variable and is *decoupled* from the space variables.
- Conditions in space: typically boundary conditions.
  For $u(t, x, y)$, boundary conditions over the entire curve in the $x$-$y$ plane that encloses the domain. For second order PDE's,
  - Dirichlet condition: value of the function
  - Neumann condition: derivative normal to the boundary
  - Mixed (Robin) condition

Dirichlet, Neumann and Cauchy problems



# Introduction



**Method of separation of variables**

For $u(x, y)$, propose a solution in the form

$$u(x, y) = X(x)Y(y)$$

and substitute

$$u_x = X'Y,\ u_y = XY',\ u_{xx} = X''Y,\ u_{xy} = X'Y',\ u_{yy} = XY''$$

to cast the equation into the form

$$\phi(x, X, X', X'') = \psi(y, Y, Y', Y'').$$

If the manoeuvre succeeds then, $x$ and $y$ being independent variables, it implies

$$\phi(x, X, X', X'') = \psi(y, Y, Y', Y'') = k.$$

Nature of the *separation constant* $k$ is decided based on the context, resulting ODE's are solved in consistency with the boundary conditions and assembled to construct $u(x, y)$.

# Hyperbolic Equations



**Transverse vibrations of a string**

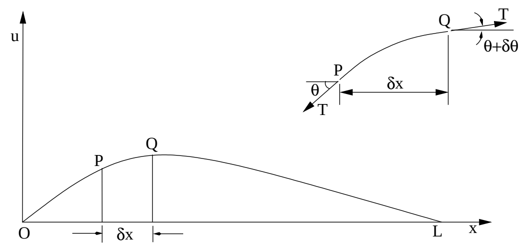



Figure: Transverse vibration of a stretched string

Small deflection and slope: $\cos\theta \approx 1,\ \sin\theta \approx \theta \approx \tan\theta$

Horizontal (longitudinal) forces on $PQ$ balance.
From Newton's second law, vertical (transverse) deflection $u(x,t)$:

$$T\sin(\theta+\delta\theta) - T\sin\theta = \rho\delta x \frac{\partial^2 u}{\partial t^2}$$



# Hyperbolic Equations



Under the assumptions, denoting $c^2 = \frac{T}{\rho}$,

$$\delta x \frac{\partial^2 u}{\partial t^2} = c^2 \left[ \left.\frac{\partial u}{\partial x}\right|_Q - \left.\frac{\partial u}{\partial x}\right|_P \right].$$

In the limit, as $\delta x \to 0$, PDE of transverse vibration:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$

one-dimensional wave equation

Boundary conditions (in this case): $u(0,t) = u(L,t) = 0$

Initial configuration and initial velocity:

$u(x,0) = f(x)$ and $u_t(x,0) = g(x)$

**Cauchy problem**: Determine $u(x,t)$ for $0 \le x \le L,\ t \ge 0$.



# Hyperbolic Equations



**Solution by separation of variables**

$u_{tt} = c^2 u_{xx},\ u(0,t) = u(L,t) = 0,\ u(x,0) = f(x),\ u_t(x,0) = g(x)$

Assuming

$$u(x,t) = X(x)T(t),$$

and substituting $u_{tt} = XT''$ and $u_{xx} = X''T$, variables are separated as

$$\frac{T''}{c^2 T} = \frac{X''}{X} = -p^2.$$

The PDE splits into two ODE's

$$X'' + p^2 X = 0 \quad \text{and} \quad T'' + c^2 p^2 T = 0.$$

Eigenvalues of BVP $X'' + p^2 X = 0,\ X(0) = X(L) = 0$ are $p = \frac{n\pi}{L}$ and eigenfunctions

$$X_n(x) = \sin px = \sin \frac{n\pi x}{L} \quad \text{for} \quad n = 1, 2, 3, \cdots.$$

Second ODE: $T'' + \lambda_n^2 T = 0$, with $\lambda_n = \frac{cn\pi}{L}$



# Hyperbolic Equations



*Corresponding* solution:

$$T_n(t) = A_n \cos\lambda_n t + B_n \sin\lambda_n t$$

Then, for $n = 1, 2, 3, \cdots$,

$$u_n(x,t) = X_n(x)T_n(t) = (A_n \cos\lambda_n t + B_n \sin\lambda_n t)\sin\frac{n\pi x}{L}$$

satisfies the PDE and the boundary conditions.

Since the PDE and the BC's are homogeneous, by superposition,

$$u(x,t) = \sum_{n=1}^{\infty} [A_n \cos\lambda_n t + B_n \sin\lambda_n t] \sin\frac{n\pi x}{L}.$$

**Question:** How to determine coefficients $A_n$ and $B_n$?

**Answer:** By imposing the initial conditions.

# Hyperbolic Equations



Initial conditions: Fourier sine series of $f(x)$ and $g(x)$

$$u(x,0) = f(x) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L}$$

$$u_t(x,0) = g(x) = \sum_{n=1}^{\infty} \lambda_n B_n \sin\frac{n\pi x}{L}$$

Hence, coefficients:

$$A_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}dx \quad \text{and} \quad B_n = \frac{2}{cn\pi}\int_0^L g(x)\sin\frac{n\pi x}{L}dx$$

**Related problems:**

- Different boundary conditions: other kinds of series
- Long wire: infinite domain, continuous frequencies and solution from Fourier integrals
  Alternative: Reduce the problem using Fourier transforms.
- General wave equation in 3-d: $u_{tt} = c^2\nabla^2 u$
- Membrane equation: $u_{tt} = c^2(u_{xx} + u_{yy})$



# Hyperbolic Equations



**D'Alembert's solution of the wave equation**

**Method of characteristics**

**Canonical form**

By coordinate transformation from $(x, y)$ to $(\xi, \eta)$, with $U(\xi, \eta) = u[x(\xi, \eta), y(\xi, \eta)]$,

hyperbolic equation: $U_{\xi\eta} = \Phi$

parabolic equation: $U_{\xi\xi} = \Phi$

elliptic equation: $U_{\xi\xi} + U_{\eta\eta} = \Phi$

in which $\Phi(\xi, \eta, U, U_\xi, U_\eta)$ is free from second derivatives.

For a hyperbolic equation, entire domain becomes a network of $\xi$-$\eta$ coordinate curves, known as *characteristic curves*,

*along which decoupled solutions can be tracked!*



# Hyperbolic Equations



For a hyperbolic equation in the form

$$a\frac{\partial^2 u}{\partial x^2} + 2b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} = F(x, y, u, u_x, u_y),$$

roots of $am^2 + 2bm + c$ are

$$m_{1,2} = \frac{-b \pm \sqrt{b^2 - ac}}{a},$$

real and distinct.

Coordinate transformation

$$\xi = y + m_1 x, \quad \eta = y + m_2 x$$

leads to $U_{\xi\eta} = \Phi(\xi, \eta, U, U_\xi, U_\eta)$.

For the BVP

$$u_{tt} = c^2 u_{xx}, \; u(0,t) = u(L,t) = 0, \; u(x,0) = f(x), \; u_t(x,0) = g(x),$$

canonical coordinate transformation:

$$\xi = x - ct, \; \eta = x + ct, \quad \text{with} \quad x = \frac{1}{2}(\xi + \eta), \; t = \frac{1}{2c}(\eta - \xi).$$



# Hyperbolic Equations



Substitution of derivatives

$$u_x = U_\xi \xi_x + U_\eta \eta_x = U_\xi + U_\eta \quad \Rightarrow \quad u_{xx} = U_{\xi\xi} + 2U_{\xi\eta} + U_{\eta\eta}$$

$$u_t = U_\xi \xi_t + U_\eta \eta_t = -cU_\xi + cU_\eta \quad \Rightarrow \quad u_{tt} = c^2 U_{\xi\xi} - 2c^2 U_{\xi\eta} + c^2 U_{\eta\eta}$$

into the PDE $u_{tt} = c^2 u_{xx}$ gives

$$c^2(U_{\xi\xi} - 2U_{\xi\eta} + U_{\eta\eta}) = c^2(U_{\xi\xi} + 2U_{\xi\eta} + U_{\eta\eta}).$$

Canonical form: $U_{\xi\eta} = 0$

Integration:

$$U_\xi = \int U_{\xi\eta} d\eta + \psi(\xi) = \psi(\xi)$$

$$\Rightarrow U(\xi, \eta) = \int \psi(\xi) d\xi + f_2(\eta) = f_1(\xi) + f_2(\eta)$$

**D'Alembert's solution:** $u(x,t) = f_1(x - ct) + f_2(x + ct)$

# Hyperbolic Equations



**Physical insight from D'Alembert's solution:**

$f_1(x-ct)$: a *progressive wave* in forward direction with speed $c$

Reflection at boundary:

*in a manner depending upon the boundary condition*

Reflected wave $f_2(x+ct)$: another *progressive wave*, this one in backward direction with speed $c$

Superposition of two waves: complete solution (response)

**Note:** Components of the earlier solution: with $\lambda_n = \frac{cn\pi}{L}$,

$$\begin{aligned}\cos\lambda_n t \sin\frac{n\pi x}{L} &= \frac{1}{2}\left[\sin\frac{n\pi}{L}(x-ct) + \sin\frac{n\pi}{L}(x+ct)\right]\\ \sin\lambda_n t \sin\frac{n\pi x}{L} &= \frac{1}{2}\left[\cos\frac{n\pi}{L}(x-ct) - \cos\frac{n\pi}{L}(x+ct)\right]\end{aligned}$$



# Parabolic Equations



**Heat conduction equation** or **diffusion equation**:

$$\frac{\partial u}{\partial t} = c^2\nabla^2 u$$

One-dimensional heat (diffusion) equation:

$$u_t = c^2 u_{xx}$$

**Heat conduction in a finite bar:** For a thin bar of length $L$ with end-points at zero temperature,

$$u_t = c^2 u_{xx},\quad u(0,t) = u(L,t) = 0,\quad u(x,0) = f(x).$$

Assumption $u(x,t) = X(x)T(t)$ leads to

$$XT' = c^2X''T \;\Rightarrow\; \frac{T'}{c^2T} = \frac{X''}{X} = -p^2,$$

giving rise to two ODE's as

$$X'' + p^2X = 0 \quad\text{and}\quad T' + c^2p^2T = 0.$$



# Parabolic Equations



BVP in the space coordinate $X'' + p^2X = 0,\ X(0) = X(L) = 0$ has solutions

$$X_n(x) = \sin\frac{n\pi x}{L}.$$

With $\lambda_n = \frac{cn\pi}{L}$, the ODE in $T(t)$ has the *corresponding solutions*

$$T_n(t) = A_n e^{-\lambda_n^2 t}.$$

By superposition,

$$u(x,t) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L} e^{-\lambda_n^2 t},$$

coefficients being determined from initial condition as

$$u(x,0) = f(x) = \sum_{n=1}^{\infty} A_n \sin\frac{n\pi x}{L},$$

a Fourier sine series.
As $t \to \infty$, $u(x,t) \to 0$ (steady state)



# Parabolic Equations



**Non-homogeneous boundary conditions**:

$$u_t = c^2u_{xx},\quad u(0,t) = u_1,\ u(L,t) = u_2,\quad u(x,0) = f(x).$$

For $u_1 \neq u_2$, with $u(x,t) = X(x)T(t)$, BC's do not separate!
Assume

$$u(x,t) = U(x,t) + u_{ss}(x),$$

where component $u_{ss}(x)$, steady-state temperature (distribution), does not enter the differential equation.

$$u_{ss}''(x) = 0,\quad u_{ss}(0) = u_1,\quad u_{ss}(L) = u_2 \;\Rightarrow\; u_{ss}(x) = u_1 + \frac{u_2-u_1}{L}x$$

Substituting into the BVP,

$$U_t = c^2U_{xx},\quad U(0,t) = U(L,t) = 0,\quad U(x,0) = f(x) - u_{ss}(x).$$

Final solution:

$$u(x,t) = \sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L} e^{-\lambda_n^2 t} + u_{ss}(x),$$

$B_n$ being coefficients of Fourier sine series of $f(x) - u_{ss}(x)$.

# Parabolic Equations



**Heat conduction in an infinite wire**

$$u_t = c^2 u_{xx}, \quad u(x,0) = f(x)$$

In place of $\frac{n\pi}{L}$, now we have continuous frequency $p$.
Solution as superposition of all frequencies:

$$u(x,t) = \int_0^\infty u_p(x,t)dp = \int_0^\infty [A(p)\cos px + B(p)\sin px]e^{-c^2p^2t}dp$$

Initial condition

$$u(x,0) = f(x) = \int_0^\infty [A(p)\cos px + B(p)\sin px]dp$$

gives the Fourier integral of $f(x)$ and amplitude functions

$$A(p) = \frac{1}{\pi}\int_{-\infty}^\infty f(v)\cos pv\, dv \quad \text{and} \quad B(p) = \frac{1}{\pi}\int_{-\infty}^\infty f(v)\sin pv\, dv.$$



# Parabolic Equations



**Solution using Fourier transforms**

$$u_t = c^2 u_{xx}, \quad u(x,0) = f(x)$$

Using derivative formula of Fourier transforms,

$$\mathcal{F}(u_t) = c^2(iw)^2\mathcal{F}(u) \;\Rightarrow\; \frac{\partial \hat{u}}{\partial t} = -c^2w^2\hat{u},$$

since variables $x$ and $t$ are independent.
Initial value problem in $\hat{u}(w,t)$:

$$\frac{\partial \hat{u}}{\partial t} = -c^2w^2\hat{u}, \quad \hat{u}(0) = \hat{f}(w)$$

Solution: $\hat{u}(w,t) = \hat{f}(w)e^{-c^2w^2t}$
Inverse Fourier transform gives solution of the original problem as

$$\begin{aligned} u(x,t) &= \mathcal{F}^{-1}\{\hat{u}(w,t)\} = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \hat{f}(w)e^{-c^2w^2t}e^{iwx}dw \\ \Rightarrow u(x,t) &= \frac{1}{\pi}\int_{-\infty}^\infty f(v)\int_0^\infty \cos(wx - wv)e^{-c^2w^2t}dw\,dv. \end{aligned}$$



# Elliptic Equations



Heat flow in a plate: two-dimensional heat equation

$$\frac{\partial u}{\partial t} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

Steady-state temperature distribution:

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$

Laplace's equation

**Steady-state heat flow in a rectangular plate:**

$$u_{xx} + u_{yy} = 0, \quad u(0,y) = u(a,y) = u(x,0) = 0, \; u(x,b) = f(x);$$

a Dirichlet problem over the domain $0 \le x \le a, 0 \le y \le b$.
Proposal $u(x,y) = X(x)Y(y)$ leads to

$$X''Y + XY'' = 0 \;\Rightarrow\; \frac{X''}{X} = -\frac{Y''}{Y} = -p^2.$$

Separated ODE's:

$$X'' + p^2X = 0 \quad \text{and} \quad Y'' - p^2Y = 0$$



# Elliptic Equations



From BVP $X'' + p^2X = 0$, $X(0) = X(a) = 0$, $X_n(x) = \sin\frac{n\pi x}{a}$.
Corresponding solution of $Y'' - p^2Y = 0$:

$$Y_n(y) = A_n\cosh\frac{n\pi y}{a} + B_n\sinh\frac{n\pi y}{a}$$

Condition $Y(0) = 0 \;\Rightarrow\; A_n = 0$, and

$$u_n(x,y) = B_n\sin\frac{n\pi x}{a}\sinh\frac{n\pi y}{a}$$

The complete solution:

$$u(x,y) = \sum_{n=1}^\infty B_n\sin\frac{n\pi x}{a}\sinh\frac{n\pi y}{a}$$

The last boundary condition $u(x,b) = f(x)$ fixes the coefficients from the Fourier sine series of $f(x)$.

**Note:** In the example, BC's on three sides were homogeneous.
How did it help? What if there are more non-homogeneous BC's?

# Elliptic Equations

**Steady-state heat flow with internal heat generation**

$$\nabla^2 u = \phi(x, y)$$

Poisson's equation

Separation of variables impossible!

Consider function $u(x, y)$ as

$$u(x, y) = u_h(x, y) + u_p(x, y)$$

Sequence of steps

- one particular solution $u_p(x, y)$ that may or may not satisfy some or all of the boundary conditions
- solution of the corresponding homogeneous equation, namely $u_{xx} + u_{yy} = 0$ for $u_h(x, y)$
  - such that $u = u_h + u_p$ satisfies all the boundary conditions



# Two-Dimensional Wave Equation

**Transverse vibration of a rectangular membrane:**

$$\frac{\partial^2 u}{\partial t^2} = c^2 \left( \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

A Cauchy problem of the membrane:

$$u_{tt} = c^2(u_{xx} + u_{yy}); \quad u(x, y, 0) = f(x, y), \quad u_t(x, y, 0) = g(x, y);$$
$$u(0, y, t) = u(a, y, t) = u(x, 0, t) = u(x, b, t) = 0.$$

Separate the time variable from the space variables:

$$u(x, y, t) = F(x, y)T(t) \;\Rightarrow\; \frac{F_{xx} + F_{yy}}{F} = \frac{T''}{c^2 T} = -\lambda^2$$

**Helmholtz equation:**

$$F_{xx} + F_{yy} + \lambda^2 F = 0$$



# Two-Dimensional Wave Equation

Assuming $F(x, y) = X(x)Y(y)$,

$$\frac{X''}{X} = -\frac{Y'' + \lambda^2 Y}{Y} = -\mu^2$$

$$\Rightarrow X'' + \mu^2 X = 0 \quad \text{and} \quad Y'' + \nu^2 Y = 0,$$

such that $\lambda = \sqrt{\mu^2 + \nu^2}$.

With BC's $X(0) = X(a) = 0$ and $Y(0) = Y(b) = 0$,

$$X_m(x) = \sin \frac{m\pi x}{a} \quad \text{and} \quad Y_n(y) = \sin \frac{n\pi y}{b}.$$

Corresponding values of $\lambda$ are

$$\lambda_{mn} = \sqrt{\left(\frac{m\pi}{a}\right)^2 + \left(\frac{n\pi}{b}\right)^2}$$

with solutions of $T'' + c^2\lambda^2 T = 0$ as

$$T_{mn}(t) = A_{mn} \cos c\lambda_{mn} t + B_{mn} \sin c\lambda_{mn} t.$$



# Two-Dimensional Wave Equation

Composing $X_m(x)$, $Y_n(y)$ and $T_{mn}(t)$ and superposing,

$$u(x, y, t) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} [A_{mn} \cos c\lambda_{mn} t + B_{mn} \sin c\lambda_{mn} t] \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b},$$

coefficients being determined from the double Fourier series

$$f(x, y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} A_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}$$

$$\text{and} \quad g(x, y) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} c\lambda_{mn} B_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}.$$

**BVP's modelled in polar coordinates**

For domains of circular symmetry, important in many practical systems, the BVP is conveniently modelled in polar coordinates,

*the separation of variables quite often producing*

- *Bessel's equation, in cylindrical coordinates, and*
- *Legendre's equation, in spherical coordinates*

## Points to note

- PDE's in physically relevant contexts
- Initial and boundary conditions
- Separation of variables
- Examples of boundary value problems with hyperbolic, parabolic and elliptic equations
  - Modelling, solution and interpretation
- Cascaded application of separation of variables for problems with more than two independent variables

Necessary Exercises: **1,2,4,7,9,10**



## Outline





## Analyticity of Complex Functions

**Function $f$ of a complex variable $z$**

*gives a rule to associate a unique complex number $w = u + iv$ to every $z = x + iy$ in a set.*

**Limit:** If $f(z)$ is defined in a neighbourhood of $z_0$ (except possibly at $z_0$ itself) and $\exists l \in C$ such that $\forall\ \epsilon > 0,\ \exists\ \delta > 0$ such that

$$0 < |z - z_0| < \delta \Rightarrow |f(z) - l| < \epsilon,$$

then

$$l = \lim_{z \to z_0} f(z).$$

Crucial difference from real functions: $z$ can approach $z_0$ in all possible manners in the complex plane.

*Definition of the limit is more restrictive.*

**Continuity:** $\lim_{z\to z_0} f(z) = f(z_0)$
Continuity in a domain $D$: continuity at every point in $D$



## Analyticity of Complex Functions

**Derivative of a complex function:**

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} = \lim_{\delta z \to 0} \frac{f(z_0 + \delta z) - f(z_0)}{\delta z}$$

When this limit exists, function $f(z)$ is said to be *differentiable*.

*Extremely restrictive definition!*

**Analytic function**

*A function $f(z)$ is called analytic in a domain $D$ if it is defined and differentiable at all points in $D$.*

Points to be settled later:

- Derivative of an analytic function is also analytic.
- An analytic function possesses derivatives of all orders.

A great **qualitative** difference between functions of a real variable and those of a complex variable!

## Analyticity of Complex Functions



**Cauchy-Riemann conditions**

If $f(z) = u(x, y) + iv(x, y)$ is analytic then

$$f'(z) = \lim_{\delta x, \delta y \to 0} \frac{\delta u + i\delta v}{\delta x + i\delta y}$$

along all paths of approach for $\delta z = \delta x + i\delta y \to 0$ or $\delta x, \delta y \to 0$.

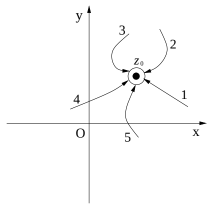



Figure: Paths approaching $z_0$

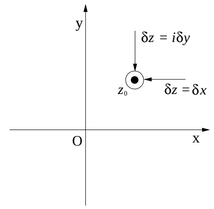



Figure: Paths in C-R equations

Two expressions for the derivative:

$$f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = \frac{\partial v}{\partial y} - i\frac{\partial u}{\partial y}$$



## Analyticity of Complex Functions



Cauchy-Riemann equations or conditions

$$\boxed{\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} \quad \text{and} \quad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}}$$

are *necessary* for analyticity.

**Question:** Do the C-R conditions *imply* analyticity?

Consider $u(x, y)$ and $v(x, y)$ having continuous first order partial derivatives that satisfy the Cauchy-Riemann conditions.

By mean value theorem,

$$\delta u = u(x + \delta x, y + \delta y) - u(x, y) = \delta x \frac{\partial u}{\partial x}(x_1, y_1) + \delta y \frac{\partial u}{\partial y}(x_1, y_1)$$

with $x_1 = x + \xi\delta x, y_1 = y + \xi\delta y$ for some $\xi \in [0, 1]$; and

$$\delta v = v(x + \delta x, y + \delta y) - v(x, y) = \delta x \frac{\partial v}{\partial x}(x_2, y_2) + \delta y \frac{\partial v}{\partial y}(x_2, y_2)$$

with $x_2 = x + \eta\delta x, y_2 = y + \eta\delta y$ for some $\eta \in [0, 1]$.

Then,

$$\delta f = \left[\delta x \frac{\partial u}{\partial x}(x_1, y_1) + i\delta y \frac{\partial v}{\partial y}(x_2, y_2)\right] + i\left[\delta x \frac{\partial v}{\partial x}(x_2, y_2) - i\delta y \frac{\partial u}{\partial y}(x_1, y_1)\right]$$



## Analyticity of Complex Functions



Using C-R conditions $\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$,

$$\begin{aligned} \delta f &= (\delta x + i\delta y)\frac{\partial u}{\partial x}(x_1, y_1) + i\delta y\left[\frac{\partial u}{\partial x}(x_2, y_2) - \frac{\partial u}{\partial x}(x_1, y_1)\right] \\ &\quad + i(\delta x + i\delta y)\frac{\partial v}{\partial x}(x_1, y_1) + i\delta x\left[\frac{\partial v}{\partial x}(x_2, y_2) - \frac{\partial v}{\partial x}(x_1, y_1)\right] \\ \Rightarrow \frac{\delta f}{\delta z} &= \frac{\partial u}{\partial x}(x_1, y_1) + i\frac{\partial v}{\partial x}(x_1, y_1) + \\ &\quad i\frac{\delta x}{\delta z}\left[\frac{\partial v}{\partial x}(x_2, y_2) - \frac{\partial v}{\partial x}(x_1, y_1)\right] + i\frac{\delta y}{\delta z}\left[\frac{\partial u}{\partial x}(x_2, y_2) - \frac{\partial u}{\partial x}(x_1, y_1)\right] \end{aligned}$$

Since $\left|\frac{\delta x}{\delta z}\right|, \left|\frac{\delta y}{\delta z}\right| \leq 1$, as $\delta z \to 0$, the limit exists and

$$f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x} = -i\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}.$$

*Cauchy-Riemann conditions are necessary and sufficient for function $w = f(z) = u(x, y) + iv(x, y)$ to be analytic.*



## Analyticity of Complex Functions



**Harmonic function**

Differentiating C-R equations $\frac{\partial v}{\partial y} = \frac{\partial u}{\partial x}$ and $\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}$,

$$\frac{\partial^2 u}{\partial x^2} = \frac{\partial^2 v}{\partial x \partial y}, \quad \frac{\partial^2 u}{\partial y^2} = -\frac{\partial^2 v}{\partial y \partial x}, \quad \frac{\partial^2 u}{\partial y \partial x} = \frac{\partial^2 v}{\partial y^2}, \quad \frac{\partial^2 u}{\partial x \partial y} = -\frac{\partial^2 v}{\partial x^2}$$

$$\Rightarrow \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 = \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2}.$$

*Real and imaginary components of an analytic functions are* harmonic functions.

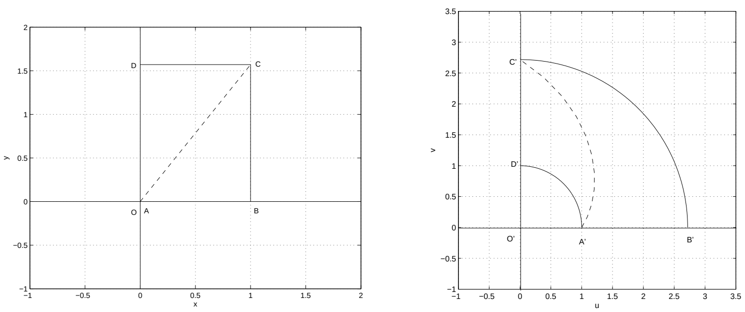
**Conjugate** harmonic function of $u(x, y)$: $v(x, y)$

Families of curves $u(x, y) = c$ and $v(x, y) = k$ are mutually orthogonal, except possibly at points where $f'(z) = 0$.

**Question:** If $u(x, y)$ is given, then how to develop the complete analytic function $w = f(z) = u(x, y) + iv(x, y)$?

# Conformal Mapping

Function: mapping of elements in domain to their images in range
Depiction of a complex variable requires a plane with two axes.
*Mapping of a complex function $w = f(z)$ is shown in two planes.*
**Example:** mapping of a rectangle under transformation $w = e^z$




(a) The $z$-plane (b) The $w$-plane

Figure: Mapping corresponding to function $w = e^z$



# Conformal Mapping

**Conformal mapping:** a mapping that preserves the angle between any two directions in magnitude and sense.
**Verify:** $w = e^z$ defines a conformal mapping.

> *Through* relative orientations *of curves at the points of intersection, 'local' shape of a figure is preserved.*

Take curve $z(t), z(0) = z_0$ and image $w(t) = f[z(t)], w_0 = f(z_0)$.
For analytic $f(z)$, $\dot{w}(0) = f'(z_0)\dot{z}(0)$, implying

$$|\dot{w}(0)| = |f'(z_0)|\,|\dot{z}(0)| \quad \text{and} \quad \arg \dot{w}(0) = \arg f'(z_0) + \arg \dot{z}(0).$$

For several curves through $z_0$,

> *image curves pass through $w_0$ and all of them turn by the same angle* $\arg f'(z_0)$.

Cautions

- $f'(z)$ varies from point to point. Different scaling and turning effects take place at different points. 'Global' shape changes.
- For $f'(z) = 0$, argument is undefined and conformality is lost.



# Conformal Mapping

> *An analytic function defines a conformal mapping except at its critical points where its derivative vanishes.*

Except at critical points, an analytic function is invertible.

> *We can establish an inverse of any conformal mapping.*

**Examples**

- Linear function $w = az + b$ (for $a \neq 0$)
- Linear fractional transformation

$$w = \frac{az + b}{cz + d}, \quad ad - bc \neq 0$$

- Other elementary functions like $z^n, e^z$ etc

Special significance of conformal mappings:

> *A harmonic function $\phi(u, v)$ in the $w$-plane is also a harmonic function, in the form $\phi(x, y)$ in the $z$-plane, as long as the two planes are related through a conformal mapping.*



# Potential Theory

**Riemann mapping theorem:** Let $D$ be a simply connected domain in the $z$-plane bounded by a closed curve $C$. Then there exists a conformal mapping that gives a one-to-one correspondence between $D$ and the unit disc $|w| < 1$ as well as between $C$ and the unit circle $|w| = 1$, bounding the unit disc.

**Application to boundary value problems**

- First, establish a conformal mapping between the given domain and a domain of simple geometry.
- Next, solve the BVP in this simple domain.
- Finally, using the inverse of the conformal mapping, construct the solution for the given domain.

**Example:** Dirichlet problem with Poisson's integral formula

$$f(re^{i\theta}) = \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2) f(Re^{i\phi})}{R^2 - 2Rr\cos(\theta - \phi) + r^2} d\phi$$

# Potential Theory

**Two-dimensional potential flow**

- Velocity potential $\phi(x, y)$ gives velocity components $V_x = \frac{\partial \phi}{\partial x}$ and $V_y = \frac{\partial \phi}{\partial y}$.
- A streamline is a curve in the flow field, the tangent to which at any point is along the local velocity vector.
- Stream function $\psi(x, y)$ remains constant along a streamline.
- $\psi(x, y)$ is the conjugate harmonic function of $\phi(x, y)$.
- Complex potential function $\Phi(z) = \phi(x, y) + i\psi(x, y)$ defines the flow.

If a flow field encounters a solid boundary of a complicated shape,

*transform the boundary conformally to a simple boundary*

to facilitate the study of the flow pattern.



# Points to note

- Analytic functions and Cauchy-Riemann conditions
- Conformality of analytic functions
- Applications in solving BVP's and flow description

Necessary Exercises: **1,2,3,4,7,9**



# Outline





# Line Integral

For $w = f(z) = u(x, y) + iv(x, y)$, over a smooth curve $C$,

$$\int_C f(z)dz = \int_C (u+iv)(dx+idy) = \int_C (udx-vdy)+i\int_C (vdx+udy).$$

Extension to piecewise smooth curves is obvious.

With parametrization, for $z = z(t), a \leq t \leq b$, with $\dot{z}(t) \neq 0$,

$$\int_C f(z)dz = \int_a^b f[z(t)]\dot{z}(t)dt.$$

Over a simple closed curve, *contour integral:* $\oint_C f(z)dz$

**Example:** $\oint_C z^n dz$ for integer $n$, around circle $z = \rho e^{i\theta}$

$$\oint_C z^n dz = i\rho^{n+1}\int_0^{2\pi} e^{i(n+1)\theta}d\theta = \begin{cases} 0 & \text{for } n \neq -1, \\ 2\pi i & \text{for } n = -1. \end{cases}$$

**The** $M$-$L$ **inequality:** If $C$ is a curve of finite length $L$ and $|f(z)| < M$ on $C$, then

$$\left|\int_C f(z)dz\right| \leq \int_C |f(z)|\,|dz| < M\int_C |dz| = ML.$$

## Cauchy's Integral Theorem

- $C$ is a simple closed curve in a simply connected domain $D$.
- Function $f(z) = u + iv$ is analytic in $D$.

Contour integral $\oint_C f(z)dz = ?$
If $f'(z)$ is continuous, then by Green's theorem in the plane,

$$\oint_C f(z)dz = \int_R\int\left(-\frac{\partial v}{\partial x}-\frac{\partial u}{\partial y}\right)dxdy + i\int_R\int\left(\frac{\partial u}{\partial x}-\frac{\partial v}{\partial y}\right)dxdy,$$

where $R$ is the region enclosed by $C$.
From C-R conditions, $\boxed{\oint_C f(z)dz = 0.}$
**Proof by Goursat:** *without* the hypothesis of continuity of $f'(z)$

**Cauchy-Goursat theorem**

*If $f(z)$ is analytic in a simply connected domain $D$, then $\oint_C f(z)dz = 0$ for every simple closed curve $C$ in $D$.*

Importance of Goursat's contribution:

- continuity of $f'(z)$ appears as *consequence*!



## Cauchy's Integral Theorem

**Principle of path independence**
Two points $z_1$ and $z_2$ on the close curve $C$

- two open paths $C_1$ and $C_2$ from $z_1$ to $z_2$

Cauchy's theorem on $C$, comprising of $C_1$ in the forward direction and $C_2$ in the reverse direction:

$$\int_{C_1} f(z)dz - \int_{C_2} f(z)dz = 0 \Rightarrow \int_{z_1}^{z_2} f(z)dz = \int_{C_1} f(z)dz = \int_{C_2} f(z)dz$$

*For an analytic function $f(z)$ in a simply connected domain $D$, $\int_{z_1}^{z_2} f(z)dz$ is independent of the path and depends only on the end-points, as long as the path is completely contained in $D$.*

Consequence: Definition of the function

$$F(z) = \int_{z_0}^{z} f(\xi)d\xi$$

What does the formulation *suggest*?



## Cauchy's Integral Theorem

**Indefinite integral**
**Question:** Is $F(z)$ analytic? Is $F'(z) = f(z)$?

$$\begin{aligned}\frac{F(z+\delta z)-F(z)}{\delta z} - f(z) &= \frac{1}{\delta z}\left[\int_{z_0}^{z+\delta z} f(\xi)d\xi - \int_{z_0}^{z} f(\xi)d\xi\right] - f(z)\\ &= \frac{1}{\delta z}\int_{z}^{z+\delta z}[f(\xi)-f(z)]d\xi\end{aligned}$$

$f$ is continuous $\Rightarrow \forall\epsilon, \exists\delta$ such that $|\xi - z| < \delta \Rightarrow |f(\xi) - f(z)| < \epsilon$
Choosing $\delta z < \delta$,

$$\left|\frac{F(z+\delta z)-F(z)}{\delta z} - f(z)\right| < \frac{\epsilon}{\delta z}\int_{z}^{z+\delta z} d\xi = \epsilon.$$

*If $f(z)$ is analytic in a simply connected domain $D$, then there exists an analytic function $F(z)$ in $D$ such that*

$$F'(z) = f(z) \;\; \textit{and} \;\; \int_{z_1}^{z_2} f(z)dz = F(z_2) - F(z_1).$$



## Cauchy's Integral Theorem

**Principle of deformation of paths**

$f(z)$ analytic everywhere other than isolated points $s_1$, $s_2$, $s_3$

$$\int_{C_1} f(z)dz = \int_{C_2} f(z)dz = \int_{C_3} f(z)dz$$

Not so for path $C^*$.

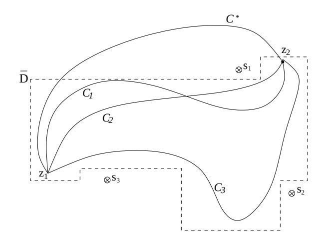



Figure: Path deformation

*The line integral remains unaltered through a continuous deformation of the path of integration with fixed end-points, as long as the sweep of the deformation includes no point where the integrand is non-analytic.*

# Cauchy's Integral Theorem

**Cauchy's theorem in multiply connected domain**

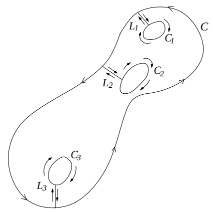



Figure: Contour for multiply connected domain

$$\oint_C f(z)dz - \oint_{C_1} f(z)dz - \oint_{C_2} f(z)dz - \oint_{C_3} f(z)dz = 0.$$

*If $f(z)$ is analytic in a region bounded by the contour $C$ as the outer boundary and non-overlapping contours $C_1$, $C_2$, $C_3$, $\cdots$, $C_n$ as inner boundaries, then*

$$\oint_C f(z)dz = \sum_{i=1}^{n} \oint_{C_i} f(z)dz.$$



# Cauchy's Integral Formula

$f(z)$: analytic function in a simply connected domain $D$

For $z_0 \in D$ and simple closed curve $C$ in $D$,

$$\oint_C \frac{f(z)}{z - z_0} dz = 2\pi i f(z_0).$$

Consider $C$ as a circle with centre at $z_0$ and radius $\rho$,

*with no loss of generality (why?).*

$$\oint_C \frac{f(z)}{z - z_0} dz = f(z_0) \oint_C \frac{dz}{z - z_0} + \oint_C \frac{f(z) - f(z_0)}{z - z_0} dz$$

From continuity of $f(z)$, $\exists \delta$ such that for any $\epsilon$,

$$|z - z_0| < \delta \Rightarrow |f(z) - f(z_0)| < \epsilon \text{ and } \left| \frac{f(z) - f(z_0)}{z - z_0} \right| < \frac{\epsilon}{\rho},$$

with $\rho < \delta$. From *M-L* inequality, the second integral vanishes.



# Cauchy's Integral Formula

**Direct applications**

- **Evaluation of contour integral:**
  - If $g(z)$ is analytic on the contour and in the enclosed region, the Cauchy's theorem implies $\oint_C g(z)dz = 0$.
  - If the contour encloses a singularity at $z_0$, then Cauchy's formula supplies a non-zero contribution to the integral, if $f(z) = g(z)(z - z_0)$ is analytic.
- **Evaluation of function at a point:** If finding the integral on the left-hand-side is relatively simple, then we use it to *evaluate* $f(z_0)$.

*Significant in the solution of boundary value problems!*

Example: Poisson's integral formula

$$u(r,\theta) = \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2)u(R,\phi)}{R^2 - 2Rr\cos(\theta - \phi) + r^2} d\phi$$

for the Dirichlet problem over a circular disc.



# Cauchy's Integral Formula

**Poisson's integral formula**

Taking $z_0 = re^{i\theta}$ and $z = Re^{i\phi}$ (with $r < R$) in Cauchy's formula,

$$2\pi i f(re^{i\theta}) = \int_0^{2\pi} \frac{f(Re^{i\phi})}{Re^{i\phi} - re^{i\theta}} (iRe^{i\phi}) d\phi.$$

How to get rid of imaginary quantities from the expression?
Develop a complement. With $\frac{R^2}{r}$ in place of $r$,

$$0 = \int_0^{2\pi} \frac{f(Re^{i\phi})}{Re^{i\phi} - \frac{R^2}{r} e^{i\theta}} (iRe^{i\phi}) d\phi = \int_0^{2\pi} \frac{f(Re^{i\phi})}{re^{-i\theta} - Re^{-i\phi}} (ire^{-i\theta}) d\phi.$$

Subtracting,

$$\begin{aligned} 2\pi i f(re^{i\theta}) &= i \int_0^{2\pi} f(Re^{i\phi}) \left[ \frac{Re^{i\phi}}{Re^{i\phi} - re^{i\theta}} + \frac{re^{-i\theta}}{Re^{-i\phi} - re^{-i\theta}} \right] d\phi \\ &= i \int_0^{2\pi} \frac{(R^2 - r^2) f(Re^{i\phi})}{(Re^{i\phi} - re^{i\theta})(Re^{-i\phi} - re^{-i\theta})} d\phi \\ \Rightarrow f(re^{i\theta}) &= \frac{1}{2\pi} \int_0^{2\pi} \frac{(R^2 - r^2) f(Re^{i\phi})}{R^2 - 2Rr\cos(\theta - \phi) + r^2} d\phi. \end{aligned}$$

## Cauchy's Integral Formula



Cauchy's integral formula evaluates contour integral of $g(z)$,

*if the contour encloses a point $z_0$ where $g(z)$ is non-analytic but $g(z)(z-z_0)$ is analytic.*

If $g(z)(z-z_0)$ is also non-analytic, but $g(z)(z-z_0)^2$ is analytic?

$$\begin{aligned}
f(z_0) &= \frac{1}{2\pi i}\oint_C \frac{f(z)}{z-z_0}dz,\\
f'(z_0) &= \frac{1}{2\pi i}\oint_C \frac{f(z)}{(z-z_0)^2}dz,\\
f''(z_0) &= \frac{2!}{2\pi i}\oint_C \frac{f(z)}{(z-z_0)^3}dz,\\
\cdots &= \cdots \quad \cdots \quad \cdots,\\
f^{(n)}(z_0) &= \frac{n!}{2\pi i}\oint_C \frac{f(z)}{(z-z_0)^{n+1}}dz.
\end{aligned}$$

The formal expressions can be established through differentiation under the integral sign.



## Cauchy's Integral Formula



$$\begin{aligned}
\frac{f(z_0+\delta z)-f(z_0)}{\delta z} &= \frac{1}{2\pi i\delta z}\oint_C f(z)\left[\frac{1}{z-z_0-\delta z}-\frac{1}{z-z_0}\right]dz\\
&= \frac{1}{2\pi i}\oint_C \frac{f(z)dz}{(z-z_0-\delta z)(z-z_0)}
\end{aligned}$$

$$= \frac{1}{2\pi i}\oint_C \frac{f(z)dz}{(z-z_0)^2} + \frac{1}{2\pi i}\oint_C f(z)\left[\frac{1}{(z-z_0-\delta z)(z-z_0)}-\frac{1}{(z-z_0)^2}\right]dz$$

$$= \frac{1}{2\pi i}\oint_C \frac{f(z)dz}{(z-z_0)^2} + \frac{1}{2\pi i}\delta z\oint_C \frac{f(z)dz}{(z-z_0-\delta z)(z-z_0)^2}$$

If $|f(z)| < M$ on $C$, $L$ is path length and $d_0 = \min|z-z_0|$,

$$\left|\delta z\oint_C \frac{f(z)dz}{(z-z_0-\delta z)(z-z_0)^2}\right| < \frac{ML|\delta z|}{d_0^2(d_0-|\delta z|)} \to 0 \quad \text{as} \quad \delta z \to 0.$$

*An analytic function possesses derivatives of all orders at every point in its domain.*

Analyticity implies much more than mere differentiability!



## Points to note



- Concept of line integral in complex plane
- Cauchy's integral theorem
- Consequences of analyticity
- Cauchy's integral formula
- Derivatives of arbitrary order for analytic functions

Necessary Exercises: **1,2,5,7**



## Outline

## Series Representations of Complex Functions

**Taylor's series** of function $f(z)$, analytic in a neighbourhood of $z_0$:

$$f(z)=\sum_{n=0}^{\infty}a_n(z-z_0)^n=a_0+a_1(z-z_0)+a_2(z-z_0)^2+a_3(z-z_0)^3+\cdots,$$

with coefficients

$$a_n=\frac{1}{n!}f^{(n)}(z_0)=\frac{1}{2\pi i}\oint_C\frac{f(w)dw}{(w-z_0)^{n+1}},$$

where $C$ is a circle with centre at $z_0$.

Form of the series and coefficients: similar to real functions

*The series representation is* convergent *within a disc $|z-z_0|<R$, where radius of convergence $R$ is the distance of the nearest singularity from $z_0$.*

**Note:** No valid power series representation around $z_0$, i.e. in powers of $(z-z_0)$, if f(z) is not analytic at $z_0$

**Question:** In that case, what about a series representation that includes *negative* powers of $(z-z_0)$ as well?



## Series Representations of Complex Functions

**Laurent's series:** If $f(z)$ is analytic on circles $C_1$ (outer) and $C_2$ (inner) with centre at $z_0$, and in the annulus in between, then

$$f(z)=\sum_{n=-\infty}^{\infty}a_n(z-z_0)^n=\sum_{m=0}^{\infty}b_m(z-z_0)^m+\sum_{m=1}^{\infty}\frac{c_m}{(z-z_0)^m};$$

with coefficients

$$a_n=\frac{1}{2\pi i}\oint_C\frac{f(w)dw}{(w-z_0)^{n+1}};$$

or, $$b_m=\frac{1}{2\pi i}\oint_C\frac{f(w)dw}{(w-z_0)^{m+1}},\quad c_m=\frac{1}{2\pi i}\oint_C f(w)(w-z_0)^{m-1}dw;$$

the contour $C$ lying in the annulus and enclosing $C_2$.

*Validity of this series representation:* in annular region obtained by growing $C_1$ and shrinking $C_2$ till $f(z)$ ceases to be analytic.

*Observation:* If $f(z)$ is analytic inside $C_2$ as well, then $c_m=0$ and Laurent's series reduces to Taylor's series.



## Series Representations of Complex Functions

**Proof of Laurent's series**

Cauchy's integral formula for any point $z$ in the annulus,

$$f(z)=\frac{1}{2\pi i}\oint_{C_1}\frac{f(w)dw}{w-z}-\frac{1}{2\pi i}\oint_{C_2}\frac{f(w)dw}{w-z}.$$

Organization of the series:

$$\frac{1}{w-z}=\frac{1}{(w-z_0)[1-(z-z_0)/(w-z_0)]}$$
$$\frac{1}{w-z}=-\frac{1}{(z-z_0)[1-(w-z_0)/(z-z_0)]}$$

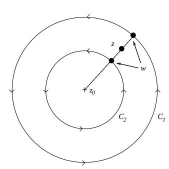


Figure: The annulus

Using the expression for the sum of a geometric series,

$$1+q+q^2+\cdots+q^{n-1}=\frac{1-q^n}{1-q}\Rightarrow\frac{1}{1-q}=1+q+q^2+\cdots+q^{n-1}+\frac{q^n}{1-q}.$$

We use $q=\frac{z-z_0}{w-z_0}$ for integral over $C_1$ and $q=\frac{w-z_0}{z-z_0}$ over $C_2$.



## Series Representations of Complex Functions

**Proof of Laurent's series** (contd)

Using $q=\frac{z-z_0}{w-z_0}$,

$$\frac{1}{w-z}=\frac{1}{w-z_0}+\frac{z-z_0}{(w-z_0)^2}+\cdots+\frac{(z-z_0)^{n-1}}{(w-z_0)^n}+\left(\frac{z-z_0}{w-z_0}\right)^n\frac{1}{w-z}$$

$$\Rightarrow\frac{1}{2\pi i}\oint_{C_1}\frac{f(w)dw}{w-z}=a_0+a_1(z-z_0)+\cdots+a_{n-1}(z-z_0)^{n-1}+T_n,$$

with coefficients as required and

$$T_n=\frac{1}{2\pi i}\oint_{C_1}\left(\frac{z-z_0}{w-z_0}\right)^n\frac{f(w)}{w-z}dw.$$

Similarly, with $q=\frac{w-z_0}{z-z_0}$,

$$-\frac{1}{2\pi i}\oint_{C_2}\frac{f(w)dw}{w-z}=a_{-1}(z-z_0)^{-1}+\cdots+a_{-n}(z-z_0)^{-n}+T_{-n},$$

with appropriate coefficients and the remainder term

$$T_{-n}=\frac{1}{2\pi i}\oint_{C_2}\left(\frac{w-z_0}{z-z_0}\right)^n\frac{f(w)}{z-w}dw.$$

## Series Representations of Complex Functions

**Convergence of Laurent's series**

$$f(z) = \sum_{k=-n}^{n-1} a_k (z-z_0)^k + T_n + T_{-n},$$

*where*

$$T_n = \frac{1}{2\pi i}\oint_{C_1}\left(\frac{z-z_0}{w-z_0}\right)^n \frac{f(w)}{w-z}dw$$

$$\text{and}\quad T_{-n} = \frac{1}{2\pi i}\oint_{C_2}\left(\frac{w-z_0}{z-z_0}\right)^n \frac{f(w)}{z-w}dw.$$

- $f(w)$ is bounded
- $\left|\frac{z-z_0}{w-z_0}\right| < 1$ over $C_1$ and $\left|\frac{w-z_0}{z-z_0}\right| < 1$ over $C_2$

Use *M-L* inequality to show that

*remainder terms $T_n$ and $T_{-n}$ approach zero as $n \to \infty$.*

**Remark:** For actually developing Taylor's or Laurent's series of a function, algebraic manipulation of known facts are employed quite often, rather than evaluating so many contour integrals!



## Zeros and Singularities

**Zeros** of an analytic function: points where the function vanishes

If, at a point $z_0$,

*a function $f(z)$ vanishes along with first $m-1$ of its derivatives, but $f^{(m)}(z_0) \neq 0$;*

then $z_0$ is a zero of $f(z)$ of order $m$, giving the Taylor's series as

$$f(z) = (z-z_0)^m g(z).$$

An *isolated* zero has a neighbourhood containing no other zero.

*For an analytic function, not identically zero, every point has a neighbourhood free of zeros of the function, except possibly for that point itself. In particular, zeros of such an analytic function are always isolated.*

**Implication:** If $f(z)$ has a zero in every neighbourhood around $z_0$ then it cannot be analytic at $z_0$, unless it is the zero function [i.e. $f(z) = 0$ everywhere].



## Zeros and Singularities

**Entire function:** A function which is analytic everywhere

Examples: $z^n$ (for positive integer $n$), $e^z$, $\sin z$ etc.

*The Taylor's series of an entire function has an infinite radius of convergence.*

**Singularities:** points where a function ceases to be analytic

Removable singularity: If $f(z)$ is not defined at $z_0$, but has a limit.
Example: $f(z) = \frac{e^z - 1}{z}$ at $z = 0$.

Pole: If $f(z)$ has a Laurent's series around $z_0$, with a finite number of terms with negative powers. If $a_n = 0$ for $n < -m$, but $a_{-m} \neq 0$, then $z_0$ is a pole of order $m$, $\lim_{z\to z_0}(z-z_0)^m f(z)$ being a non-zero finite number.
A simple pole: a pole of order one.

Essential singularity: A singularity which is neither a removable singularity nor a pole. If the function has a Laurent's series, then it has infinite terms with negative powers. Example: $f(z) = e^{1/z}$ at $z = 0$.



## Zeros and Singularities

**Zeros and poles**: complementary to each other

- Poles are necessarily *isolated* singularities.
- A zero of $f(z)$ of order $m$ is a pole of $\frac{1}{f(z)}$ of the same order and vice versa.
- If $f(z)$ has a zero of order $m$ at $z_0$ where $g(z)$ has a pole of the same order, then $f(z)g(z)$ is either analytic at $z_0$ or has a removable singularity there.
- **Argument theorem:**

  *If $f(z)$ is analytic inside and on a simple closed curve $C$ except for a finite number of poles inside and $f(z) \neq 0$ on $C$, then*

$$\frac{1}{2\pi i}\oint_C \frac{f'(z)}{f(z)}dz = N - P,$$

  *where $N$ and $P$ are total numbers of zeros and poles inside $C$ respectively, counting multiplicities (orders).*

## Residues

Term by term integration of Laurent's series: $\oint_C f(z)dz = 2\pi i a_{-1}$

**Residue**: $\underset{z_0}{\text{Res}} f(z) = a_{-1} = \frac{1}{2\pi i}\oint_C f(z)dz$

If $f(z)$ has a pole (of order $m$) at $z_0$, then

$$(z-z_0)^m f(z) = \sum_{n=-m}^{\infty} a_n (z-z_0)^{m+n}$$

is analytic at $z_0$, and

$$\frac{d^{m-1}}{dz^{m-1}}[(z-z_0)^m f(z)] = \sum_{n=-1}^{\infty} \frac{(m+n)!}{(n+1)!} a_n (z-z_0)^{n+1}$$

$$\Rightarrow \quad \underset{z_0}{\text{Res}} f(z) = a_{-1} = \frac{1}{(m-1)!}\lim_{z\to z_0}\frac{d^{m-1}}{dz^{m-1}}[(z-z_0)^m f(z)].$$

**Residue theorem:** If $f(z)$ is analytic inside and on simple closed curve $C$, with singularities at $z_1, z_2, z_3, \cdots, z_k$ inside $C$; then

$$\oint_C f(z)dz = 2\pi i \sum_{i=1}^{k} \underset{z_i}{\text{Res}} f(z).$$



## Evaluation of Real Integrals

**General strategy**

- Identify the required integral as a contour integral of a complex function, or a part thereof.
- If the domain of integration is infinite, then extend the contour infinitely, without enclosing new singularities.

**Example:**

$$I = \int_0^{2\pi} \phi(\cos\theta, \sin\theta)d\theta$$

With $z = e^{i\theta}$ and $dz = izd\theta$,

$$I = \oint_C \phi\left[\frac{1}{2}\left(z+\frac{1}{z}\right), \frac{1}{2i}\left(z-\frac{1}{z}\right)\right]\frac{dz}{iz} = \oint_C f(z)dz,$$

where $C$ is the unit circle centred at the origin.
Denoting poles falling inside the unit circle $C$ as $p_j$,

$$I = 2\pi i \sum_j \underset{p_j}{\text{Res}} f(z).$$



## Evaluation of Real Integrals

**Example:** For real rational function $f(x)$,

$$I = \int_{-\infty}^{\infty} f(x)dx,$$

denominator of $f(x)$ being of degree two higher than numerator.

Consider contour $C$ enclosing semi-circular region $|z| \le R, y \ge 0$, large enough to enclose all singularities above the $x$-axis.

$$\oint_C f(z)dz = \int_{-R}^{R} f(x)dx + \int_S f(z)dz$$

For finite $M$, $|f(z)| < \frac{M}{R^2}$ on $C$

$$\left|\int_S f(z)dz\right| < \frac{M}{R^2}\pi R = \frac{\pi M}{R}.$$

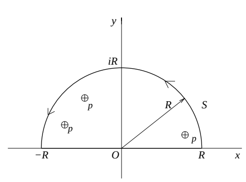



Figure: The contour

$$I = \int_{-\infty}^{\infty} f(x)dx = 2\pi i \sum_j \underset{p_j}{\text{Res}} f(z) \quad \text{as} \quad R \to \infty.$$



## Evaluation of Real Integrals

**Example:** Fourier integral coefficients

$$A(s) = \int_{-\infty}^{\infty} f(x)\cos sx\, dx \quad \text{and} \quad B(s) = \int_{-\infty}^{\infty} f(x)\sin sx\, dx$$

Consider

$$I = A(s) + iB(s) = \int_{-\infty}^{\infty} f(x)e^{isx}dx.$$

Similar to the previous case,

$$\oint_C f(z)e^{isz}dz = \int_{-R}^{R} f(x)e^{isx}dx + \int_S f(z)e^{isz}dz.$$

As $|e^{isz}| = |e^{isx}|\,|e^{-sy}| = |e^{-sy}| \le 1$ for $y \ge 0$, we have

$$\left|\int_S f(z)e^{isz}dz\right| < \frac{M}{R^2}\pi R = \frac{\pi M}{R},$$

which yields, as $R \to \infty$,

$$I = 2\pi i \sum_j \underset{p_j}{\text{Res}}[f(z)e^{isz}].$$

# Points to note

- Taylor's series and Laurent's series
- Zeros and poles of analytic functions
- Residue theorem
- Evaluation of real integrals through contour integration of suitable complex functions

Necessary Exercises: **1,2,3,5,8,9,10**



# Outline





# Introduction

Consider a particle moving on a smooth surface $z = \psi(q_1, q_2)$.

With position $\mathbf{r} = [q_1(t)\ q_2(t)\ \psi(q_1(t), q_2(t))]^T$ on the surface and $\delta\mathbf{r} = [\delta q_1\ \delta q_2\ (\nabla\psi)^T\delta\mathbf{q}]^T$ in the tangent plane, length of the path from $\mathbf{q}_i = \mathbf{q}(t_i)$ to $\mathbf{q}_f = \mathbf{q}(t_f)$ is

$$l = \int \|\delta\mathbf{r}\| = \int_{t_i}^{t_f} \|\dot{\mathbf{r}}\| dt = \int_{t_i}^{t_f} \left[\dot{q}_1^2 + \dot{q}_2^2 + (\nabla\psi^T\dot{\mathbf{q}})^2\right]^{1/2} dt.$$

For shortest path or geodesic, minimize the path length $l$.

**Question:** What are the variables of the problem?

**Answer:** The entire curve or function $\mathbf{q}(t)$.

**Variational problem:**
Optimization of a function of *functions*, i.e. a *functional*.



# Introduction

**Functionals and their extremization**

Suppose that a candidate curve is represented as a sequence of points $\mathbf{q}_j = \mathbf{q}(t_j)$ at time instants

$$t_i = t_0 < t_1 < t_2 < t_3 < \cdots < t_{N-1} < t_N = t_f.$$

**Geodesic problem:** a multivariate optimization problem with the $2(N-1)$ variables in $\{\mathbf{q}_j, 1 \le j \le N-1\}$.

*With $N \to \infty$, we obtain the actual function.*

First order necessary condition: Functional is stationary with respect to *arbitrary* small variations in $\{\mathbf{q}_j\}$.

*[Equivalent to vanishing of the gradient]*

This gives *equations* for the stationary points.

Here, these equations are *differential equations*!

## Introduction



**Examples of variational problems**

Geodesic path: Minimize $I = \int_a^b \|\mathbf{r}'(t)\| dt$

Minimal surface of revolution: Minimize
$$S = \int 2\pi y ds = 2\pi \int_a^b y\sqrt{1+y'^2}dx$$

The brachistochrone problem: To find the curve along which the descent is fastest.
Minimize $T = \int \frac{ds}{v} = \int_a^b \sqrt{\frac{1+y'^2}{2gy}}dx$

Fermat's principle: Light takes the fastest path.
Minimize $T = \int_{u_1}^{u_2} \frac{\sqrt{x'^2+y'^2+z'^2}}{c(x,y,z)}du$

Isoperimetric problem: Largest area in the plane enclosed by a closed curve of given perimeter. By extension, extremize a functional under one or more equality constraints.

Hamilton's principle of least action: Evolution of a dynamic system through the minimization of the action
$$s = \int_{t_1}^{t_2} L dt = \int_{t_1}^{t_2} (K-P)dt$$



## Euler's Equation



Find out a function $y(x)$, that will make the functional
$$I[y(x)] = \int_{x_1}^{x_2} f[x, y(x), y'(x)]dx$$
stationary, with boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$.
Consider variation $\delta y(x)$ with $\delta y(x_1) = \delta y(x_2) = 0$ and *consistent* variation $\delta y'(x)$.
$$\delta I = \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial y'}\delta y'\right) dx$$
Integration of the second term by parts:
$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\delta y' dx = \int_{x_1}^{x_2} \frac{\partial f}{\partial y'}\frac{d}{dx}(\delta y)dx = \left[\frac{\partial f}{\partial y'}\delta y\right]_{x_1}^{x_2} - \int_{x_1}^{x_2} \frac{d}{dx}\frac{\partial f}{\partial y'}\delta y\, dx$$
With $\delta y(x_1) = \delta y(x_2) = 0$, the first term vanishes identically, and
$$\delta I = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'}\right]\delta y\, dx.$$



## Euler's Equation



For $\delta I$ to vanish for arbitrary $\delta y(x)$,
$$\boxed{\frac{d}{dx}\frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0.}$$

**Functions involving higher order derivatives**
$$I[y(x)] = \int_{x_1}^{x_2} f\left(x, y, y', y'', \cdots, y^{(n)}\right) dx$$
with prescribed boundary values for $y, y', y'', \cdots, y^{(n-1)}$
$$\delta I = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial y}\delta y + \frac{\partial f}{\partial y'}\delta y' + \frac{\partial f}{\partial y''}\delta y'' + \cdots + \frac{\partial f}{\partial y^{(n)}}\delta y^{(n)}\right] dx$$
**Working rule:** Starting from the last term, integrate one term at a time by parts, using consistency of variations and BC's.
Euler's equation:
$$\frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} + \frac{d^2}{dx^2}\frac{\partial f}{\partial y''} - \cdots + (-1)^n\frac{d^n}{dx^n}\frac{\partial f}{\partial y^{(n)}} = 0,$$
an ODE of order $2n$, in general.



## Euler's Equation



**Functionals of a vector function**
$$I[\mathbf{r}(t)] = \int_{t_1}^{t_2} f(t, \mathbf{r}, \dot{\mathbf{r}})dt$$
In terms of partial gradients $\frac{\partial f}{\partial \mathbf{r}}$ and $\frac{\partial f}{\partial \dot{\mathbf{r}}}$,
$$\begin{aligned}\delta I &= \int_{t_1}^{t_2} \left[\left(\frac{\partial f}{\partial \mathbf{r}}\right)^T \delta\mathbf{r} + \left(\frac{\partial f}{\partial \dot{\mathbf{r}}}\right)^T \delta\dot{\mathbf{r}}\right] dt \\ &= \int_{t_1}^{t_2} \left(\frac{\partial f}{\partial \mathbf{r}}\right)^T \delta\mathbf{r}dt + \left[\left(\frac{\partial f}{\partial \dot{\mathbf{r}}}\right)^T \delta\mathbf{r}\right]_{t_1}^{t_2} - \int_{t_1}^{t_2} \frac{d}{dt}\left(\frac{\partial f}{\partial \dot{\mathbf{r}}}\right)^T \delta\mathbf{r}dt \\ &= \int_{t_1}^{t_2} \left[\frac{\partial f}{\partial \mathbf{r}} - \frac{d}{dt}\frac{\partial f}{\partial \dot{\mathbf{r}}}\right]^T \delta\mathbf{r}dt.\end{aligned}$$
Euler's equation: a system of second order ODE's
$$\frac{d}{dt}\frac{\partial f}{\partial \dot{\mathbf{r}}} - \frac{\partial f}{\partial \mathbf{r}} = \mathbf{0} \quad \text{or} \quad \frac{d}{dt}\frac{\partial f}{\partial \dot{r}_i} - \frac{\partial f}{\partial r_i} = 0 \text{ for each } i.$$

## Euler's Equation



**Functionals of functions of several variables**

$$I[u(x,y)] = \int_D \int f(x,y,u,u_x,u_y)dx\,dy$$

Euler's equation: $\frac{\partial}{\partial x}\frac{\partial f}{\partial u_x} + \frac{\partial}{\partial y}\frac{\partial f}{\partial u_y} - \frac{\partial f}{\partial u} = 0$

**Moving boundaries**

Revision of the basic case: allowing non-zero $\delta y(x_1)$, $\delta y(x_2)$

At an end-point, $\frac{\partial f}{\partial y'}\delta y$ has to vanish for *arbitrary* $\delta y(x)$.

$\frac{\partial f}{\partial y'}$ *vanishes at the boundary.*

| Euler boundary condition or natural boundary condition |
|---|

**Equality constraints and isoperimetric problems**

Minimize $I = \int_{x_1}^{x_2} f(x,y,y')dx$ subject to $J = \int_{x_1}^{x_2} g(x,y,y')dx = J_0$.

In another level of generalization, constraint $\phi(x,y,y') = 0$.

Operate with $f^*(x,y,y',\lambda) = f(x,y,y') + \lambda(x)g(x,y,y')$.



## Direct Methods



**Finite difference method**

With given boundary values $y(a)$ and $y(b)$,

$$I[y(x)] = \int_a^b f[x, y(x), y'(x)]dx$$

- Represent $y(x)$ by its values over $x_i = a + ih$ with $i = 0, 1, 2, \cdots, N$, where $b - a = Nh$.
- Approximate the functional by

$$I[y(x)] \approx \phi(y_1, y_2, y_3, \cdots, y_{N-1}) = \sum_{i=1}^{N} f(\bar{x}_i, \bar{y}_i, \bar{y}'_i)h,$$

  where $\bar{x}_i = \frac{x_i + x_{i-1}}{2}$, $\bar{y}_i = \frac{y_i + y_{i-1}}{2}$ and $\bar{y}'_i = \frac{y_i - y_{i-1}}{h}$.
- Minimize $\phi(y_1, y_2, y_3, \cdots, y_{N-1})$ with respect to $y_i$; for example, by solving $\frac{\partial \phi}{\partial y_i} = 0$ for all $i$.

**Exercise:** Show that $\frac{\partial \phi}{\partial y_i} = 0$ is equivalent to Euler's equation.



## Direct Methods



**Rayleigh-Ritz method**

In terms of a set of basis functions, express the solution as

$$y(x) = \sum_{i=1}^{N} \alpha_i w_i(x).$$

Represent functional $I[y(x)]$ as a multivariate function $\phi(\boldsymbol{\alpha})$.

Optimize $\phi(\boldsymbol{\alpha})$ to determine $\alpha_i$'s.

**Note:** As $N \to \infty$, the numerical solution approaches exactitude. For a particular tolerance, one can truncate appropriately.

**Observation:** With these direct methods, no need to *reduce* the variational (optimization) problem to Euler's equation!

**Question:** Is it possible to reformulate a BVP as a variational problem and then use a direct method?



## Direct Methods



**The inverse problem:** From

$$I[y(x)] \approx \phi(\boldsymbol{\alpha}) = \int_a^b f\left(x, \sum_{i=1}^{N} \alpha_i w_i(x), \sum_{i=1}^{N} \alpha_i w'_i(x)\right) dx,$$

$$\frac{\partial \phi}{\partial \alpha_i} = \int_a^b \left[\frac{\partial f}{\partial y}\left(x, \sum_{i=1}^{N}\alpha_i w_i, \sum_{i=1}^{N}\alpha_i w'_i\right) w_i(x) + \frac{\partial f}{\partial y'}\left(x, \sum_{i=1}^{N}\alpha_i w_i, \sum_{i=1}^{N}\alpha_i w'_i\right) w'_i(x)\right] dx.$$

Integrating the second term by parts and using $w_i(a) = w_i(b) = 0$,

$$\frac{\partial \phi}{\partial \alpha_i} = \int_a^b \mathcal{R}\left[\sum_{i=1}^{N} \alpha_i w_i\right] w_i(x)dx,$$

where $\mathcal{R}[y] \equiv \frac{\partial f}{\partial y} - \frac{d}{dx}\frac{\partial f}{\partial y'} = 0$ is the Euler's equation of the variational problem.

Def.: $\mathcal{R}[z(x)]$: *residual* of the differential equation $\mathcal{R}[y] = 0$ operated over the function $z(x)$

*Residual of the Euler's equation of a variational problem operated upon the solution obtained by Rayleigh-Ritz method is orthogonal to basis functions $w_i(x)$.*

# Direct Methods



**Galerkin method**

**Question:** What if we cannot find a 'corresponding' variational problem for the differential equation?

**Answer:** Work with the residual directly and demand

$$\int_a^b \mathcal{R}[z(x)]w_i(x)dx = 0.$$

Freedom to choose two *different* families of functions as basis functions $\psi_j(x)$ and trial functions $w_i(x)$:

$$\int_a^b \mathcal{R}\left[\sum_j \alpha_j \psi_j(x)\right] w_i(x)dx = 0$$

A singular case of the Galerkin method:

*delta functions, at discrete points, as trial functions*

Satisfaction of the differential equation *exactly* at the chosen points, known as collocation points:

Collocation method



# Direct Methods



**Finite element methods**

- discretization of the domain into elements of simple geometry
- basis functions of low order polynomials with local scope
- design of basis functions so as to achieve enough order of continuity or smoothness across element boundaries
- piecewise continuous/smooth basis functions for entire domain, with a built-in sparse structure
- some weighted residual method to frame the algebraic equations
- solution gives coefficients which are actually the nodal values

Suitability of finite element analysis in software environments

- effectiveness and efficiency
- neatness and modularity



# Points to note



- Optimization with respect to a *function*
- Concept of a functional
- Euler's equation
- Rayleigh-Ritz and Galerkin methods
- Optimization and equation-solving in the infinite-dimensional function space: practical methods and connections

Necessary Exercises: **1,2,4,5**



# Outline

## Epilogue

Source for further information:
*http://home.iitk.ac.in/~dasgupta/MathBook*

Destination for feedback:
*dasgupta@iitk.ac.in*

Some general courses in immediate continuation

- Advanced Mathematical Methods
- Scientific Computing
- Advanced Numerical Analysis
- Optimization
- Advanced Differential Equations
- Partial Differential Equations
- Finite Element Methods



## Epilogue

Some specialized courses in immediate continuation

- Linear Algebra and Matrix Theory
- Approximation Theory
- Variational Calculus and Optimal Control
- Advanced Mathematical Physics
- Geometric Modelling
- Computational Geometry
- Computer Graphics
- Signal Processing
- Image Processing



## Outline





## Selected References I


F. S. Acton.
*Numerical Methods that usually Work*.
The Mathematical Association of America (1990).

C. M. Bender and S. A. Orszag.
*Advanced Mathematical Methods for Scientists and Engineers*.
Springer-Verlag (1999).

G. Birkhoff and G.-C. Rota.
*Ordinary Differential Equations*.
John Wiley and Sons (1989).

G. H. Golub and C. F. Van Loan.
*Matrix Computations*.
The John Hopkins University Press (1983).

## Selected References II


- M. T. Heath.
  *Scientific Computing*.
  Tata McGraw-Hill Co. Ltd (2000).
- E. Kreyszig.
  *Advanced Engineering Mathematics*.
  John Wiley and Sons (2002).
- E. V. Krishnamurthy and S. K. Sen.
  *Numerical Algorithms*.
  Affiliated East-West Press Pvt Ltd (1986).
- D. G. Luenberger.
  *Linear and Nonlinear Programming*.
  Addison-Wesley (1984).
- P. V. O'Neil.
  *Advanced Engineering Mathematics*.
  Thomson Books (2004).




## Selected References III


- W. H. Press, S. A. Teukolsky, W. T. Vellerling and B. P. Flannery.
  *Numerical Recipes*.
  Cambridge University Press (1998).
- G. F. Simmons.
  *Differential Equations with Applications and Historical Notes*.
  Tata McGraw-Hill Co. Ltd (1991).
- J. Stoer and R. Bulirsch.
  *Introduction to Numerical Analysis*.
  Springer-Verlag (1993).
- C. R. Wylie and L. C. Barrett.
  *Advanced Engineering Mathematics*.
  Tata McGraw-Hill Co. Ltd (2003).