



Mathematical Foundations for Data Science

BITS Pilani
Pilani Campus

MFDS Team



DSECL ZC416, MFDS

Lecture 0

Agenda

- Matrices and their types
- REF and RREF
- Rank, its computation and properties
- Determinant, its computation and properties
- Consistency and inconsistency of linear systems
- Significant digits and rounding arithmetic

Matrices

- A **matrix** is a **rectangular array of numbers or functions** which we will enclose in brackets. For example,

$$\begin{bmatrix} 0.3 & 1 & -5 \\ 0 & -0.2 & 16 \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad (1)$$

$$\begin{bmatrix} e^{-x} & 2x^2 \\ e^{6x} & 4x \end{bmatrix}, \quad \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}, \quad \begin{bmatrix} 4 \\ 1 \\ \frac{1}{2} \end{bmatrix}$$

- The numbers (or functions) are called **entries** or, less commonly, *elements* of the matrix.
- The first matrix in (1) has two **rows**, which are the horizontal lines of entries.

Matrix – Notations

- We shall denote matrices by capital boldface letters \mathbf{A} , \mathbf{B} , \mathbf{C} , ..., or by writing the general entry in brackets; thus $\mathbf{A} = [a_{jk}]$, and so on.
- By an $m \times n$ **matrix** (read *m by n matrix*) we mean a **matrix with m rows and n columns**—rows always come first! $m \times n$ is called the **size** of the matrix. Thus an $m \times n$ matrix is of the form

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}. \quad (2)$$

Vectors

- A **vector** is a **matrix with only one row or column**. Its entries are called the **components** of the vector.
- We shall denote vectors by *lowercase boldface letters* **a**, **b**, ... or by its general component in brackets, $\mathbf{a} = [a_j]$, and so on. Our special vectors in (1) suggest that a (general) **row vector** is of the form

$$\mathbf{a} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}. \quad \text{For instance, } \mathbf{a} = \begin{bmatrix} -2 & 5 & 0.8 & 0 & 1 \end{bmatrix}.$$

A **column vector**

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \quad \text{For instance, } \mathbf{b} = \begin{bmatrix} 4 \\ 0 \\ -7 \end{bmatrix}.$$

Equality of Matrices

- Two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ are **equal**, written $\mathbf{A} = \mathbf{B}$, if and only if (1) they have the same size and (2) the corresponding entries are equal, that is, $a_{11} = b_{11}$, $a_{12} = b_{12}$, and so on.
- Matrices that are not equal are called **different**. Thus, matrices of different sizes are always different.

Matrix Multiplication

Multiplication of a Matrix by a Matrix

- The **product** $\mathbf{C} = \mathbf{AB}$ (in this order) of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ times an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is **defined if and only if $r = n$** and is then the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$(3) \quad c_{jk} = \sum_{l=1}^n a_{jl} b_{lk} = a_{j1} b_{1k} + a_{j2} b_{2k} + \cdots + a_{jn} b_{nk} \quad \begin{matrix} j = 1, \dots, m \\ k = 1, \dots, p. \end{matrix}$$

- The condition $r = n$ means that the second factor, \mathbf{B} , must have as many rows as the first factor has columns, namely n . A diagram of sizes that shows when matrix multiplication is possible is as follows:

$$\begin{array}{ccc} \mathbf{A} & \mathbf{B} & = \mathbf{C} \\ [m \times n] & [n \times p] & = [m \times p]. \end{array}$$

Matrix Multiplication

Matrix Multiplication Is *Not Commutative*, $AB \neq BA$ in General

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

but $\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} = \begin{bmatrix} 99 & 99 \\ -99 & -99 \end{bmatrix}.$

- It is interesting that this also shows that $AB = 0$ does *not* necessarily imply $BA = 0$ or $A = 0$ or $B = 0$.

Transposition of Matrices & Vectors



- The transpose of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the $n \times m$ matrix \mathbf{A}^T (read *A transpose*) that has the first *row* of \mathbf{A} as its first *column*, the second *row* of \mathbf{A} as its second *column*, and so on.
 - As a special case, transposition converts row vectors to column vectors and conversely.
 - Rules of transposition
 - (a) $(\mathbf{A}^T)^T = \mathbf{A}$
 - (b) $(\mathbf{A} + \mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$
 - (c) $(c\mathbf{A})^T = c\mathbf{A}^T$
 - (d) $(\mathbf{AB})^T = \mathbf{B}^T \mathbf{A}^T$.
-

Special Matrices

- **Symmetric:** $a_{ij} = a_{ji}$ Eg:
$$\begin{bmatrix} 1 & 1 & -1 \\ 1 & 2 & 0 \\ -1 & 0 & 5 \end{bmatrix}$$
- **Skew Symmetric :** $a_{ij} = - a_{ji}$ Eg:
$$\begin{bmatrix} 0 & 1 & -2 \\ -1 & 0 & 3 \\ 2 & -3 & 0 \end{bmatrix}$$
- **Triangular:** Upper Triangular $\rightarrow a_{ij} = 0$ for all $i > j$
 Lower Triangular $\rightarrow a_{ij} = 0$ for all $i < j$

Upper triangular matrix: U

$$\begin{bmatrix} 1 & 1/2 & 3 & 0 \\ 0 & 5 & 0 & 1 \\ 0 & 0 & 4 & -2 \\ 0 & 0 & 0 & 3 \end{bmatrix}$$

 Lower triangular matrix: L

$$\begin{bmatrix} 1 & 0 & 0 \\ 3 & 3 & 0 \\ 1 & -2 & 0 \end{bmatrix}$$
- **Diagonal Matrix:** $a_{ij} = 0$ for all $i \neq j$ Eg:

$$\begin{bmatrix} \text{---} & 0 & \dots & 0 \\ 0 & \text{---} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \text{---} \end{bmatrix}$$

$$\begin{bmatrix} 0 & 0 & 3 & 0 & 4 \\ 0 & 0 & 5 & 7 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 2 & 6 & 0 & 0 \end{bmatrix}$$
- **Sparse Matrix:** Many zeroes and few non-zero entities
- **Orthogonal matrix:** A such that $A^T = A^{-1}$

Elementary Row Operations

- *Interchange of two rows*
- *Addition of a constant multiple of one row to another row*
- *Multiplication of a row by a nonzero constant c*

CAUTION! These operations are for rows, *not for columns!*

Row Echelon Form (REF) of a matrix



- Any rows of all zeros are below any other non zero rows.
- Each leading entry of a row is in a column to the right of the leading entry of the row above it.
- All entries in a column below a leading entry are zeros
- [Example](#)

$$\begin{bmatrix} 3 & 2 & 0 & 7 & 9 \\ 0 & 4 & 5 & 10 & 0 \\ 0 & 0 & 0 & -4 & 1 \\ 0 & 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Row Reduced Echelon Form (RREF)



- We say that a matrix is in Reduced Row Echelon Form if it is in Echelon form and additionally,
 1. The leading entry in each row is 1.
 2. Each leading 1 is the only non zero entry in its column

Example

$$\begin{bmatrix} 1 & 0 & 3 & 0 & 9 \\ 0 & 1 & 4 & 0 & -6 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Uniqueness of Row Reduced Echelon Form



- We can transform any matrix into a matrix in reduced echelon form by using elementary row operations.
- No matter what sequence of row operations we use each matrix is row equivalent to one and only one reduced echelon matrix

Rank of a Matrix

Definition

- The **rank** of a matrix A is the number of non-zero row in the RREF of A .
- It is denoted by $\text{rank } A$.
- We call a matrix A_1 **row-equivalent** to a matrix A_2 if A_1 can be obtained from A_2 by (finitely many!) elementary row operations.

Theorem 1

- *Row-equivalent matrices have the same rank.*

Determination of Rank

$$A = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \quad (\text{given})$$

$$= \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & -21 & -14 & -29 \end{bmatrix} \quad \begin{array}{l} \text{Row 2} + 2 \text{ Row 1} \\ \text{Row 3} - 7 \text{ Row 1} \end{array}$$

$$= \begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \text{Row 3} + \frac{1}{2} \text{ Row 2.}$$

- The last matrix is in row-echelon form and has two nonzero rows.
Hence rank $A = 2$.

Minor

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Each element in \mathbf{A} has a minor

Delete first row and column from \mathbf{A} .

The determinant of the remaining 2×2 submatrix is the minor of a_{11}

$$m_{11} = \begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix}$$

Minor

$$m_{12} = \begin{vmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{vmatrix}$$

And the minor for a_{13} is:

$$m_{13} = \begin{vmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{vmatrix}$$

Cofactor

The cofactor C_{ij} of an element a_{ij} is defined as:

$$C_{ij} = (-1)^{i+j} m_{ij}$$

When the sum of a row number i and column j is even, $c_{ij} = m_{ij}$ and when $i+j$ is odd, $c_{ij} = -m_{ij}$

$$c_{11}(i=1, j=1) = (-1)^{1+1} m_{11} = +m_{11}$$

$$c_{12}(i=1, j=2) = (-1)^{1+2} m_{12} = -m_{12}$$

$$c_{13}(i=1, j=3) = (-1)^{1+3} m_{13} = +m_{13}$$

Determinant

The determinant of an $n \times n$ matrix \mathbf{A} can now be defined as

$$|A| = \det A = a_{11}c_{11} + a_{12}c_{12} + \dots + a_{1n}c_{1n}$$

The determinant of \mathbf{A} is therefore the sum of the products of the elements of the first row of \mathbf{A} and their corresponding cofactors.

(It is possible to define $|A|$ in terms of any other row or column but for simplicity, the first row only is used)

Adjoint

The adjoint matrix of \mathbf{A} , denoted by $\text{adj } \mathbf{A}$, is the transpose of its cofactor matrix

$$\text{adj} \mathbf{A} = \mathbf{C}^T$$

It can be shown that:

$$\mathbf{A}(\text{adj } \mathbf{A}) = (\text{adj} \mathbf{A}) \mathbf{A} = |\mathbf{A}| \mathbf{I}$$

Example: $A = \begin{bmatrix} 1 & 2 \\ -3 & 4 \end{bmatrix}$

$$|A| = (1)(4) - (2)(-3) = 10$$

$$\text{adj} \mathbf{A} = \mathbf{C}^T = \begin{bmatrix} 4 & -2 \\ 3 & 1 \end{bmatrix}$$

Adjoint

$$A(adjA) = \begin{bmatrix} 1 & 2 \\ -3 & 4 \end{bmatrix} \begin{bmatrix} 4 & -2 \\ 3 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = 10I$$

$$(adjA)A = \begin{bmatrix} 4 & -2 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ -3 & 4 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} = 10I$$

Properties of Determinants

1. $\det(AB) = \det(A) * \det(B)$
2. $\det(A)$ nonzero implies there exists a matrix B such that $AB=BA=I$
3. Two Rows Equal $\rightarrow \det = 0$ (Singular)
4. R_i and R_j swapped $\rightarrow \det$ gets a minus sign ($i \neq j$)
5. $\det(A) = \det(A^T)$
6. $R_i \leftarrow cR_j \rightarrow \det A \leftarrow c \det A$

Orthogonal matrices: $\det(A) = -1$ or 1
$$\begin{bmatrix} \cos q & -\sin q \\ \sin q & \cos q \end{bmatrix}$$

Inverse

- $A^{-1} = \frac{\text{adj}(A)}{\det(A)}$ where $\det(A) \neq 0$

Reiterate $\det(A) \neq 0 \rightarrow A$ is Non singular

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

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

Inverse

The result can be checked using

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

The determinant of a matrix must not be zero for the inverse to exist as there will not be a solution

Non-singular matrices have non-zero determinants

Singular matrices have zero determinants

Linear System

- A **linear system of m equations in n unknowns** x_1, \dots, x_n is a set of equations of the form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \dots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{1}$$

- The system is called *linear* because each variable x_j appears in the first power only, just as in the equation of a straight line.

Linear System

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \dots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned} \tag{1}$$

- a_{11}, \dots, a_{mn} are given numbers, called the **coefficients** of the system.
- b_1, \dots, b_m on the right are also given numbers.
- If all the b_j are zero, then (1) is called a **homogeneous system**.
- If at least one b_j is not zero, then (1) is called a **non-homogeneous system**.

Coefficient Matrix

Matrix Form of the Linear System (1).

- From the definition of matrix multiplication we see that the m equations of (1) may be written as a single vector equation

$$\mathbf{Ax} = \mathbf{b} \quad (2)$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $m \times n$ matrix are column vectors.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

Matrix Form of Linear System

Matrix Form of the Linear System (1). (continued)

- We assume that the coefficients a_{jk} are not all zero, so that \mathbf{A} is not a zero matrix. Note that \mathbf{x} has n components, whereas \mathbf{b} has m components. The matrix

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array} \right]$$

- is called the **augmented matrix** of the system (1).
- The dashed vertical line could be omitted, as we shall do later. It is merely a reminder that the last column of $\tilde{\mathbf{A}}$ did not come from matrix \mathbf{A} but came from vector \mathbf{b} . Thus, we *augmented* the matrix \mathbf{A} .

Solution to System of Linear Equations



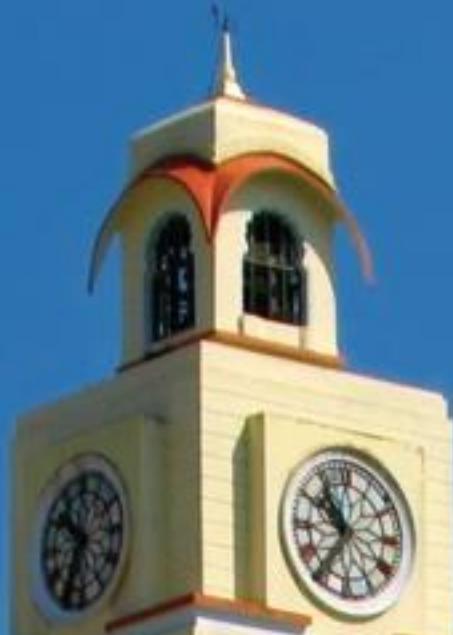
- A linear system (1) is called **overdetermined** if it has more equations than unknowns, **determined** if $m = n$, and **underdetermined** if it has fewer equations than unknowns.
- A linear system is **consistent** if $\text{rank}(A) = \text{rank}(A|b)$
- A **consistent** system has at least one solution (thus, one solution or infinitely many solutions), but **inconsistent** has no solutions at all, as $x_1 + x_2 = 1, x_1 + x_2 = 0$.

Significant Digits

In a **floating-point** system, the **significant digits** of a number c is any given digit of c , except possibly for zeros to the left of the first nonzero digit; these zeros serve only to fix the position of the decimal point. For instance, **13600**, **1.3600**, **0.0013600**, all have 5 significant digits.

Round the number 1.23454621 to (a) 2 S, (b) 3 S, (c) 4 S, (d) 5 S and (e) 6 S

- a) 1.2 b) 1.23 c) 1.235 d) 1.2345 e) 1.23455



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No.1

Agenda

- Solution of linear systems – an overview
 - Gauss elimination methods
 - sensitivity to changes in A
 - pivoting and
 - operations count
 - LU decomposition methods
 - Doolittle's method
 - Crout's method
-

Linear System

Matrix Form of the Linear System (1).

From the definition of matrix multiplication we see that the m equations of (1) may be written as a single vector equation

$$(2) \quad \mathbf{Ax} = \mathbf{b}$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $m \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

are column vectors.

Matrix Form of Linear System

Matrix Form of the Linear System (1). (continued)

We assume that the coefficients a_{jk} are not all zero, so that \mathbf{A} is not a zero matrix. Note that \mathbf{x} has n components, whereas \mathbf{b} has m components. The matrix

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array} \right]$$

is called the **augmented matrix** of the system (1). The dashed vertical line could be omitted, as we shall do later. It is merely a reminder that the last column of $\tilde{\mathbf{A}}$ did not come from matrix \mathbf{A} but came from vector \mathbf{b} . Thus, we *augmented* the matrix \mathbf{A} .

Gauss Elimination and Back Substitution



Triangular form:

Triangular means that all the nonzero entries of the corresponding coefficient matrix lie above the diagonal and form an upside-down 90° triangle. Then we can solve the system by **back substitution**.

Since a linear system is completely determined by its augmented matrix, *Gauss elimination can be done by merely considering the matrices.*

(We do this again in the next example, emphasizing the matrices by writing them first and the equations behind them, just as a help in order not to lose track.)

Gauss Elimination

At the end of the Gauss elimination (before the back substitution), the row echelon form of the augmented matrix will be

- in upper triangular form
- having the first r rows non-zero
- Exactly $(m - r)$ rows would be zero rows
- the rhs would also have the last $(m-r)$ rows zero
- any one of the $(m-r)$ last rows in non-zero would imply inconsistency
- complexity is $O(n^3)$, where n is the number of rows
- facilitates the back substitution

$$\left[\begin{array}{cccc|c} F_{11} & F_{12} & \cdots & F_{1n} & f_1 \\ 0 & F_{22} & \cdots & F_{2n} & f_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & F_{rr} & f_r \\ 0 & 0 & \cdots & 0 & f_{r+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & f_m \end{array} \right].$$

Solution

The number of nonzero rows, r , in the row-reduced coefficient matrix \mathbf{R} is called the **rank of \mathbf{R}** and also the **rank of \mathbf{A}** . Here is the method for determining whether $\mathbf{Ax} = \mathbf{b}$ has solutions and what they are:

(a) No solution. If r is less than m (meaning that \mathbf{R} actually has at least one row of all 0s) *and* at least one of the numbers $f_{r+1}, f_{r+2}, \dots, f_m$ is not zero, then the system $\mathbf{Rx} = \mathbf{f}$ is inconsistent: No solution is possible. Therefore the system $\mathbf{Ax} = \mathbf{b}$ is inconsistent as well.

Solution

If the system is consistent (either $r = m$, or $r < m$ and all the numbers $f_{r+1}, f_{r+2}, \dots, f_m$ are zero), then there are solutions.

- (b) Unique solution.** If the system is consistent and $r = n$, there is exactly one solution, which can be found by back substitution.
- (c) Infinitely many solutions.** To obtain any of these solutions, choose values of x_{r+1}, \dots, x_n arbitrarily. Then solve the r th equation for x_r (in terms of those arbitrary values), then the $(r - 1)$ st equation for x_{r-1} , and so on up the line.

Gauss Elimination

Solve $Ax = b$

Consists of two phases:
Forward elimination
Back substitution

Forward Elimination
 reduces $Ax = b$ to an
 upper triangular system
 $Tx = b'$

Back substitution can then
 solve $Tx = b'$ for x

$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{array} \right]$$



$$\left[\begin{array}{ccc|c} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & a_{22} & a_{23} & b_2' \\ 0 & 0 & a_{33}'' & b_3'' \end{array} \right]$$



$$x_3 = \frac{b_3''}{a_{33}''} \quad x_2 = \frac{b_2' - a_{23}'x_3}{a_{22}'}$$

$$x_1 = \frac{b_1 - a_{13}x_3 - a_{12}x_2}{a_{11}}$$

Forward
Elimination

Back
Substitution

Pitfalls of Gauss Elimination

Division by zero

It is possible that during both elimination and back-substitution phases a division by zero can occur.

For example:

$$\begin{aligned} 2x_2 + 3x_3 &= 8 \\ 4x_1 + 6x_2 + 7x_3 &= -3 \\ 2x_1 + x_2 + 6x_3 &= 5 \end{aligned}$$

A =

0	2	3
4	6	7
2	1	6

**a₁₁ = 0
(the pivot element)**

It is possible that during both elimination and back-substitution phases, a division by zero can occur.

Solution: Pivoting

Pitfalls of Gauss Elimination

Round-off errors

Because computers carry only a limited number of significant figures, round-off errors will occur and they will *propagate* from one iteration to the next.

This problem is especially important when **large** numbers of equations (100 or more) are to be solved.

Always use **double-precision** numbers/arithmetic. It is slow but needed for correctness!

It is also a good idea to substitute your results back into the original equations and check whether a substantial error has occurred.

Common norms

- The most commonly used vector norms belong to the family of ℓ_p -norms, which are defined by

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

- The following ℓ_p -norms are of particular interest:

- $p = 1$: The ℓ_1 -norm

$$\|x\|_1 = |x_1| + |x_2| + \dots + |x_n|$$

- $p = 2$: The ℓ_2 -norm or Euclidean norm

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2} = \sqrt{x^T x}$$

- $p = \infty$: The ℓ_∞ -norm

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

Matrix Norms

Some of the commonly used matrix norms are given below

Matrix norm corresponding to vector 1-norm is maximum absolute column sum

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

Matrix norm corresponding to vector ∞ - norm is maximum absolute row sum,

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

$\|A\|_2$ is the *Frobenius norm*

$$\|A\|_2 = \|A\|_F = \sqrt{\sum_{i=1}^M \sum_{j=1}^N |a_{ij}|^2}$$

Ill conditioned systems

- Systems where small changes in coefficients result in large change in solution

$$x_1 + 2x_2 = 10$$

$$1.1x_1 + 2x_2 = 10.4$$

$$\rightarrow x_1 = 4.0 \text{ & } x_2 = 3.0$$

$$x_1 + 2x_2 = 10$$

$$1.05x_1 + 2x_2 = 10.4$$

$$\rightarrow x_1 = 8.0 \text{ & } x_2 = 1.0$$

Condition Number

Condition number of a non singular matrix A is defined by

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

By convention, $\text{cond}(A) = \infty$ if A is singular

Example: $A = \begin{bmatrix} 2 & -1 & 1 \\ 1 & 0 & 1 \\ 3 & -1 & 4 \end{bmatrix}$ $\|A\|_1 = 6$ $\|A\|_\infty = 8$

$$A^{-1} = \begin{bmatrix} 0.5 & 1.5 & -0.5 \\ -0.5 & 2.5 & -0.5 \\ -0.5 & -0.5 & 0.5 \end{bmatrix}$$
 $\|A^{-1}\|_1 = 4.5$ $\|A^{-1}\|_\infty = 3.5$

$$\text{cond}_1(A) = 6 \times 4.5 = 27$$

$$\text{cond}_\infty(A) = 8 \times 3.5 = 28$$

1. There is no sharp dividing line between “well-conditioned” and “ill-conditioned,” but generally the situation will get worse as we go from systems with small $\kappa(A)$ to systems with larger $\kappa(A)$. Now always $\kappa(A) \geq 1$, so that values of 10 or 20 or so give no reason for concern, whereas $\kappa(A) = 100$, say, calls for caution, and systems such as those in Examples 1 and 2 are extremely ill-conditioned.

2. If $\kappa(A)$ is large (or small) in one norm, it will be large (or small, respectively) in any other norm. See Example 5.

Techniques for Improving the solution



Use of more significant figures – double precision arithmetic

Pivoting

If a pivot element is zero, normalization step leads to division by zero. The same problem may arise, when the pivot element is close to zero. Problem can be avoided:

Partial pivoting

- Switching the rows below so that the largest element is the pivot element.

Complete pivoting

- Searching for the largest element in all rows and columns then switching.
- This is rarely used because switching columns changes the order of x' s and adds significant complexity and overhead → costly

Scaling - used to reduce the round-off errors and improve accuracy

Partial Pivoting – Example

Pivoting Example

Example 14: Solve the following system using Gauss Elimination with pivoting.

$$\begin{aligned}
 2x_2 + & \quad + x_4 = 0 \\
 2x_1 + 2x_2 + 3x_3 + 2x_4 & = -2 \\
 4x_1 - 3x_2 & \quad + x_4 = -7 \\
 6x_1 + x_2 - 6x_3 - 5x_4 & = 6
 \end{aligned}$$

Step 0: Form the augmented matrix

0	2	0	1		0
2	2	3	2		-2
4	-3	0	1		-7
6	1	-6	-5		6

Step 1: Forward Elimination

(1.1) Eliminate x_1 . But the pivot element is 0. We have to interchange the 1st row with one of the rows below it. Interchange it with the 4th row because 6 is the largest possible pivot.

Partial Pivoting – Example

(1.1) Eliminate x_1 . But the pivot element is 0. We have to interchange the 1st row with one of the rows below it. Interchange it with the 4th row because 6 is the largest possible pivot.

6	1	-6	-5		6
2	2	3	2		-2
4	-3	0	1		-7
0	2	0	1		0

Now eliminate x_1

6	1	-6	-5		6
0	1.6667	5	3.6667		-4
0	-3.6667	4	4.3333		-11
0	2	0	1		0

(1.2) Eliminate x_2 from the 3rd and 4th eqns. Pivot element is 1.6667. There is no division by zero problem. Still we will perform pivoting to reduce round-off errors. Interchange the 2nd and 3rd rows. Note that complete pivoting would interchange 2nd and 3rd columns.

6	1	-6	-5		6
0	-3.6667	4	4.3333		-11
0	1.6667	5	3.6667		-4
0	2	0	1		0

Eliminate x_2

6	1	-6	-5		6
0	-3.6667	4	4.3333		-11
0	0	6.8182	5.6364		-9.0001
0	0	2.1818	3.3636		-5.9999

(1.3) Eliminate x_3 . $6.8182 > 2.1818$, therefore no pivoting is necessary.

6	1	-6	-5		6
0	-3.6667	4	4.3333		-11
0	0	6.8182	5.6364		-9.0001
0	0	0	1.5600		-3.1199

Partial Pivoting – Example

Step 2: Back substitution

$$x_4 = -3.1199 / 1.5600 = \mathbf{-1.9999}$$

$$x_3 = [-9.0001 - 5.6364 * (-1.9999)] / 6.8182 = \mathbf{0.33325}$$

$$x_2 = [-11 - 4.3333 * (-1.9999) - 4 * 0.33325] / -3.6667 = \mathbf{1.0000}$$

$$x_1 = [6 - (-5) * (-1.9999) - (-6) * 0.33325 - 1 * 1.0000] / 6 = \mathbf{-0.50000}$$

Exact solution is $x = [-2 \quad 1/3 \quad 1 \quad -0.5]^T$. Use more than 5 sig. figs. to reduce round-off errors.

Gauss Elimination with Rounding

$$0.0004x_1 + 1.402x_2 = 1.406$$

$$0.4003x_1 - 1.502x_2 = 2.501$$

Original solution of the system is $x_1 = 10$, $x_2 = 1$

Picking the first of given equation as pivot equation , we have to multiply this equation by $m = 0.4003/0.0004 = 1001$ and subtract result from the second equation , obtaining

$$-1405x_2 = -1404 \rightarrow x_2 = 0.9993$$

From first equation we get $x_1 = 12.5$

The failure occurs because $|a_{11}|$ is small compared to $|a_{12}|$ so that a small round off error in x_2 led to a large error in x_1

Gauss Elimination with Rounding

$$0.0004x_1 + 1.402x_2 = 1.406$$

$$0.4003x_1 - 1.502x_2 = 2.501$$

Picking the second of the given equations as the pivot equation, we have to multiply this equation by $0.0004/0.4003 = 0.0009993$ and subtract the result from the first equation obtaining

$$1.404x_2 = 1.404$$

$$x_2 = 1 \text{ and } x_1 = 10$$

Note $|a_{21}|$ is not very small compared to $|a_{22}|$ so that a small round off error in x_2 would not lead to a large error in x_1

Operation Count – Gauss Elimination



Important factors in judging the quality of a numerical method are

- Amount of storage
- Amount of time (= number of operations)

Consider Augmented Matrix of $Ax = b$, where $a_{in+1} = b_i$

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n+1} \\ a_{21} & a_{22} & \dots & a_{2n+1} \\ \vdots & \ddots & & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn+1} \end{bmatrix}$$

Operation Count – Gauss Elimination



In elimination procedure to get the rank we will make all elements below main diagonal zero.

Total number of multiplications and additions required to determine the rank by elimination procedure are

$$2 \cdot \sum_{k=1}^{n-1} (n-k)(n-k+1) = O(n^3)$$

Total number of divisions is

$$\sum_{k=1}^{n-1} (n-k) = O(n^2)$$

Operation Count – Gauss Elimination



In back substitution total number of additions, multiplications and divisions required are

$$\left(2 \cdot \sum_{k=1}^n (n - k) \right) + n = O(n^2)$$

If an operation takes 10^{-9} sec, then

Algorithm	$n = 1000$	$n = 10000$
Elimination	0.7 sec	11 min
Back substitution	0.001 sec	0.1 sec

LU Factorization

We write square matrix A as

$$A = LU$$

Doolittle's Method : L is lower triangular matrix
 $\text{diag}(L) = 1$, $l_{ii} = 1$ and U is
upper triangular matrix

Crout's Method : U is upper triangular matrix with
 $\text{diag}(U) = 1$, $u_{ii} = 1$ and L is lower
triangular matrix

Cholesky's Method: $U = L^T$

Benefits of LU Decomposition

$A = LU$, Thus, the system $Ax = b$, is

$$LUx = b$$

Let $Ux = y$, then

$$Ly = b$$

Algorithm :-

Step-I Solve $Ly = b$, to find y .

Step-II Then solve $Ux = y$ to find x

•Methods of LU Factorization

Doolittle Method: The Factors L, U are defined as

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix} \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}$$

$l_{ij} = 1$, for $i = j$
 $l_{ij} = 0$, for $i < j$
 $u_{ij} = 0$, for $i > j$

Crout's Method: The Factors L, U are defined as

$$L = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \quad U = \begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$l_{ij} = 0$, for $i < j$
 $u_{ij} = 1$, for $i = j$
 $u_{ij} = 0$, for $i > j$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 2

Agenda

- LU Decomposition (continued)
 - Motivation for numerical methods
 - Matrix Norms
 - Iterative Solution
 - Gauss Seidel Method
 - Gauss Jacobi Method
 - Convergence Criteria
-

Crout's Method

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = LU$$

where $L = \begin{bmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{bmatrix}$ $U = \begin{bmatrix} 1 & U_{12} & U_{13} \\ 0 & 1 & U_{23} \\ 0 & 0 & 1 \end{bmatrix}$

.

Crout's Method

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 3 & 8 & 14 \\ 2 & 6 & 13 \end{bmatrix} = LU$$

where $L = \begin{bmatrix} 1 & 0 & 0 \\ 3 & 2 & 0 \\ 2 & 2 & 3 \end{bmatrix}$ $U = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$

Cholesky Method

- Cholesky decomposition is a technique that is designed for a system where the matrix \mathbf{A} is symmetric and positive definite.
- The symmetric matrix refers to the matrix with the element of $a_{ij} = a_{ji}$ for all $i \neq j$. In other words, $\mathbf{A} = \mathbf{A}^T$.
- Cholesky decomposition method offer computational advantages because only half of the storage and computation time are required.
- In Cholesky method, a symmetric matrix \mathbf{A} is decomposed as

$$\mathbf{A} = \mathbf{U}^T \mathbf{U}$$

Cholesky Method

- Decompose A such that $A = U^T U$. Hence, we may have $U^T U x = b$
- Set up and solve $U^T d = b$, where d can be obtained by using forward substitution
- Set up and solve $Ux = d$, where x can be obtained by using backward substitution

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2}$$
$$u_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj}}{u_{ii}} \quad \text{for } j = i+1, \dots, n$$

Computational Complexity

The LU decomposition is computed directly without solving simultaneous equations

- It is more economical to produce the LU Factorization
- This is followed by solving two simpler linear systems

1. To perform LU Factorization , we need about $\frac{n^3}{3}$ operations
2. To solve the Lower triangular system $Ly=b$ we need $O(n^2)$ operations
3. To solve the Upper triangular system $Ux=y$ we need $O(n^2)$ operations

Motivation

Problem: Solve $7x = x + 18$. (**Solution $x = 3$**)

Iterative Procedure: $x_{i+1} = f(x_i)$, with x_1 given

$$7x_{i+1} = x_i + 18, x_1 = 1$$

$$x_2 = 2.7143$$

$$x_3 = 2.9592$$

$$x_4 = 2.9942$$

$$x_5 = 2.9992$$

$$x_6 = 2.9999$$

Converges to $x = 3$

$$x_{i+1} = 7x_i - 18, x_1 = 1$$

$$x_2 = -11$$

$$x_3 = -95$$

$$x_4 = -683$$

$$x_5 = -4799$$

$$x_6 = -33611$$

Diverges

- a) Extension to Linear Systems?
- b) Criteria for convergence (Similar to $\|r\| < 1$)

Matrix Norms

Some of the commonly used matrix norms are given below

Matrix norm corresponding to vector 1-norm is maximum absolute column sum

$$\|A\|_1 = \max_j \sum_{i=1}^n |a_{ij}|$$

Matrix norm corresponding to vector ∞ - norm is maximum absolute row sum,

$$\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$$

$\|A\|_2$ is the *Frobenius norm*

$$\|A\|_2 = \|A\|_F = \sqrt{\sum_{i=1}^M \sum_{j=1}^N |a_{ij}|^2}$$

Iterative Methods for Systems

Consider the normalized system $Ax = b$

Write $A = I + L + U$, where

- I is an identity matrix
- L is lower triangular matrix with diagonal entries 0
- U is upper triangular matrix with diagonal entries 0

$$Ax = (I + L + U)x = b$$

- $x^{k+1} = -Lx^k - Ux^k + b$ (Gauss Jacobi)
- $x^{k+1} = -(I + L)^{-1}Ux^k + (I + L)^{-1}b$ (Gauss Seidel)

Converges if $\|(I + L)^{-1}U\| < 1$

Gauss Seidel Method

- Given the system $[A]\{x\}=\{B\}$ and starting values $\{x\}^0$, Gauss-Seidel uses the first equation to solve for x_1 , second for x_2 , etc.

$$x_1 = (b_1 - a_{12}x_2 - a_{13}x_3 - \dots - a_{1n}x_n) / a_{11}$$

$$x_2 = (b_2 - a_{21}x_1 - a_{23}x_3 - \dots - a_{2n}x_n) / a_{22}$$

.....

$$x_n = (b_n - a_{n1}x_1 - a_{n2}x_2 - \dots - a_{(n-1)(n-1)}x_{n-1}) / a_{nn}$$

After the first iteration you get $\{x\}^1$. Use these values to start a new iteration. Repeat until the tolerance is satisfied as

$$|\varepsilon_{a,i}| = \left| \frac{x_i^k - x_i^{k-1}}{x_i^k} \right| \leq 100\% < \varepsilon_s$$

for all the unknowns ($i=1,\dots,n$), where k and $k-1$ represent the present and previous iterations.

Gauss Seidel Method

Solve the following system using the Gauss-Seidel Method

$$\begin{aligned}
 6x_1 - 2x_2 + x_3 &= 11 \\
 -2x_1 + 7x_2 + 2x_3 &= 5 \\
 x_1 + 2x_2 - 5x_3 &= -1
 \end{aligned}
 \quad \text{starting with } x_1^0 = x_2^0 = x_3^0 = 0.0$$

- Rearrange the equations

$$\begin{aligned}
 x_1 &= (11 + 2x_2 - x_3) / 6 \\
 x_2 &= (5 + 2x_1 - 2x_3) / 7 \\
 x_3 &= (1 + x_1 + 2x_2) / 5
 \end{aligned}$$

- First iteration

$$\begin{aligned}
 x_1^1 &= (11 + 2x_2^0 - x_3^0) / 6 = (11 + 0 - 0) / 6 = 1.833 \\
 x_2^1 &= (5 + 2x_1^1 - 2x_3^0) / 7 = (5 + 2*1.8333 - 0) / 7 = 1.238 \\
 x_3^1 &= (1 + x_1^1 + 2x_2^1) / 5 = (1 + 1.8333 + 2*1.2381) / 5 = 1.062
 \end{aligned}$$

- Second iteration

$$\begin{aligned}
 x_1^2 &= (11 + 2x_2^1 - x_3^1) / 6 = (11 + 2*1.238 - 1.062) / 6 = 2.069 \\
 x_2^2 &= (5 + 2x_1^2 - 2x_3^1) / 7 = (5 + 2*2.069 - 2*1.062) / 7 = 1.002 \\
 x_3^2 &= (1 + x_1^2 + 2x_2^2) / 5 = (1 + 2.069 + 2*1.002) / 5 = 1.015
 \end{aligned}$$

Gauss Seidel Method

Note: Always check the convergence using specified tolerance

	Zeroth	First	Second	Third	Fourth	Fifth
x_1	0.000	1.833	2.069	1.998	1.999	2.000
x_2	0.000	1.238	1.002	0.995	1.000	1.000
x_3	0.000	1.062	1.015	0.998	1.000	1.000

Diagonally Dominant Matrix

Definition of Strictly Diagonally Dominant Matrix Theorem

An $n \times n$ matrix A is **strictly diagonally dominant** if the absolute value of each entry on the main diagonal is greater than the sum of the absolute values of the other entries in the same row. That is,

$$\begin{aligned}|a_{11}| &> |a_{12}| + |a_{13}| + \cdots + |a_{1n}| \\|a_{22}| &> |a_{21}| + |a_{23}| + \cdots + |a_{2n}| \\&\vdots \\|a_{nn}| &> |a_{n1}| + |a_{n2}| + \cdots + |a_{n,n-1}|\end{aligned}$$

Which of the following systems of linear equations has a strictly diagonally dominant coefficient matrix?

(a) $3x_1 - x_2 = -4$
 $2x_1 + 5x_2 = 2$?

(b) $4x_1 + 2x_2 - x_3 = -1$
 ~~x_1~~ + $2x_3 = -4$?
 $3x_1 - 5x_2 + x_3 = 3$

Gauss Jacobi Method

- Gauss- Seidel always uses the newest available x values. Jacobi Method uses x values from the previous iteration.

Example 19: Repeat the previous example using the Jacobi method.

- Rearrange the equations

$$\begin{aligned}x_1 &= (11 + 2x_2 - x_3) / 6 \\x_2 &= (5 + 2x_1 - 2x_3) / 7 \\x_3 &= (1 + x_1 + 2x_2) / 5\end{aligned}$$

- First iteration (use x^0 values)

$$\begin{aligned}x_1^1 &= (11 + 2x_2^0 - x_3^0) / 6 = (11 + 0 - 0) / 6 = 1.833 \\x_2^1 &= (5 + 2x_1^0 - 2x_3^0) / 7 = (5 + 0 - 0) / 7 = 0.714 \\x_3^1 &= (1 + x_1^0 + 2x_2^0) / 5 = (1 + 0 + 0) / 5 = 0.200\end{aligned}$$

- Second iteration (use x^1 values)

$$\begin{aligned}x_1^2 &= (11 + 2x_2^1 - x_3^1) / 6 = (11 + 2*0.714 - 0.200) / 6 = 2.038 \\x_2^2 &= (5 + 2x_1^1 - 2x_3^1) / 7 = (5 + 2*1.833 - 2*0.200) / 7 = 1.181 \\x_3^2 &= (1 + x_1^1 + 2x_2^1) / 5 = (1 + 1.833 + 2*0.714) / 5 = 0.852\end{aligned}$$

Convergence of Jacobi Method

Show that for each of the following matrices \mathbf{A} , the system $\mathbf{Ax} = \mathbf{b}$ can be solved by Jacobi iteration with guaranteed convergence.

$$(a) \begin{bmatrix} 5 & -1 & 3 \\ 2 & -8 & 1 \\ -2 & 0 & 4 \end{bmatrix}$$

$$(b) \begin{bmatrix} -2 & 0 & 4 \\ 2 & -8 & 1 \\ 5 & -1 & 3 \end{bmatrix}$$

$$(c) \begin{bmatrix} 4 & 2 & -2 \\ 0 & 4 & 2 \\ 1 & 0 & 4 \end{bmatrix}$$

(a) This matrix is diagonally dominated since $|5| > |-1| + |3|$, $|-8| > |2| + |1|$ and $|4| > |-2| + |0|$ are all true. From the above theorem we know that the Jacobi 'iteration matrix' \mathbf{P} must have an ∞ -norm that is strictly less than 1, but this is also easy to verify directly:

$$\mathbf{P} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$$

$$= -\begin{bmatrix} 0.2 & 0 & 0 \\ 0 & -0.125 & 0 \\ 0 & 0 & 0.25 \end{bmatrix} \begin{bmatrix} 0 & -1 & 3 \\ 2 & 0 & 1 \\ -2 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0.2 & -0.6 \\ 0.25 & 0 & 0.125 \\ 0.5 & 0 & 0 \end{bmatrix}$$

Hence $\|\mathbf{P}\|_{\infty} = 0.8$, which as expected is < 1 .

Convergence of Jacobi Method

$$(b) \begin{bmatrix} -2 & 0 & 4 \\ 2 & -8 & 1 \\ 5 & -1 & 3 \end{bmatrix}$$

(b) This matrix is not diagonally dominated since $|-2| > |0| + |4|$ is false; the third row also violates. However, the first and third equations of $\mathbf{Ax} = \mathbf{b}$ can be swapped to give a new system $\mathbf{A}'\mathbf{x} = \mathbf{b}'$ with

$$\mathbf{A}' = \begin{bmatrix} 5 & -1 & 3 \\ 2 & -8 & 1 \\ -2 & 0 & 4 \end{bmatrix}$$

which is diagonally dominated; in fact it is the matrix of part (a). The reordered system is suitable for Jacobi iteration.

Convergence of Jacobi Method

(c)
$$\begin{bmatrix} 4 & 2 & -2 \\ 0 & 4 & 2 \\ 1 & 0 & 4 \end{bmatrix}$$
 Diagonal dominance is not necessary but sufficient condition

(c) This matrix is not diagonally dominated since $|4| > |2| + |-2|$ is false (we need strict satisfaction of the inequality in each row). In this case, we cannot achieve diagonal dominance by suitably reordering the rows. We therefore press on to compute the Jacobi 'iteration matrix'

$$\mathbf{P} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$$

$$= -\begin{bmatrix} 0.25 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 0.25 \end{bmatrix} \begin{bmatrix} 0 & 2 & -2 \\ 0 & 0 & 2 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -0.5 & 0.5 \\ 0 & 0 & -0.5 \\ -0.25 & 0 & 0 \end{bmatrix}$$

and see what its norms look like. The row-sum norm $\|\mathbf{P}\|_{\infty} = 1$: Column sum

norm $\|\mathbf{P}\|_1 = 1$. We are still looking for a matrix norm of \mathbf{P} that is strictly less than 1. The Frobenius norm is our last chance, and happily we find

$$\|\mathbf{P}\|_{\text{Fro}} = \sqrt{(-0.5)^2 + (0.5)^2 + (-0.5)^2 + (-0.25)^2} = 0.901$$

Example of Divergence

Ex 1 Apply the Jacobi method to the system

$$x_1 - 5x_2 = -4$$

$$7x_1 - x_2 = 6,$$

using the initial approximation $(x_1, x_2) = (0, 0)$, and show that the method diverges.

As usual, begin by rewriting the given system in the form

$$x_1 = -4 + 5x_2$$

$$x_2 = -6 + 7x_1.$$

Then the initial approximation $(0, 0)$ produces

$$x_1 = -4 + 5(0) = -4$$

$$x_2 = -6 + 7(0) = -6$$

as the first approximation. Repeated iterations produce the sequence of approximations shown in Table 10.3.

TABLE 10.3

n	0	1	2	3	4	5	6	7
x_1	0	-4	-34	-174	-1244	-6124	-42,874	-214,374
x_2	0	-6	-34	-244	-1244	-8574	-42,874	-300,124

Example of Divergence

TABLE 10.3

JACOBI METHOD

n	0	1	2	3	4	5	6	7
x_1	0	-4	-34	-174	-1244	-6124	-42,874	-214,374
x_2	0	-6	-34	-244	-1244	-8574	-42,874	-300,124

For this particular system of linear equations you can determine that the actual solution is $x_1 = 1$ and $x_2 = 1$. So you can see from Table 10.3 that the approximations given by the Jacobi method become progressively *worse* instead of better, and you can conclude that the method diverges.

TABLE 10.4

GAUSS SEIDEL METHOD

n	0	1	2	3	4	5
x_1	0	-4	-174	-6124	-214,374	-7,503,124
x_2	0	-34	-1224	-42,874	-1,500,624	-52,521,874

Diagonally Dominant Matrix

Ex 1 Begin by interchanging the two rows of the given system to obtain

$$7x_1 - x_2 = 6$$

$$x_1 - 5x_2 = -4.$$

Note that the coefficient matrix of this system is strictly diagonally dominant. Then solve for x_1 and x_2 as follows.

$$x_1 = \frac{6}{7} + \frac{1}{7}x_2$$

$$x_2 = \frac{4}{5} + \frac{1}{5}x_1$$

Using the initial approximation $(x_1, x_2) = (0, 0)$, you can obtain the sequence of approximations shown in Table 10.5.

TABLE 10.5

n	0	1	2	3	4	5
x_1	0.0000	0.8571	0.9959	0.9999	1.000	1.000
x_2	0.0000	0.9714	0.9992	1.000	1.000	1.000

So you can conclude that the solution is $x_1 = 1$ and $x_2 = 1$.

Diagonally Dominant Matrix

Do not conclude that strict diagonal dominance is a necessary condition for convergence of the Gauss Jacobi or Gauss- Seidel methods. For instance, the coefficient matrix of the system

$$\begin{aligned} -4x_1 + 5x_2 &= 1 \\ x_1 + 2x_2 &= 3 \end{aligned}$$

is not a strictly diagonally dominant matrix, and yet both methods converge to the solution $x_1 = 1$ and $x_2 = 1$ when you use an initial approximation of $(x_1, x_2) = (0,0)$.



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No.3

Agenda

- Fields
 - Vector spaces and subspaces
 - Linear independence and dependence
 - Bases and dimensions
 - Linear transformations
-

Field – Definition and Examples



Group : $(G, *)$ is a group if

- i.* is closed
- ii.* is associative
- iii.* has an identity
- iv.* has an inverse

Eg: $\langle \mathbb{R}, + \rangle, \langle \mathbb{R}, * \rangle$

$\langle G, * \rangle$ is **Abelian** if $a * b = b * a \quad \forall a, b \in G$

Eg: $\langle \mathbb{R}, + \rangle, \langle \mathbb{R}, * \rangle$ are Abelian

$\langle F, +, . \rangle$ is a **Field** if $\langle F, + \rangle$ and $\langle F, . \rangle$ are Abelian

Eg: $\langle \mathbb{R}, +, . \rangle, \langle \mathbb{C}, +, . \rangle, \langle \mathbb{Q}, +, . \rangle$

Vector Space

Real Vector Space

A non-empty set V of elements $\mathbf{a}, \mathbf{b}, \dots$ is called a **real vector space** (or *real linear space*) over a field F , and these elements are called **vectors** (regardless of their nature, which will come out from the context or will be left arbitrary) if, in V , there are defined two algebraic operations (called *vector addition* and *scalar multiplication*) as follows.

I. Vector addition associates with every pair of vectors \mathbf{a} and \mathbf{b} of V a unique vector of V , called the *sum* of \mathbf{a} and \mathbf{b} and denoted by $\mathbf{a} + \mathbf{b}$, such that the following axioms are satisfied.

Vector Space

Real Vector Space (continued 1)

I.1 Commutativity. For any two vectors \mathbf{a} and \mathbf{b} of V ,

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}.$$

I.2 Associativity. For any three vectors \mathbf{a} , \mathbf{b} , \mathbf{c} of V ,

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}) \quad (\text{written } \mathbf{a} + \mathbf{b} + \mathbf{c}).$$

I.3 There is a unique vector in V , called the *zero vector* and denoted by $\mathbf{0}$, such that for every \mathbf{a} in V ,

$$\mathbf{a} + \mathbf{0} = \mathbf{a}.$$

I.4 For every \mathbf{a} in V , there is a unique vector in V that is denoted by $-\mathbf{a}$ and is such that

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}.$$

Vector Space

Real Vector Space (continued 2)

II. Scalar multiplication. The real numbers are called **scalars**. Scalar multiplication associates with every \mathbf{a} in V and every scalar c a unique vector of V , called the *product* of c and \mathbf{a} and denoted by $c\mathbf{a}$ (or $\mathbf{a}c$) such that the following axioms are satisfied.

II.1 Distributivity. For every scalar c and vectors \mathbf{a} and \mathbf{b} in V ,

$$c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}.$$

II.2 Distributivity. For all scalars c and k and every \mathbf{a} in V ,

$$(c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a}.$$

II.3 Associativity. For all scalars c and k and every \mathbf{a} in V ,

$$c(k\mathbf{a}) = (ck)\mathbf{a} \quad (\text{written } cka).$$

II.4 For every \mathbf{a} in V ,

$$1\mathbf{a} = \mathbf{a}.$$

Subspace

By a **subspace** of a vector space V we mean

“a nonempty subset of V (including V itself) that forms a vector space with respect to the two algebraic operations (addition and scalar multiplication) defined for the vectors of V . ”

- Space $(W, +, \cdot)$: within a vector space
- $W \neq \Phi$ and $W \subseteq (V, +, \cdot)$ over F is a subspace if
 - $0 \in W, \alpha w_1 + w_2 \in W$
- Ex: $V = \{(x_1, x_2) \mid x_1, x_2 \in R\}$ over R , $W = \{(x_1, 0) \mid x_1 \in R\}$
- Set of singular matrices is not a subspace of $M_{2 \times 2}$

Linear Dependence and Independence of Vectors

Given any set of m vectors $\mathbf{a}_{(1)}, \dots, \mathbf{a}_{(m)}$ (with the same number of components), a **linear combination** of these vectors is an expression of the form

$$c_1 \mathbf{a}_{(1)} + c_2 \mathbf{a}_{(2)} + \dots + c_m \mathbf{a}_{(m)}$$

where c_1, c_2, \dots, c_m are any scalars.

Now consider the equation

$$(1) \quad c_1 \mathbf{a}_{(1)} + c_2 \mathbf{a}_{(2)} + \dots + c_m \mathbf{a}_{(m)} = \mathbf{0}$$

Clearly, *this vector equation (1) holds if we choose all c_j 's zero, because then it becomes $\mathbf{0} = \mathbf{0}$.*

If this is the only m -tuple of scalars for which (1) holds, then our vectors $\mathbf{a}_{(1)}, \dots, \mathbf{a}_{(m)}$ are said to form a *linearly independent set* or, more briefly, we call them **linearly independent**.

Linear Dependence and Independence of Vectors



Otherwise, if (1) also holds with scalars not all zero, we call these vectors **linearly dependent**.

This means that we can express at least one of the vectors as a linear combination of the other vectors. For instance, if (1) holds with, say, $c_1 \neq 0$, we can solve (1) for $\mathbf{a}_{(1)}$:

$$\mathbf{a}_{(1)} = k_2 \mathbf{a}_{(2)} + \dots + k_m \mathbf{a}_{(m)} \text{ where } k_j = -c_j/c_1.$$

The **rank** of a matrix \mathbf{A} is the maximum number of linearly independent row vectors of \mathbf{A} .

It is denoted by $\text{rank } \mathbf{A}$.

Basis and Dimension

*A linearly independent set in V consisting of a maximum possible number of vectors in V is called a **basis** for V . The number of vectors of a basis for V equals $\dim V$.*

The vector space R^n over R consisting of all vectors with n components (n real numbers) has dimension n .

- R over $R \rightarrow$ One dimensional vector space
 - C over $C \rightarrow$ One dimensional vector space
 - C over $R \rightarrow$ Two dimensional vector space
 - If $W = \{0\}$, then W is generated by Φ and hence $\dim(W) = 0$
-

Row Space and Column Space



- If A is an $m \times n$ matrix
 - the subspace of R^n spanned by the row vectors of A is called the **row space** of A
 - the subspace of R^m spanned by the column vectors is called the **column space** of A

The solution space of the homogeneous system of equation $Ax = 0$, which is a subspace of R^n , is called the nullspace of A .

$$A_{m \times n} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

$$\mathbf{c}_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{m1} \end{bmatrix}, \quad \mathbf{c}_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{m2} \end{bmatrix}, \dots, \mathbf{c}_n = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{bmatrix}$$

Basis for Row Space and Column Space

If a matrix R is in row echelon form

- the row vectors with the leading 1's (i.e., the nonzero row vectors) form a basis for the row space of R
- the column vectors with the leading 1's of the row vectors form a basis for the column space of R

Basis for Row Space

Find a basis of row space of

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

Sol:

$$A = \begin{bmatrix} 1 & 3 & 1 & 3 \\ 0 & 1 & 1 & 0 \\ -3 & 0 & 6 & -1 \\ 3 & 4 & -2 & 1 \\ 2 & 0 & -4 & 2 \end{bmatrix} \xrightarrow{\text{G.E.}} B = \begin{bmatrix} 1 & 3 & 1 & 3 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \mathbf{w}_3 \\ \mathbf{b}_1 \\ \mathbf{b}_2 \\ \mathbf{b}_3 \\ \mathbf{b}_4 \end{matrix}$$

a basis for $RS(A) = \{\text{the nonzero row vectors of } B\}$
 $= \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3\} = \{(1, 3, 1, 3), (0, 1, 1, 0), (0, 0, 0, 1)\}$

Basis for Column Space

Find a basis for the column space
of the matrix A .

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

$$\begin{matrix} a_1 & a_2 & a_3 & a_4 & a_5 \end{matrix}$$

Reduce A to the reduced row- echelon form

$$E = \begin{bmatrix} 1 & 2 & 0 & 3 & 0 \\ 0 & 0 & 1 & 5 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} = [e_1 \ e_2 \ e_3 \ e_4 \ e_5]$$

$$e_2 = 2e_1 \rightarrow a_2 = 2a_1$$

$$e_4 = 3e_1 + 5e_3 \rightarrow a_4 = 3a_1 + 5a_3$$

$\{a_1, a_3, a_5\}$ is a basis for column space of A

Solution Space/ Null Space

Find the solution space of a homogeneous system $A\mathbf{x} = \mathbf{0}$.

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

$$A = \begin{bmatrix} 1 & 2 & -2 & 1 \\ 3 & 6 & -5 & 4 \\ 1 & 2 & 0 & 3 \end{bmatrix} \xrightarrow{\text{G.J.E}} \begin{bmatrix} 1 & 2 & 0 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad x_1 = -2s - 3t, x_2 = s, \\ x_3 = -t, x_4 = t$$

$$\Rightarrow \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} -2s - 3t \\ s \\ -t \\ t \end{bmatrix} = s \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + t \begin{bmatrix} -3 \\ 0 \\ -1 \\ 1 \end{bmatrix} = s\mathbf{v}_1 + t\mathbf{v}_2 \\ \Rightarrow NS(A) = \{s\mathbf{v}_1 + t\mathbf{v}_2 \mid s, t \in R\}$$

Nullity(A) = dim(NS(A)) = 2

RS, CS, Rank(A)

Theorem : (Row and column space have equal dimensions)

- If A is an $m \times n$ matrix, then the row space and the column space of A have the same dimension.

$$\dim(RS(A)) = \dim(CS(A))$$

- **Rank:**

The dimension of the row (or column) space of a matrix A is called the rank of A .

$$\text{rank}(A) = \dim(RS(A)) = \dim(CS(A))$$

Notes: $\text{rank}(A^T) = \dim(RS(A^T)) = \dim(CS(A)) = \text{rank}(A)$

Therefore $\text{rank}(A^T) = \text{rank}(A)$

Rank Nullity Theorem

- **Nullity:** The dimension of the nullspace of A is called the nullity of A

$$\text{nullity}(A) = \dim(NS(A))$$

- **Theorem :** (Dimension of the solution space)

If A is an $m \times n$ matrix of rank r , then the dimension of the solution space of $Ax = 0$ is $n - r$. That is

$$\text{nullity}(A) = n - \text{rank}(A) = n - r$$

$$n = \text{rank}(A) + \text{nullity}(A)$$

Rank Nullity Theorem for Matrix

Rank and Nullity of Matrix

If A is an $m \times n$ matrix and $\text{rank}(A) = r$, then

Fundamental Space	Dimension
-------------------	-----------

$$RS(A) = CS(A^T) \quad r$$

$$CS(A) = RS(A^T) \quad r$$

$$NS(A) \quad n - r$$

$$NS(A^T) \quad m - r$$

Linear Transformation

Let X and Y be any vector spaces. To each vector \mathbf{x} in X we assign a unique vector \mathbf{y} in Y . Then we say that a **mapping** (or **transformation** or **operator**) of X into Y is given.

Such a mapping is denoted by a capital letter, say F . The vector \mathbf{y} in Y assigned to a vector \mathbf{x} in X is called the **image** of \mathbf{x} under F and is denoted by $F(\mathbf{x})$ [or $F\mathbf{x}$, without parentheses].

F is called a **linear mapping** or **linear transformation** if, for all vectors \mathbf{v} and \mathbf{x} in X and scalars c ,

$$(10) \quad F(\mathbf{v} + \mathbf{x}) = F(\mathbf{v}) + F(\mathbf{x})$$
$$F(c\mathbf{x}) = cF(\mathbf{x}).$$

Linear Transformation of Space R^n into Space R^m

From now on we let $X = R^n$ and $Y = R^m$. Then any real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ gives a transformation of R^n into R^m ,

$$(11) \quad \mathbf{y} = \mathbf{Ax}.$$

Since $\mathbf{A}(\mathbf{u} + \mathbf{x}) = \mathbf{Au} + \mathbf{Ax}$ and $\mathbf{A}(c\mathbf{x}) = c\mathbf{Ax}$, this transformation is linear.

If \mathbf{A} in (11) is square, $n \times n$, then (11) maps R^n into R^n . If this \mathbf{A} is non-singular, so that \mathbf{A}^{-1} exists (see Sec. 7.8), then multiplication of (11) by \mathbf{A}^{-1} from the left and use of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives the **inverse transformation**

$$(14) \quad \mathbf{x} = \mathbf{A}^{-1} \mathbf{y}.$$

It maps every $\mathbf{y} = \mathbf{y}_0$ onto that \mathbf{x} , which by (11) is mapped onto \mathbf{y}_0 . *The inverse of a linear transformation is itself linear,* because it is given by a matrix, as (14) shows.

Range and Kernel

- Let $T : V \rightarrow W$ be linear transformation
- $\text{Range}(T) = \{T(v) \mid v \in V\}$ is subspace of W
- $\text{Kernel}(T) = \{v \in V \mid T(v) = \mathbf{0}\}$ is subspace of V
- $\text{Nullity}(T) = \dim(\text{Kernel}(T))$
- $\boxed{\text{Rank}(T) = \dim(\text{Range}(T))}$
- Rank Nullity Theorem for Linear Transformation
 $\dim(\text{Kernel}(T)) + \dim(\text{Range}(T)) = \dim V$

Rank Nullity Theorem Example

$T : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, $T(x) = Ax$ where $A = \begin{bmatrix} 1 & 2 \\ 3 & -1 \end{bmatrix}$ Find the rank and nullity of linear transformation and verify the Rank Nullity Theorem

Let $v = [x, y]$ be a vector of \mathbb{R}^2 , $v \in \ker(T)$

$$Av = 0$$

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

Augmented Matrix
$$\left[\begin{array}{cc|c} 1 & 2 & 0 \\ 3 & -1 & 0 \end{array} \right]$$

$$x = 0, y = 0 \rightarrow \ker T = [0, 0]$$

Range Space of T is $R(T) = \text{col}(A)$

Rank Nullity Theorem Example

Range Space of T is $R(T) = \text{col}(A)$

$$= \left\{ x_1 \begin{bmatrix} 1 \\ 3 \end{bmatrix} + x_2 \begin{bmatrix} 2 \\ -1 \end{bmatrix} \mid x_1, x_2 \in \mathbb{R} \right\}$$

Nullity $T = 0$

$$\text{Basis for } R(T) = \left\{ \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \end{bmatrix} \right\}$$

Rank $T = \dim [R(T)] = 2$

Rank T + nullity $T = 2 + 0 = 2 = \dim V$

$\dim V = \dim \mathbb{R}^2 = 2$

Hence Rank Nullity Theorem is verified



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 4

Agenda

- Definition and examples of eigenvalues
 - Computation of eigenvalues and eigenvectors
 - Diagonalization and Jordan Canonical Form
 - Gershgorin's theorem
 - Properties of eigenvalues and eigenvectors
-

Eigenvalue Problem

A matrix eigenvalue problem considers the vector equation

$$(1) \quad \mathbf{Ax} = \lambda \mathbf{x},$$

where \mathbf{A} is a given square matrix, λ an unknown scalar (real or complex), and \mathbf{x} an unknown vector.

The task is to determine λ 's and \mathbf{x} 's (dependent on λ 's) that satisfy (1).

Since $\mathbf{x} = \mathbf{0}$ is always a solution for any λ , we only admit solutions with $\mathbf{x} \neq \mathbf{0}$.

The solutions to (1) are given the following names: The λ 's that satisfy (1) are called **eigenvalues of \mathbf{A}** and the corresponding nonzero \mathbf{x} 's that also satisfy (1) are called **eigenvectors of \mathbf{A}** .

Eigen Value Analysis

This example illustrates the general case as follows. Equation (1) written in components is

$$a_{11}x_1 + \cdots + a_{1n}x_n = \lambda x_1$$

$$a_{21}x_1 + \cdots + a_{2n}x_n = \lambda x_2$$

.....

$$a_{n1}x_1 + \cdots + a_{nn}x_n = \lambda x_n.$$

Transferring the terms on the right side to the left side, we have

$$(2) \quad \begin{aligned} (a_{11} - \lambda)x_1 + & \quad a_{12}x_2 + \cdots + & a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + & \cdots + & a_{2n}x_n &= 0 \\ \dots & & & \\ a_{n1}x_1 + & \quad a_{n2}x_2 + \cdots + (a_{nn} - \lambda)x_n & &= 0. \end{aligned}$$

Eigen Value Analysis

In matrix notation,

$$(3) \quad (\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}.$$

By Cramer's theorem in Sec. 7.7, this homogeneous linear system of equations has a nontrivial solution if and only if the corresponding determinant of the coefficients is zero:

$$(4) \quad D(\lambda) = \det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Eigenvalue Analysis

$A - \lambda I$ is called the **characteristic matrix** and $D(\lambda)$ the **characteristic determinant** of A . Equation (4) is called the **characteristic equation** of A . By developing $D(\lambda)$ we obtain a polynomial of n th degree in λ . This is called the **characteristic polynomial** of A .

Eigenvalues

Theorem 1

Eigenvalues

The eigenvalues of a square matrix \mathbf{A} are the roots of the characteristic equation (4) of \mathbf{A} .

Hence an $n \times n$ matrix has at least one eigenvalue and at most n numerically different eigenvalues.

The eigenvalues must be determined first.

Once these are known, corresponding eigenvectors are obtained from the system (2), for instance, by the Gauss elimination, where λ is the eigenvalue for which an eigenvector is wanted.

Eigen Space

Theorem 2

Eigenvectors, Eigenspace

If \mathbf{w} and \mathbf{x} are eigenvectors of a matrix \mathbf{A} corresponding to the same eigenvalue λ , so are $\mathbf{w} + \mathbf{x}$ (provided $\mathbf{x} \neq -\mathbf{w}$) and $k\mathbf{x}$ for any $k \neq 0$.

Proof: follows from definition.

Hence the eigenvectors corresponding to one and the same eigenvalue λ of \mathbf{A} , together with $\mathbf{0}$, form a vector space called the eigenspace of \mathbf{A} corresponding to that λ .

Spectrum

The set of all the eigenvalues of \mathbf{A} is called the **spectrum** of \mathbf{A} . We shall see that the spectrum consists of at least one eigenvalue and at most of n numerically different eigenvalues.

The largest of the absolute values of the eigenvalues of \mathbf{A} is called the *spectral radius* of \mathbf{A} .

Determination of Eigen Value and Eigen Vector

$$\mathbf{A} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}.$$

Solution.

(a) *Eigenvalues.* These must be determined *first*.

Equation (1) is in components

$$\mathbf{Ax} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda \begin{bmatrix} x_1 \\ x_2 \end{bmatrix};$$

$$-5x_1 + 2x_2 = \lambda x_1$$

$$2x_1 - 2x_2 = \lambda x_2.$$

Example

Solution. (continued 1)

(a) *Eigenvalues. (continued 1)*

Transferring the terms on the right to the left, we get

$$(2^*) \quad \begin{aligned} (-5 - \lambda)x_1 + & \qquad 2x_2 = 0 \\ 2x_1 + (-2 - \lambda)x_2 = & 0 \end{aligned}$$

This can be written in matrix notation

$$(3^*) \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$$

Because (1) is $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{Ax} - \lambda\mathbf{Ix} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = 0$,
which gives (3*).

Example

Solution. (continued 2)

(a) *Eigenvalues.* (continued 2)

We see that this is a *homogeneous* linear system. By Cramer's theorem in Sec. 7.7 it has a nontrivial solution (an eigenvector of \mathbf{A} we are looking for) if and only if its coefficient determinant is zero, that is,

$$\begin{aligned}
 D(\lambda) &= \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -5-\lambda & 2 \\ 2 & -2-\lambda \end{vmatrix} \\
 (4^*) &= (-5-\lambda)(-2-\lambda) - 4 = \lambda^2 + 7\lambda + 6 = 0.
 \end{aligned}$$

Example

Solution. (continued 3)

(a) *Eigenvalues.* (continued 3)

We call $D(\lambda)$ the **characteristic determinant** or, if expanded, the **characteristic polynomial**, and $D(\lambda) = 0$ the **characteristic equation** of A . The solutions of this quadratic equation are $\lambda_1 = -1$ and $\lambda_2 = -6$. These are the eigenvalues of A .

(b₁) *Eigenvector of A corresponding to λ_1 .* This vector is obtained from (2*) with $\lambda = \lambda_1 = -1$, that is,

$$-4x_1 + 2x_2 = 0$$

$$2x_1 - x_2 = 0.$$

Example

Solution. (continued 4)

(b₁) *Eigenvector of A corresponding to λ₁.* (continued)

A solution is $x_2 = 2x_1$, as we see from either of the two equations, so that we need only one of them. This determines an eigenvector corresponding to $λ_1 = -1$ up to a scalar multiple. If we choose $x_1 = 1$, we obtain the eigenvector

$$v = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \text{ Check: } Av = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix} = (-1)v = λ_1 v.$$

Example

Solution. (continued 5)

(b₂) *Eigenvector of A corresponding to λ₂.*

For λ = λ₂ = -6, equation (2*) becomes

$$x_1 + 2x_2 = 0$$

$$2x_1 + 4x_2 = 0.$$

A solution is x₂ = -x₁/2 with arbitrary x₁. If we choose x₁ = 2, we get x₂ = -1. Thus an eigenvector of A corresponding to λ₂ = -6 is

$$w = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \text{ Check: } Aw = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -12 \\ 6 \end{bmatrix} = (-6)w = \lambda_2 w.$$

Multiple Eigen Values

Example 2: Find the eigenvalues and eigenvectors of

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

Solution.

For our matrix, the characteristic determinant gives the characteristic equation

$$-\lambda^3 - \lambda^2 + 21\lambda + 45 = 0.$$

The roots (eigenvalues of \mathbf{A}) are $\lambda_1 = 5$, $\lambda_2 = \lambda_3 = -3$.

Multiple Eigen Values

Solution. (continued 1)

To find eigenvectors, we apply the Gauss elimination (Sec. 7.3) to the system $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$, first with $\lambda = 5$ and then with $\lambda = -3$. For $\lambda = 5$ the characteristic matrix is

$$\mathbf{A} - \lambda \mathbf{I} = \mathbf{A} - 5\mathbf{I} = \begin{bmatrix} -7 & 2 & -3 \\ 2 & -4 & -6 \\ -1 & -2 & -5 \end{bmatrix}.$$

It row-reduces to

$$\begin{bmatrix} -7 & 2 & -3 \\ 0 & -24/7 & -48/7 \\ 0 & 0 & 0 \end{bmatrix}.$$

Multiple Eigen Values

Solution. (continued 2)

Hence it has rank 2. Choosing $x_3 = -1$ we have $x_2 = 2$ from $-\frac{24}{7}x_2 - \frac{48}{7}x_3 = 0$ and then $x_1 = 1$ from $-7x_1 + 2x_2 - 3x_3 = 0$.

Hence an eigenvector of \mathbf{A} corresponding to $\lambda = 5$ is

$$\mathbf{x}_1 = [1 \ 2 \ -1]^T.$$

For $\lambda = -3$ the characteristic matrix

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} + 3\mathbf{I} = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & -6 \\ -1 & -2 & 3 \end{bmatrix}$$

row-reduces to

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

Multiple Eigenvalues

Solution. (continued 3)

Hence it has rank 1.

From $x_1 + 2x_2 - 3x_3 = 0$ we have $x_1 = -2x_2 + 3x_3$. Choosing $x_2 = 1, x_3 = 0$ and $x_2 = 0, x_3 = 1$, we obtain two linearly independent eigenvectors of \mathbf{A} corresponding to $\lambda = -3$ [as they must exist by (5), Sec. 7.5, with rank = 1 and $n = 3$],

$$\mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_3 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}.$$

Jordan Canonical Form

- Suppose A has an eigenvalue λ_i is repeated k times, we can find the nullspace of $(A - \lambda_i I)^j$ $1 \leq j \leq k$ and form P
- In this case $P^{-1}AP = J$, where J is an upper triangular matrix with superdiagonal 1

$$\begin{pmatrix} \mathbf{J}_{\lambda_1} & & & & \\ & \mathbf{J}_{\lambda_1} & & & \mathbf{O} \\ & & \ddots & & \\ & \mathbf{O} & & \mathbf{J}_{\lambda_{k-1}} & \\ & & & & \mathbf{J}_{\lambda_k} \end{pmatrix}$$

- Exemplification

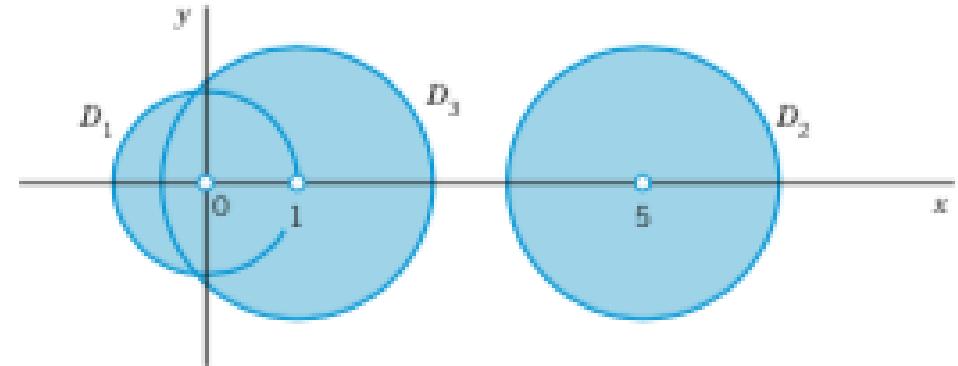
Gerschgorin's Theorem

Theorem gives the bound on Eigenvalues

Every eigenvalue of matrix $A_{n \times n}$ satisfies :

$$\lambda - |a_{ii}| \leq \sum_{j \neq i} |a_{ij}|, \quad i = 1, 2, \dots, n$$

Example : $A = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 5 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix}$



We get Gerschgorin disks D_1 : Centre 0, radius 1

D_2 : Centre 5, radius 1.5

D_3 : Centre 1, radius 1.5

The centers are main diagonal entries of A. These would be the eigenvalues of A if A were diagonal

Algebraic Multiplicity &Geometric Multiplicity

The order M_λ of an eigenvalue λ as a root of the characteristic polynomial is called the **algebraic multiplicity** of λ . The number m_λ of linearly independent eigenvectors corresponding to λ is called the **geometric multiplicity** of λ . Thus m_λ is the dimension of the eigenspace corresponding to this λ .

Since the characteristic polynomial has degree n , the sum of all the algebraic multiplicities must equal n . In Example 2 for $\lambda = -3$ we have $m_\lambda = M_\lambda = 2$. In general, $m_\lambda \leq M_\lambda$, as can be shown. The difference $\Delta_\lambda = M_\lambda - m_\lambda$ is called the **defect** of λ . Thus $\Delta_{-3} = 0$ in Example 2, but positive defects Δ_λ can easily occur.

Special cases

Theorem 3

Eigenvalues of the Transpose

The transpose \mathbf{A}^T of a square matrix \mathbf{A} has the same eigenvalues as \mathbf{A} .

Basis of Eigenvectors

If an $n \times n$ matrix \mathbf{A} has n distinct eigenvalues, then \mathbf{A} has a basis of eigenvectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ for \mathbb{R}^n .

Eigenvalue properties

- If λ is eigenvalue of A , then λ^k is eigenvalue of A^k .
- If λ is eigenvalue of A and A is invertible, then $\frac{1}{\lambda}$ is eigenvalue of A^{-1} .
- $\text{Trace}(A) = \text{Sum of eigenvalues of matrix}$

$$= \lambda_1 + \lambda_2 + \dots + \lambda_n = \sum_{j=1}^n \lambda_j$$
- $\text{Determinant}(A) = \text{Product of eigenvalues of matrix}$

$$= \lambda_1 \lambda_2 \dots \lambda_n = \prod_{j=1}^n \lambda_j$$

Eigenvalue properties

Theorem 4

The eigenvalues of

- a) A symmetric matrix (Hermitian) is real
- b) A skew-symmetric matrix (skew-Hermitian) is either 0 or purely imaginary
- c) An orthogonal matrix has absolute value 1

Proof.: Observe that $\lambda = \frac{x^T Ax}{x^T x}$ for the eigenvector corresponding to λ . Replace x by \bar{x}



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 5

Agenda

- Recapitulation of concepts on Eva, Eve
 - Eva, Eve of special matrices
 - Similarity transformation
 - Diagonalization of matrices
 - Numerical method for Eva, Eve
-

Eigenvalue properties

Theorem 4

The eigenvalues of

- a) A symmetric matrix (Hermitian) is real
- b) A skew-symmetric matrix (skew-Hermitian) is either 0 or purely imaginary
- c) An orthogonal matrix has absolute value 1

Proof.: Observe that $\lambda = \frac{x^T Ax}{x^T x}$ for the eigenvector corresponding to λ . Replace x by \bar{x}

Eigenvectors

Theorem 5

Let A be a $n \times n$ matrix having distinct eigenvalues $\lambda_1, \lambda_2, \lambda_3, \dots, \lambda_n$ then the eigenvectors corresponding to these eigenvalues are linearly independent.

That is then A has a basis of eigenvectors x_1, \dots, x_n for \mathbb{R}^n .

Proof: Constructive.

Eigenvectors corresponding to Distinct Eigenvalues are Linearly Independent

Let k be the smallest positive integer such that v_1, v_2, \dots, v_k are linearly independent. If $k = p$, nothing is to be proved.

If $k < p$, then v_{k+1} is a linear combination of v_1, \dots, v_k ; that is, there exist constants c_1, c_2, \dots, c_k such that

$$v_{k+1} = c_1 v_1 + c_2 v_2 + \dots + c_k v_k.$$

Applying the matrix A to both sides, we have

$$\begin{aligned} Av_{k+1} &= \lambda_{k+1} v_{k+1} \\ &= \lambda_{k+1} (c_1 v_1 + c_2 v_2 + \dots + c_k v_k) \\ &= c_1 \lambda_k v_1 + c_2 \lambda_k v_2 + \dots + c_k \lambda_k v_k; \end{aligned}$$

$$\begin{aligned} Av_{k+1} &= A(c_1 v_1 + c_2 v_2 + \dots + c_k v_k) \\ &= c_1 Av_1 + c_2 Av_2 + \dots + c_k Av_k \\ &= c_1 \lambda_1 v_1 + c_2 \lambda_2 v_2 + \dots + c_k \lambda_k v_k. \end{aligned}$$

Thus $c_1(\lambda_{k+1} - \lambda_1)v_1 + c_2(\lambda_{k+1} - \lambda_2)v_2 + \dots + c_k(\lambda_{k+1} - \lambda_k)v_k = 0$.

Since v_1, v_2, \dots, v_k are linearly independent, we have

$$c_1(\lambda_{k+1} - \lambda_1) = c_2(\lambda_{k+1} - \lambda_2) = \dots = c_k(\lambda_{k+1} - \lambda_k) = 0.$$

Note that the eigenvalues are distinct. Hence

$$c_1 = c_2 = \dots = c_k = 0,$$

which implies that v_{k+1} is the zero vector. This is contradictory to $v_{k+1} \neq 0$.

Similarity of Matrices

Similar Matrices. Similarity Transformation

An $n \times n$ matrix $\hat{\mathbf{A}}$ is called **similar** to an $n \times n$ matrix \mathbf{A} if

$$(4) \quad \hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$$

for some (nonsingular!) $n \times n$ matrix \mathbf{P} . This transformation, which gives $\hat{\mathbf{A}}$ from \mathbf{A} , is called a **similarity transformation**.

Eigenvalues and Eigenvectors of Similar Matrices

If $\hat{\mathbf{A}}$ is similar to \mathbf{A} , then $\hat{\mathbf{A}}$ has the same eigenvalues as \mathbf{A} . Furthermore, if \mathbf{x} is an eigenvector of \mathbf{A} , then $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ is an eigenvector of $\hat{\mathbf{A}}$ corresponding to the same eigenvalue.

Diagonalization of a Matrix

Diagonalization of a Matrix

If an $n \times n$ matrix \mathbf{A} has a basis of eigenvectors, then

$$(5) \quad \mathbf{D} = \mathbf{X}^{-1}\mathbf{AX}$$

is diagonal, with the eigenvalues of \mathbf{A} as the entries on the main diagonal. Here \mathbf{X} is the matrix with these eigenvectors as column vectors. Also,

$$(5^*) \quad \mathbf{D}^m = \mathbf{X}^{-1}\mathbf{A}^m\mathbf{X} \quad (m = 2, 3, \dots).$$

Diagonalization of a Matrix

Diagonalize

$$\mathbf{A} = \begin{bmatrix} 7.3 & 0.2 & -3.7 \\ -11.5 & 1.0 & 5.5 \\ 17.7 & 1.8 & -9.3 \end{bmatrix}.$$

Solution.

The characteristic determinant gives the characteristic equation $-\lambda^3 - \lambda^2 + 12\lambda = 0$. The roots (eigenvalues of \mathbf{A}) are $\lambda_1 = 3$, $\lambda_2 = -4$, $\lambda_3 = 0$. By the Gauss elimination applied to $(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1, \lambda_2, \lambda_3$ we find eigenvectors and then \mathbf{X}^{-1} by the Gauss–Jordan elimination

Diagonalization of a Matrix

Solution. (continued 1)

The results are

$$\begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix},$$

$$\mathbf{X}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$

Diagonalization of a Matrix

Solution. (continued 2)

Calculating \mathbf{AX} and multiplying by \mathbf{X}^{-1} from the left, we thus obtain

$$\mathbf{D} = \mathbf{X}^{-1} \mathbf{AX} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} \begin{bmatrix} -3 & -4 & 0 \\ 9 & 4 & 0 \\ -3 & -12 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Dominant Eigenvalue

Let $\lambda_1, \lambda_2, \dots, \lambda_n$ be eigen values of $n \times n$ matrix A

λ_1 is called **dominant eigen value** of A if

$$|\lambda_1| > |\lambda_i| \text{ for all } i = 2, 3, \dots, n$$

The eigen vectors corresponding to λ_1 are called **dominant eigen vectors**

Note : Not every matrix has dominant eigen value

Not every matrix has a dominant eigenvalue. For instance, the matrix

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

(with eigenvalues of $\lambda_1 = 1$ and $\lambda_2 = -1$) has no dominant eigenvalue. Similarly, the matrix

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

(with eigenvalues of $\lambda_1 = 2$, $\lambda_2 = 2$, and $\lambda_3 = 1$) has no dominant eigenvalue.

Rayleigh's Quotient

Let A be an $n \times n$ real symmetric matrix

Let $x(\neq 0)$ be any real vector with n components

Let $y = Ax$, $m_0 = x^T x$, $m_1 = x^T y$, $m_2 = y^T y$

Rayleigh Quotient

$q = \frac{m_1}{m_0}$ is an approximation for an eigenvalue λ of A , and if we set $q = \lambda - \varepsilon$ so that ε is error of q , then

$$|\hat{\lambda} - q| = \sqrt{\frac{m_2}{m_0} - q^2}$$

Power Method for approximating Eigenvalues

$$\lambda^n + c_{n-1}\lambda^{n-1} + c_{n-2}\lambda^{n-2} + \cdots + c_0 = 0.$$

For large values of n , polynomial equations like this one are difficult and time-consuming to solve. Moreover, numerical techniques for approximating roots of polynomial equations of high degree are sensitive to rounding errors. In this section you will look at an alternative method for approximating eigenvalues. As presented here, the method can be used only to find the eigenvalue of A that is largest in absolute value—this eigenvalue is called the **dominant eigenvalue** of A . Although this restriction may seem severe, dominant eigenvalues are of primary interest in many physical applications.

$$\mathbf{x}_1 = A\mathbf{x}_0$$

$$\mathbf{x}_2 = A\mathbf{x}_1 = A(A\mathbf{x}_0) = A^2\mathbf{x}_0$$

$$\mathbf{x}_3 = A\mathbf{x}_2 = A(A^2\mathbf{x}_0) = A^3\mathbf{x}_0$$

$$\vdots$$

$$\mathbf{x}_k = A\mathbf{x}_{k-1} = A(A^{k-1}\mathbf{x}_0) = A^k\mathbf{x}_0.$$

Power Method

To find the largest eigenvalues of a $n \times n$ matrix A. Assume that A has dominant eigenvalue

1. STEP 1 : Choose a column vector $u_o = [1, 1, \dots, 1]^T$

2. STEP 2 : Compute $A * u_o$

3. STEP 3 : Normalize the resulting vector obtained in step 2 by dividing each component by the largest in magnitude

4. STEP 4 : Repeat steps 2 and 3 until the change in normalizing factor is negligible

CONCLUSION :

- The normalizing factor is an approximate value of eigen value
- The final vector is the corresponding eigen vector

Power Method – Example

Find the dominant eigenvalues and corresponding eigenvectors of the matrix

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

Step 1 : Choose column vector $u_o = [1, 1, 1]^T$

Step 2 : Multiply the matrix by the matrix [A] by $u_o = y_1$

$$\begin{bmatrix} 4 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix} = \begin{Bmatrix} 5 \\ 3 \\ -1 \end{Bmatrix} \Rightarrow \begin{Bmatrix} 5 \\ 3 \\ -1 \end{Bmatrix} = 5 \begin{Bmatrix} 1 \\ 0.6 \\ -0.2 \end{Bmatrix}$$

Step 3 : Normalize the resulting vector obtained in step 2 by dividing each component by the largest in magnitude

$$u_1 = y_1 / 5 = [1, 0.6, -0.2]^T \text{ Normalizing factor } m_1 = 5$$

Power Method - Example

Step 4 :

$$\begin{bmatrix} 4 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{Bmatrix} 1 \\ 0.6 \\ -0.2 \end{Bmatrix} = \begin{Bmatrix} 4.6 \\ 1 \\ 0.2 \end{Bmatrix}$$

$$\Rightarrow \begin{Bmatrix} 4.6 \\ 1 \\ 0.2 \end{Bmatrix} = 4.6 \begin{Bmatrix} 1 \\ 0.217 \\ 0.0435 \end{Bmatrix} \quad u_2 = y_2 / 4.6 \text{ (normalizing factor } m_2)$$

Power Method

Now Repeating steps 2 and 3

$$\begin{bmatrix} 4 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{Bmatrix} 1 \\ 0.217 \\ 0.0435 \end{Bmatrix} = \begin{Bmatrix} 4.2174 \\ 0.4783 \\ -0.0435 \end{Bmatrix}$$

$$\Rightarrow \begin{Bmatrix} 4.2174 \\ 0.4783 \\ -0.0435 \end{Bmatrix} = 4.2174 \begin{Bmatrix} 1 \\ 0.1134 \\ -0.0183 \end{Bmatrix}$$

$$u_3 = y_3 / 4.2174$$

(normalizing factor m_3)

Power Method – Example

Continue the process

$$\begin{bmatrix} 4 & 1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & -1 \end{bmatrix} \begin{Bmatrix} 1 \\ 0.1134 \\ -0.0183 \end{Bmatrix} = \begin{Bmatrix} 4.1134 \\ 0.2165 \\ 0.0103 \end{Bmatrix}$$

$$\Rightarrow \begin{Bmatrix} 4.1134 \\ 0.2165 \\ 0.0103 \end{Bmatrix} = 4.1134 \begin{Bmatrix} 1 \\ 0.0526 \\ 0.0025 \end{Bmatrix} \quad u_4 = y_4 / 4.1134 \\ (\text{normalizing factor } m_4)$$

m_1	m_2	m_3	m_4
5	4.6	4.2174	4.1134

Change in normalizing factor m_i 's is now negligible

LARGEST Eigenvalue is $m_4 = 4.1134$

Corresponding Eigen vector $u_4 = [1, 0, 0]^T$

Convergence of Power Method

If A is an $n \times n$ diagonalizable matrix with a dominant eigenvalue, then there exists a nonzero vector \mathbf{x}_0 such that the sequence of vectors given by

$$A\mathbf{x}_0, A^2\mathbf{x}_0, A^3\mathbf{x}_0, A^4\mathbf{x}_0, \dots, A^k\mathbf{x}_0, \dots$$

approaches a multiple of the dominant eigenvector of A .

If the eigen values are ordered such that

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|,$$

then the power method will converge quickly if $|\lambda_2|/|\lambda_1|$ is small, and slowly if $|\lambda_2|/|\lambda_1|$ is close to 1.

Convergence of Power Method

(a) The matrix

$$A = \begin{bmatrix} 4 & 5 \\ 6 & 5 \end{bmatrix}$$

has eigenvalues of $\lambda_1 = 10$ and $\lambda_2 = -1$. So the ratio $|\lambda_2|/|\lambda_1|$ is 0.1. For this matrix, only four iterations are required to obtain successive approximations that agree when rounded to three significant digits. (See Table 10.7.)

TABLE 10.7

\mathbf{x}_0	\mathbf{x}_1	\mathbf{x}_2	\mathbf{x}_3	\mathbf{x}_4
$\begin{bmatrix} 1.000 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.818 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.835 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.833 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.833 \\ 1.000 \end{bmatrix}$

Convergence of Power Method

(b) The matrix

$$A = \begin{bmatrix} -4 & 10 \\ 7 & 5 \end{bmatrix}$$

has eigenvalues of $\lambda_1 = 10$ and $\lambda_2 = -9$. For this matrix, the ratio $|\lambda_2|/|\lambda_1|$ is 0.9, and the power method does not produce successive approximations that agree to three significant digits until sixty-eight iterations have been performed, as shown in Table 10.8.

TABLE 10.8

\mathbf{x}_0	\mathbf{x}_1	\mathbf{x}_2	\dots	\mathbf{x}_{66}	\mathbf{x}_{67}	\mathbf{x}_{68}
$\begin{bmatrix} 1.000 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.500 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.941 \\ 1.000 \end{bmatrix}$	\dots	$\begin{bmatrix} 0.715 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.714 \\ 1.000 \end{bmatrix}$	$\begin{bmatrix} 0.714 \\ 1.000 \end{bmatrix}$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team

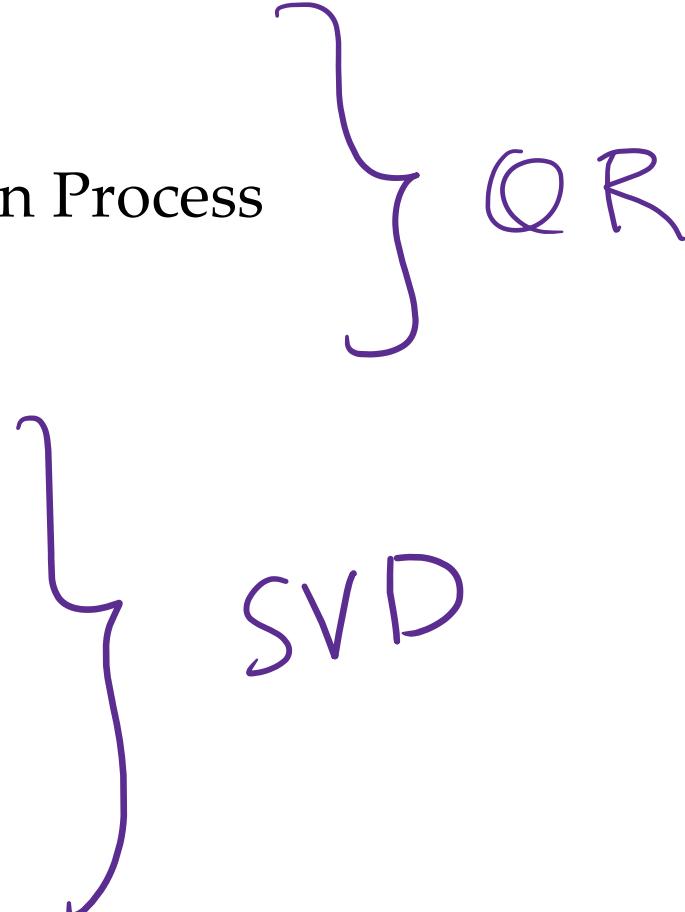


DSECL ZC416, MFDS

Lecture No. 6

Agenda

- Inner product spaces
- Gram Schmidt Orthogonalization Process
- QR decomposition
- Singular Value Decomposition
 - Derivation of SVD
 - Properties of SVD
 - Dimensionality Reduction



Inner Product Space

If u and v are vectors in R^n , then u and v are regarded as $n \times 1$ matrices. u^T is a $1 \times n$ matrix. $u^T v$ is called Inner product of u and v is denoted by $\langle u, v \rangle$

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

$$\begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = u_1 v_1 + u_2 v_2 + \cdots + u_n v_n$$

If u , v , and w are vectors in R^n and c is a scalar, then the following properties are true.

1. $u \cdot v = v \cdot u$
2. $u \cdot (v + w) = u \cdot v + u \cdot w$
3. $c(u \cdot v) = (cu) \cdot v = u \cdot (cv)$
4. $v \cdot v = \|v\|^2$
5. $v \cdot v \geq 0$, and $v \cdot v = 0$ if and only if $v = \mathbf{0}$.

A vector space V with an inner product is called Inner Product Space.

Whenever an inner product space is referred to, assume that set of scalars is the set of real numbers

Inner Product Space Example

An inner product on $M_{2 \times 2}$

Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ $B = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$ be matrices in the vector

space $M_{2 \times 2}$

Define $\langle A, B \rangle = a_{11}b_{11} + a_{21}b_{21} + a_{12}b_{12} + a_{22}b_{22}$

It is easy to verify that the operator is an inner product

For eg: $\langle A, A \rangle = a_{11}^2 + a_{12}^2 + a_{21}^2 + a_{22}^2$
 ≥ 0 always

$\langle A, A \rangle = 0 \rightarrow a_{ij} = 0 \quad \forall i, j$
 $\rightarrow A = 0$

Orthogonality

Orthogonal Set

Let V be vector space with an inner product

Non-zero vectors $v_1, v_2, \dots, v_k \in V$ form an **orthogonal set** if they are orthogonal to each other:

$$\langle v_i, v_j \rangle = 0 \text{ for } i \neq j$$

If, in addition, all vectors are of unit norm $\|v_i\| = 1$ then

v_1, v_2, \dots, v_k is called **orthonormal set**

Remark:

Any orthogonal set is linearly independent

Eg: $\{(1,1), (1,-1)\}$

$$\left(\begin{array}{c} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{array} \right), \left(\begin{array}{c} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{array} \right)$$

$$\begin{aligned} a) \langle u_i, u_j \rangle &= 0 & i \neq j \\ b) \langle u_i, u_i \rangle &= 1 \end{aligned}$$

Gram Schmidt Orthogonalization

Let V be a vector space with an inner product.

Suppose $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ is a basis for V . Let

$$\mathbf{v}_1 = \mathbf{x}_1,$$

$$\mathbf{v}_2 = \mathbf{x}_2 - \frac{\langle \mathbf{x}_2, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1,$$

$$\mathbf{v}_3 = \mathbf{x}_3 - \frac{\langle \mathbf{x}_3, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \frac{\langle \mathbf{x}_3, \mathbf{v}_2 \rangle}{\langle \mathbf{v}_2, \mathbf{v}_2 \rangle} \mathbf{v}_2,$$

.....

$$\mathbf{v}_n = \mathbf{x}_n - \frac{\langle \mathbf{x}_n, \mathbf{v}_1 \rangle}{\langle \mathbf{v}_1, \mathbf{v}_1 \rangle} \mathbf{v}_1 - \cdots - \frac{\langle \mathbf{x}_n, \mathbf{v}_{n-1} \rangle}{\langle \mathbf{v}_{n-1}, \mathbf{v}_{n-1} \rangle} \mathbf{v}_{n-1}.$$

Then $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ is an orthogonal basis for V .

Gram Schmidt Orthonormalization

Apply the Gram-Schmidt orthonormalization process to the basis for R^2 shown below.

$$B = \{\mathbf{v}_1, \mathbf{v}_2\}$$

$$B = \{(1, 1), (0, 1)\}$$

The Gram-Schmidt orthonormalization process produces

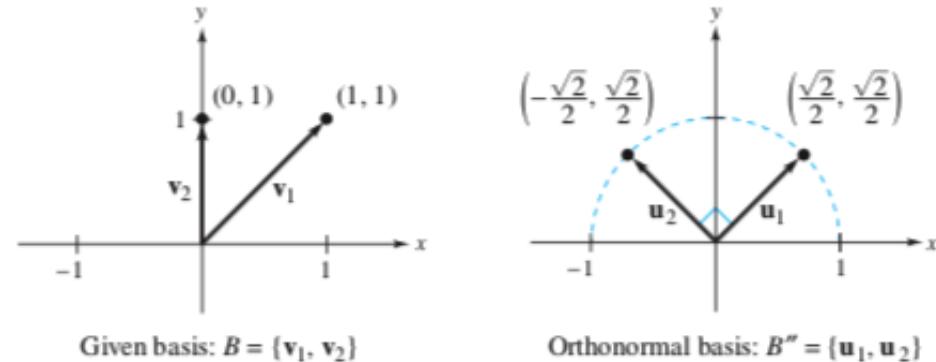
$$\mathbf{w}_1 = \mathbf{v}_1 = (1, 1)$$

$$\begin{aligned}\mathbf{w}_2 &= \mathbf{v}_2 - \frac{\mathbf{v}_2 \cdot \mathbf{w}_1}{\mathbf{w}_1 \cdot \mathbf{w}_1} \mathbf{w}_1 \\ &= (0, 1) - \frac{1}{2}(1, 1) = \left(-\frac{1}{2}, \frac{1}{2}\right).\end{aligned}$$

The set $B' = \{\mathbf{w}_1, \mathbf{w}_2\}$ is an orthogonal basis for R^2 . By normalizing each vector in B' , you obtain

$$\mathbf{u}_1 = \frac{\mathbf{w}_1}{\|\mathbf{w}_1\|} = \frac{1}{\sqrt{2}}(1, 1) = \frac{\sqrt{2}}{2}(1, 1) = \left(\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right)$$

$$\mathbf{u}_2 = \frac{\mathbf{w}_2}{\|\mathbf{w}_2\|} = \frac{1}{1/\sqrt{2}}\left(-\frac{1}{2}, \frac{1}{2}\right) = \sqrt{2}\left(-\frac{1}{2}, \frac{1}{2}\right) = \left(-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\right).$$



So, $B'' = \{\mathbf{u}_1, \mathbf{u}_2\}$ is an orthonormal basis for R^2 .

QR Factorization

If A is an $m \times n$ matrix with linearly independent columns, then A can be factored as $\mathbf{A} = \mathbf{Q}\mathbf{R}$ where \mathbf{Q} is an $m \times n$ matrix whose columns form an orthonormal basis for $\text{Col } A$ and \mathbf{R} is an $n \times n$ **upper triangular invertible matrix** with positive entries on its diagonal

$$A_{m \times n} = Q_{m \times n} R_{n \times n}$$

$$A = [q_1 \quad q_2 \quad \cdots \quad q_n] \left[\begin{array}{cccc} R_{11} & R_{12} & \cdots & R_{1n} \\ 0 & R_{22} & \cdots & R_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ddots & R_{nn} \end{array} \right]$$




vectors q_1, \dots, q_n are orthonormal m -vectors:

$$\|q_i\| = 1, \quad q_i^T q_j = 0 \quad \text{if } i \neq j$$

QR Factorization Example

Find a QR factorization of $A =$

$$A_{mxn} = Q_{mxn} \cdot R_{nxn}$$

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

1. First verify that the columns of A are LI

1. Gram Schmidt Orthogonalization to the columns of A

$\langle v_1, v_2, \dots, v_n \rangle$ such that $\langle v_i, v_j \rangle = 0$ for all $i \neq j$

3. Normalize v_i to get $u_i \rightarrow$ columns of Q

4. Use Q to get R

QR Factorization Example

$$Q = \begin{bmatrix} u_1 & u_2 & u_3 \end{bmatrix} = \begin{bmatrix} 1/2 & -3/\sqrt{12} & 0 \\ 1/2 & 1/\sqrt{12} & -2/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \\ 1/2 & 1/\sqrt{12} & 1/\sqrt{6} \end{bmatrix}$$

To find R, observe that $Q^T Q = I$ because columns of Q are orthonormal

$$Q^T A = Q^T (QR) = IR = R$$

$$R = \begin{bmatrix} 1/2 & 1/2 & 1/2 & 1/2 \\ -3/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} & 1/\sqrt{12} \\ 0 & -2/\sqrt{6} & 1/\sqrt{6} & 1/\sqrt{6} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 2 & 3/2 & 1 \\ 0 & 3/\sqrt{12} & 2/\sqrt{12} \\ 0 & 0 & 2/\sqrt{6} \end{bmatrix}$$

Singular Value Decomposition

Singular Value Decomposition(SVD) is a factorization of an $m \times n$ matrix into

U is an $m \times m$ orthogonal matrix (Its columns are Left Singular Vectors)

Σ is an $m \times n$ diagonal matrix with **singular values** on the diagonal

$$\Sigma = \begin{pmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & 0 \end{pmatrix}$$

Convention: $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_n \geq 0$

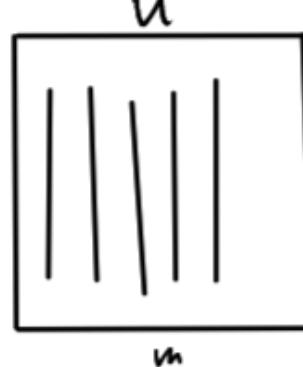
V^T is an $n \times n$ orthogonal matrix (V 's columns are called Right Singular Vectors) such that $A = U\Sigma V^T$

Singular Value Decomposition

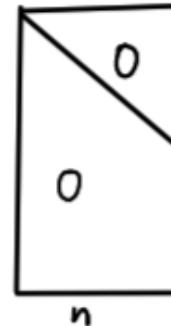
$m > n$



$=$



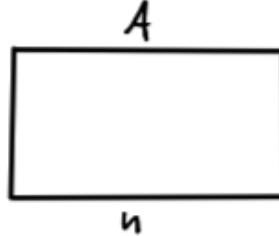
Σ



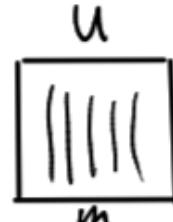
V^T



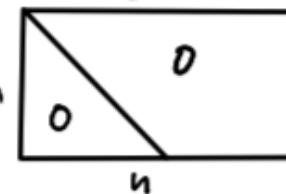
$m < n$



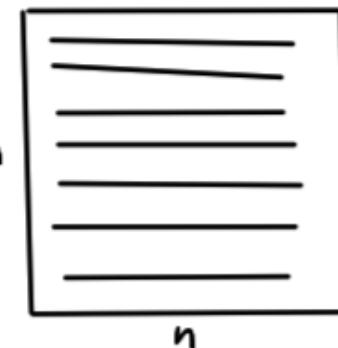
$=$



Σ



V^T



Evaluation of U and V

1. Find an orthogonal diagonalization of $A^T A$
 - Find the eigenvalues of $A^T A$ and corresponding orthonormal set of eigenvectors
 2. Set up V and Σ
 - Arrange the eigenvalues of $A^T A$ in decreasing order and compute the square roots of the eigen values. Σ will be same size as A with D (diagonal entries are non zero singular values) in upper left corner and with 0's elsewhere
 3. Derive U for $A = U \Sigma V^T$
-

Evaluation of U and V - Example

Construct singular value decomposition of $A = \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix}$

1. Find eigenvalues of $A^T A$

$$A^T A = \begin{bmatrix} 4 & 8 \\ 11 & 7 \\ 14 & -2 \end{bmatrix} \begin{bmatrix} 4 & 11 & 14 \\ 8 & 7 & -2 \end{bmatrix} = \begin{bmatrix} 80 & 100 & 40 \\ 100 & 170 & 140 \\ 40 & 140 & 200 \end{bmatrix}$$

Eigenvalues of $A^T A$ are $\lambda_1 = 360$, $\lambda_2 = 90$ and $\lambda_3 = 0$

2. Set up V and Σ

$$V = [\mathbf{v}_1 \ \mathbf{v}_2 \ \mathbf{v}_3] = \begin{bmatrix} 1/3 & -2/3 & 2/3 \\ 2/3 & -1/3 & -2/3 \\ 2/3 & 2/3 & 1/3 \end{bmatrix} \rightarrow \text{corresponding Unit eigen vectors}$$

Evaluation of U and V - Example

3. Set up Σ

The square roots of the eigen values are singular values

$$\sigma_1 = 6\sqrt{10}, \sigma_2 = 3\sqrt{10}, \sigma_3 = 0$$

$$D = \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix}, \quad \Sigma = [D \ 0] = \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix}$$

4. Construct U

$$u_1 = \frac{1}{\sigma_1} A v_1 = \frac{1}{6\sqrt{10}} \begin{bmatrix} 18 \\ 6 \end{bmatrix} = \begin{bmatrix} 3/\sqrt{10} \\ 1/\sqrt{10} \end{bmatrix}$$

$$u_2 = \frac{1}{\sigma_2} A v_2 = \frac{1}{3\sqrt{10}} \begin{bmatrix} 3 \\ -9 \end{bmatrix} = \begin{bmatrix} 1/\sqrt{10} \\ -3/\sqrt{10} \end{bmatrix}$$

Singular value decomposition of A is

$$A = \begin{bmatrix} 3/\sqrt{10} & 1/\sqrt{10} \\ 1/\sqrt{10} & -3/\sqrt{10} \end{bmatrix} \begin{bmatrix} 6\sqrt{10} & 0 & 0 \\ 0 & 3\sqrt{10} & 0 \end{bmatrix} \begin{bmatrix} 1/3 & 2/3 & 2/3 \\ -2/3 & -1/3 & 2/3 \\ 2/3 & -2/3 & 1/3 \end{bmatrix}$$

↑ U ↑ Σ ↑ V^T ■

Comparison between Eigenvalue decomposition and SVD



Eigenvalue Decomposition	Singular Value Decomposition
1. Works only for square matrices	Works for matrix of any size
2. Non diagonal matrices P and P^{-1} are inverses of each other	Non diagonal matrices U and V are not necessarily inverses of each other
3. Entries of D can be any complex number – negative, positive , imaginary	Entries in the diagonal matrix Σ are real and non negative, singular values are decreasing.
4. Vectors in eigenvalue decomposition matrix P are not necessarily orthogonal	Matrices U and V are orthonormal

Left and Right Singular Vectors

- The column vectors in V are called right singular vectors
- The column vectors in U are called left singular vectors

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$$

$$U = \begin{bmatrix} -0.2425 & -0.9701 \\ -0.9701 & 0.2425 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.4472 & 0.8944 \\ -0.8944 & -0.4472 \end{bmatrix}$$

u_1, u_2 are left singular vectors

v_1, v_2 are right singular vectors

Summation form of SVD

Let A be an m by n matrix of rank r. Let $A = U\Sigma V^T$

Then A can be expanded as

$$A = \sigma_1(u_1 v_1^T) + \sigma_2(u_2 v_2^T) + \dots + \sigma_r(u_r v_r^T)$$

Here $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r$

It is addition of rank 1 matrices

Summation Formula Example

Consider the following matrix

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$$

$$\text{Rref}(A) = \begin{bmatrix} 1 & 2 \\ 0 & 0 \end{bmatrix}$$

Rank = 1

Eigendecomposition of AA^T

$$AA^T = \begin{bmatrix} 5 & 20 \\ 20 & 80 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} -0.2425 & -0.9701 \\ -0.9701 & 0.2425 \end{bmatrix}$$

$$D_1 = \begin{bmatrix} 85 & 0 \\ 0 & 0 \end{bmatrix}$$

Eigendecomposition of $A^T A$

$$A^T A = \begin{bmatrix} 17 & 34 \\ 34 & 68 \end{bmatrix}$$

$$P_2 = \begin{bmatrix} 0.4472 & -0.8944 \\ 0.8944 & 0.4472 \end{bmatrix}$$

$$D_2 = \begin{bmatrix} 85 & 0 \\ 0 & 0 \end{bmatrix}$$

Summation Formula Example

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 8 \end{bmatrix}$$

Rank = 1

$$U = \begin{bmatrix} -0.2425 & -0.9701 \\ -0.9701 & 0.2425 \end{bmatrix}$$

$$S = \begin{bmatrix} 9.2195 & 0 \\ 0 & 0 \end{bmatrix}$$

$$V = \begin{bmatrix} -0.4472 & 0.8944 \\ -0.8944 & -0.4472 \end{bmatrix}$$

$$A = \sigma_1 u_1 v_1^T$$

$$A = 9.2195 \begin{bmatrix} -0.2425 \\ -0.9701 \end{bmatrix} \begin{bmatrix} -0.4472 & -0.8944 \end{bmatrix}$$

Singular Values of A

It may be observed that the 2 singular values of A are the square root of eigenvalues of AA^T or A^TA

- $\sigma_1 = \sqrt{\lambda_1} = 9.2195 = \sqrt{85}$
- $\sigma_2 = \sqrt{\lambda_2} = 0 = \sqrt{0}$

Applications of SVD

1. Dimension Reduction

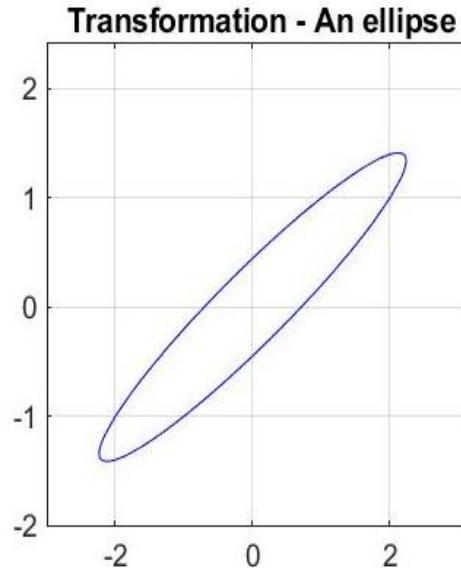
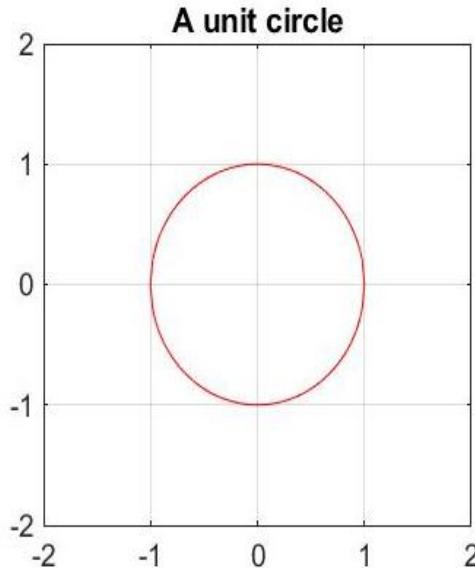
σ_i s dictate the dimension which are relevant

Suppose we consider only $\sigma_1, \sigma_2, \dots, \sigma_k$, $k < n$, then dimension is reduced to k from n.

2. Pseudo Inverse

3. Signal and Image analysis

Transformation of Circle under Matrix Operation



$Av_i = \sigma_i u_i$ where A is given matrix, v_i is orthogonal set
 σ_i are singular values
 u_i are principal axes direction

In higher dimensions

$$Av_1 = \sigma_1 u_1$$

$$Av_2 = \sigma_2 u_2$$

.

$$Av_n = \sigma_n u_n$$

Reduced SVD

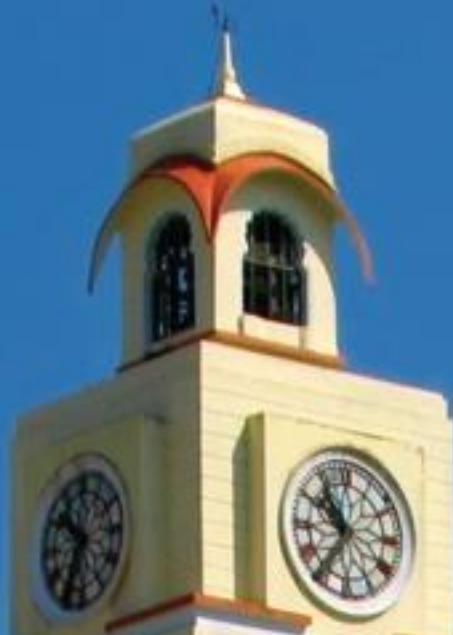
$$Av_j = \sigma_j u_j \quad 1 \leq j \leq n$$

$$A \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix},$$

$$A = \hat{U} \hat{\Sigma} V^T.$$

Reduced SVD ($m \geq n$)

$$\begin{array}{c}
 \boxed{} \\
 A \\
 m \times n
 \end{array}
 =
 \begin{array}{c}
 \boxed{} \\
 \hat{U} \\
 m \times n
 \end{array}
 \begin{array}{c}
 \boxed{\text{diagonal}} \\
 \hat{\Sigma} \\
 n \times n
 \end{array}
 \begin{array}{c}
 \boxed{} \\
 V^T \\
 n \times n
 \end{array}$$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 7

Agenda

- Reddy Mikks Problem
 - Formulation and graphical method
 - TORA
 - Urban Renewal Model
 - Formulation and feasibility
 - Currency Arbitrage Model
 - Formulation and applicability
-



Problem 1: Reddy Mikks Company

Example 2.1-1 (The Reddy Mikks Company)

Reddy Mikks produces both interior and exterior paints from two raw materials, M_1 and M_2 . The following table provides the basic data of the problem:

	Tons of raw material per ton of		Maximum daily availability (tons)
	<i>Exterior paint</i>	<i>Interior paint</i>	
Raw material, M_1	6	4	24
Raw material, M_2	1	2	6
Profit per ton (\$1000)	5	4	

A market survey indicates that the daily demand for interior paint cannot exceed that for exterior paint by more than 1 ton. Also, the maximum daily demand for interior paint is 2 tons. Reddy Mikks wants to determine the optimum (best) product mix of interior and exterior paints that maximizes the total daily profit.



For the Reddy Mikks problem, we need to determine the daily amounts to be produced of exterior and interior paints. Thus the variables of the model are defined as

x_1 = Tons produced daily of exterior paint

x_2 = Tons produced daily of interior paint

To construct the objective function, note that the company wants to *maximize* (i.e., increase as much as possible) the total daily profit of both paints. Given that the profits per ton of exterior and interior paints are 5 and 4 (thousand) dollars, respectively, it follows that

Total profit from exterior paint = $5x_1$ (thousand) dollars

Total profit from interior paint = $4x_2$ (thousand) dollars

Letting z represent the total daily profit (in thousands of dollars), the objective of the company is

$$\text{Maximize } z = 5x_1 + 4x_2$$

Next, we construct the constraints that restrict raw material usage and product demand. The raw material restrictions are expressed verbally as

$$\left(\begin{array}{l} \text{Usage of a raw material} \\ \text{by both paints} \end{array} \right) \leq \left(\begin{array}{l} \text{Maximum raw material} \\ \text{availability} \end{array} \right)$$

The daily usage of raw material $M1$ is 6 tons per ton of exterior paint and 4 tons per ton of interior paint. Thus

Usage of raw material $M1$ by exterior paint = $6x_1$ tons/day

Usage of raw material $M1$ by interior paint = $4x_2$ tons/day

Hence

Usage of raw material $M1$ by both paints = $6x_1 + 4x_2$ tons/day

In a similar manner,

Usage of raw material $M2$ by both paints = $1x_1 + 2x_2$ tons/day



Because the daily availabilities of raw materials $M1$ and $M2$ are limited to 24 and 6 tons, respectively, the associated restrictions are given as

$$6x_1 + 4x_2 \leq 24 \quad (\text{Raw material } M1)$$

$$x_1 + 2x_2 \leq 6 \quad (\text{Raw material } M2)$$

The first demand restriction stipulates that the excess of the daily production of interior over exterior paint, $x_2 - x_1$, should not exceed 1 ton, which translates to

$$x_2 - x_1 \leq 1 \quad (\text{Market limit})$$

The second demand restriction stipulates that the maximum daily demand of interior paint is limited to 2 tons, which translates to

$$x_2 \leq 2 \quad (\text{Demand limit})$$

An implicit (or “understood-to-be”) restriction is that variables x_1 and x_2 cannot assume negative values. The **nonnegativity restrictions**, $x_1 \geq 0$, $x_2 \geq 0$, account for this requirement.

The complete Reddy Mikks model is

$$\text{Maximize } z = 5x_1 + 4x_2$$

subject to

$$6x_1 + 4x_2 \leq 24 \tag{1}$$

$$x_1 + 2x_2 \leq 6 \tag{2}$$

$$-x_1 + x_2 \leq 1 \tag{3}$$

$$x_2 \leq 2 \tag{4}$$

$$x_1, x_2 \geq 0 \tag{5}$$



Problem 2: Urban Renewal Model

Example 2.3-1 (Urban Renewal Model)

The city of Erstville is faced with a severe budget shortage. Seeking a long-term solution, the city council votes to improve the tax base by condemning an inner-city housing area and replacing it with a modern development.

The project involves two phases: (1) demolishing substandard houses to provide land for the new development, and (2) building the new development. The following is a summary of the situation.

1. As many as 300 substandard houses can be demolished. Each house occupies a .25-acre lot. The cost of demolishing a condemned house is \$2000.
2. Lot sizes for new single-, double-, triple-, and quadruple-family homes (units) are .18, .28, .4, and .5 acre, respectively. Streets, open space, and utility easements account for 15% of available acreage.
3. In the new development the triple and quadruple units account for at least 25% of the total. Single units must be at least 20% of all units and double units at least 10%.
4. The tax levied per unit for single, double, triple, and quadruple units is \$1,000, \$1,900, \$2,700, and \$3,400, respectively.
5. The construction cost per unit for single-, double-, triple-, and quadruple- family homes is \$50,000, \$70,000, \$130,000, and \$160,000, respectively. Financing through a local bank can amount to a maximum of \$15 million.

How many units of each type should be constructed to maximize tax collection?



Problem 2: Urban Renewal Model

Step 1: Define **Variables** of the problem

x_1 = Number of units of single-family homes

x_2 = Number of units of double-family homes

x_3 = Number of units of triple-family homes

x_4 = Number of units of quadruple-family homes

x_5 = Number of old homes to be demolished

Step 2: Define **Objective** that we need to optimize

The objective is to maximize total tax collection from all four types of homes—that is,

$$\text{Maximize } z = 1000x_1 + 1900x_2 + 2700x_3 + 3400x_4$$



Problem 2: Urban Renewal Model

Lot sizes for new single-, double-, triple-, and quadruple-family homes (units) are .18, .28, .4, and .5 acre, respectively. Streets, open space, and utility easements account for 15% of available acreage.

The first constraint of the problem deals with land availability.

$$\left(\begin{array}{l} \text{Acreage used for new} \\ \text{home construction} \end{array} \right) \leq \left(\begin{array}{l} \text{Net available} \\ \text{acreage} \end{array} \right)$$

From the data of the problem we have

$$\text{Acreage needed for new homes} = .18x_1 + .28x_2 + .4x_3 + .5x_4$$

To determine the available acreage, each demolished home occupies a .25-acre lot, thus netting $.25x_5$ acres. Allowing for 15% open space, streets, and easements, the net acreage available is $.85(.25x_5) = .2125x_5$. The resulting constraint is

$$.18x_1 + .28x_2 + .4x_3 + .5x_4 \leq .2125x_5$$

or

$$.18x_1 + .28x_2 + .4x_3 + .5x_4 - .2125x_5 \leq 0$$



Problem 2: Urban Renewal Model

As many as 300 substandard houses can be demolished.

$$x_5 \leq 300$$

In the new development the triple and quadruple units account for at least 25% of the total. Single units must be at least 20% of all units and double units at least 10%.

(Number of single units) \geq (20% of all units)

(Number of double units) \geq (10% of all units)

(Number of triple and quadruple units) \geq (25% of all units)

These constraints translate mathematically to

$$x_1 \geq .2(x_1 + x_2 + x_3 + x_4)$$

$$x_2 \geq .1(x_1 + x_2 + x_3 + x_4)$$

$$x_3 + x_4 \geq .25(x_1 + x_2 + x_3 + x_4)$$



Problem 2: Urban Renewal Model

The only remaining constraint deals with keeping the demolition/construction cost within the allowable budget—that is,

$$(\text{Construction and demolition cost}) \leq (\text{Available budget})$$

Expressing all the costs in thousands of dollars, we get

$$(50x_1 + 70x_2 + 130x_3 + 160x_4) + 2x_5 \leq 15000$$



Problem 2: Urban Renewal Model

The complete model thus becomes

$$\text{Maximize } z = 1000x_1 + 1900x_2 + 2700x_3 + 3400x_4$$

subject to

$$.18x_1 + .28x_2 + .4x_3 + .5x_4 - .2125x_5 \leq 0$$

$$x_5 \leq 300$$

$$-.8x_1 + .2x_2 + .2x_3 + .2x_4 \leq 0$$

$$.1x_1 - .9x_2 + .1x_3 + .1x_4 \leq 0$$

$$.25x_1 + .25x_2 - .75x_3 - .75x_4 \leq 0$$

$$50x_1 + 70x_2 + 130x_3 + 160x_4 + 2x_5 \leq 15000$$

$$x_1, x_2, x_3, x_4, x_5 \geq 0$$



Problem 2: Urban Renewal Model

Total tax collection = $z = \$343,965$

Number of single homes = $x_1 = 35.83 \approx 36$ units

Number of double homes = $x_2 = 98.53 \approx 99$ units

Number of triple homes = $x_3 = 44.79 \approx 45$ units

Number of quadruple homes = $x_4 = 0$ units

Number of homes demolished = $x_5 = 244.49 \approx 245$ units

However, the feasible solution is

$x_1 = 36, x_2 = 98, x_3 = 45, x_4 = 0$ and $x_5 = 245$



Problem 3: Currency Arbitrage Model

Example 2.3-2 (Currency Arbitrage Model)

Suppose that a company has a total of 5 million dollars that can be exchanged for euros (€), British pounds (£), yen (¥), and Kuwaiti dinars (KD). Currency dealers set the following limits on the amount of any single transaction: 5 million dollars, 3 million euros, 3.5 million pounds, 100 million yen, and 2.8 million KDs. The table below provides typical spot exchange rates. The bottom diagonal rates are the reciprocal of the top diagonal rates. For example, $\text{rate}(\text{€} \rightarrow \$) = 1/\text{rate}(\$ \rightarrow \text{€}) = 1/.769 = 1.30$.

	\$	€	£	¥	KD
\$	1	.769	.625	105	.342
€	$\frac{1}{.769}$	1	.813	137	.445
£	$\frac{1}{.625}$	$\frac{1}{.813}$	1	169	.543
¥	$\frac{1}{105}$	$\frac{1}{137}$	$\frac{1}{169}$	1	.0032
KD	$\frac{1}{.342}$	$\frac{1}{.445}$	$\frac{1}{.543}$	$\frac{1}{.0032}$	1

Is it possible to increase the dollar holdings (above the initial \$5 million) by circulating currencies through the currency market?



Problem 3: Currency Arbitrage Model

Currency	\$	€	£	¥	KD
Code	1	2	3	4	5

Define

x_{ij} = Amount in currency i converted to currency j , i and $j = 1, 2, \dots, 5$

For example, x_{12} is the dollar amount converted to euros and x_{51} is the KD amount converted to dollars. We further define two additional variables representing the input and the output of the arbitrage problem:

I = Initial dollar amount (= \$5 million)

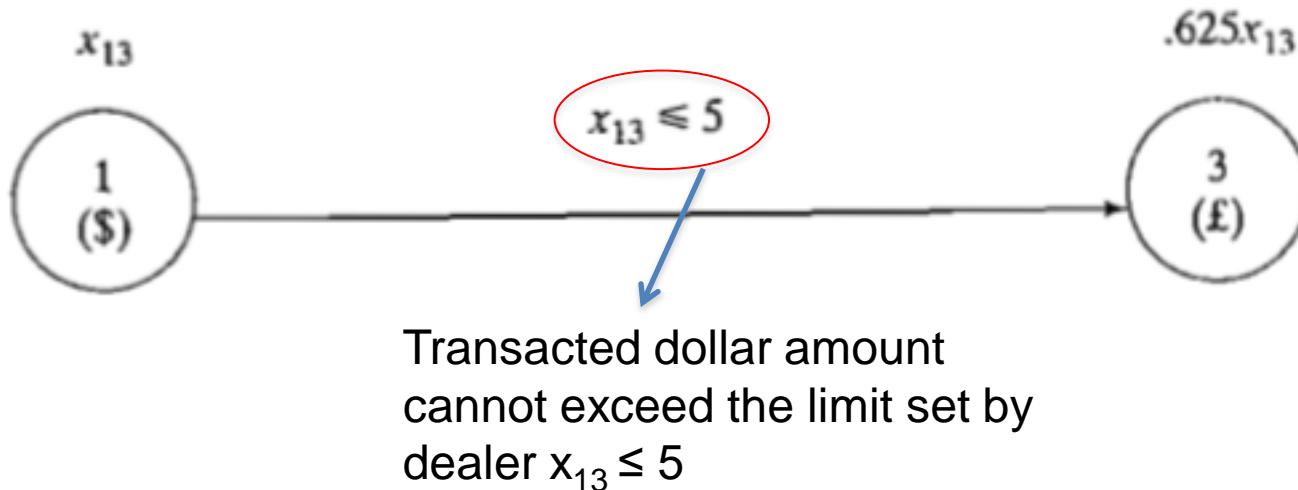
y = Final dollar holdings (to be determined from the solution)



Problem 3: Currency Arbitrage Model

Our goal is to determine the maximum final dollar holdings, y , subject to the currency flow restrictions and the maximum limits allowed for the different transactions.

$$\text{Maximize } z = y$$





To conserve the flow of money from one currency to another, each currency must satisfy the following input-output equation:

$$\begin{pmatrix} \text{Total sum available} \\ \text{of a currency (input)} \end{pmatrix} = \begin{pmatrix} \text{Total sum converted to} \\ \text{other currencies (output)} \end{pmatrix}$$

1. Dollar ($i = 1$):

$$\begin{aligned}\text{Total available dollars} &= \text{Initial dollar amount} + \\ &\quad \text{dollar amount from other currencies} \\ &= I + (\epsilon \rightarrow \$) + (\text{£} \rightarrow \$) + (\text{¥} \rightarrow \$) + (\text{KD} \rightarrow \$) \\ &= I + \frac{1}{.769}x_{21} + \frac{1}{.625}x_{31} + \frac{1}{105}x_{41} + \frac{1}{.342}x_{51}\end{aligned}$$

$$\begin{aligned}\text{Total distributed dollars} &= \text{Final dollar holdings} + \\ &\quad \text{dollar amount to other currencies} \\ &= y + (\$ \rightarrow \epsilon) + (\$ \rightarrow \text{£}) + (\$ \rightarrow \text{¥}) + (\$ \rightarrow \text{KD}) \\ &= y + x_{12} + x_{13} + x_{14} + x_{15}\end{aligned}$$

Given $I = 5$, the dollar constraint thus becomes

$$y + x_{12} + x_{13} + x_{14} + x_{15} - \left(\frac{1}{.769}x_{21} + \frac{1}{.625}x_{31} + \frac{1}{105}x_{41} + \frac{1}{.342}x_{51} \right) = 5$$



2. Euro ($i = 2$):

$$\begin{aligned}\text{Total available euros} &= (\$ \rightarrow \epsilon) + (\text{£} \rightarrow \epsilon) + (\text{¥} \rightarrow \epsilon) + (\text{KD} \rightarrow \epsilon) \\ &= .769x_{12} + \frac{1}{813}x_{32} + \frac{1}{137}x_{42} + \frac{1}{445}x_{52}\end{aligned}$$

$$\begin{aligned}\text{Total distributed euros} &= (\epsilon \rightarrow \$) + (\epsilon \rightarrow \text{£}) + (\epsilon \rightarrow \text{¥}) + (\epsilon \rightarrow \text{KD}) \\ &= x_{21} + x_{23} + x_{24} + x_{25}\end{aligned}$$

Thus, the constraint is

$$x_{21} + x_{23} + x_{24} + x_{25} - (.769x_{12} + \frac{1}{813}x_{32} + \frac{1}{137}x_{42} + \frac{1}{445}x_{52}) = 0$$



Transaction Limit

British pounds (£), yen (¥), and Kuwaiti dinars (KD). Currency dealers set the following limits on the amount of any single transaction: 5 million dollars, 3 million euros, 3.5 million pounds, 100 million yen, and 2.8 million KDs. The table below provides typical spot exchange rates. The

The only remaining constraints are the transaction limits, which are 5 million dollars, 3 million euros, 3.5 million pounds, 100 million yen, and 2.8 million KDs. These can be translated as

$$x_{1j} \leq 5, j = 2, 3, 4, 5$$

$$x_{2j} \leq 3, j = 1, 3, 4, 5$$

$$x_{3j} \leq 3.5, j = 1, 2, 4, 5$$

$$x_{4j} \leq 100, j = 1, 2, 3, 5$$

$$x_{5j} \leq 2.8, j = 1, 2, 3, 4$$



Problem 3: Currency Arbitrage Model

The complete model is now given as

$$\text{Maximize } z = y$$

subject to

$$y + x_{12} + x_{13} + x_{14} + x_{15} - \left(\frac{1}{.769}x_{21} + \frac{1}{.625}x_{31} + \frac{1}{105}x_{41} + \frac{1}{.342}x_{51} \right) = 5$$

$$x_{21} + x_{23} + x_{24} + x_{25} - \left(.769x_{12} + \frac{1}{.813}x_{32} + \frac{1}{137}x_{42} + \frac{1}{.445}x_{52} \right) = 0$$

$$x_{31} + x_{32} + x_{34} + x_{35} - \left(.625x_{13} + .813x_{23} + \frac{1}{169}x_{43} + \frac{1}{.543}x_{53} \right) = 0$$

$$x_{41} + x_{42} + x_{43} + x_{45} - \left(105x_{14} + 137x_{24} + 169x_{34} + \frac{1}{.0032}x_{54} \right) = 0$$

$$x_{51} + x_{52} + x_{53} + x_{54} - \left(.342x_{15} + .445x_{25} + .543x_{35} + .0032x_{45} \right) = 0$$

$$x_{1j} \leq 5, j = 2, 3, 4, 5$$

$$x_{2j} \leq 3, j = 1, 3, 4, 5$$

$$x_{3j} \leq 3.5, j = 1, 2, 4, 5$$

$$x_{4j} \leq 100, j = 1, 2, 3, 5$$

$$x_{5j} \leq 2.8, j = 1, 2, 3, 4$$

$$x_{ij} \geq 0, \text{ for all } i \text{ and } j$$



Problem 3: Currency Arbitrage Model

Solution	Interpretation
$y = 5.09032$	Final holdings = \$5,090,320. Net dollar gain = \$90,320, which represents a 1.8064% rate of return
$x_{12} = 1.46206$	Buy \$1,462,060 worth of euros
$x_{15} = 5$	Buy \$5,000,000 worth of KD
$x_{25} = 3$	Buy €3,000,000 worth of KD
$x_{31} = 3.5$	Buy £3,500,000 worth of dollars
$x_{32} = 0.931495$	Buy £931,495 worth of euros
$x_{41} = 100$	Buy ¥100,000,000 worth of dollars
$x_{42} = 100$	Buy ¥100,000,000 worth of euros
$x_{43} = 100$	Buy ¥100,000,000 worth of pounds
$x_{53} = 2.085$	Buy KD2,085,000 worth of pounds
$x_{54} = .96$	Buy KD960,000 worth of yen



Problem 3: Currency Arbitrage Model

At first it may appear for solution to be nonsensical as it calls for using $x_{12} + x_{15} = 1.46206 + 5 = 6.46206 = \$6,462,060$ to buy Euros or KD but initial dollar amount is only \$5 million
In practice the given solution is submitted to the currency dealer as one order, we do not wait until we accumulate enough currency of certain type before making a buy.

$$\begin{aligned}I &= y + x_{12} + x_{13} + x_{14} + x_{15} - \left(\frac{1}{.769}x_{21} + \frac{1}{.625}x_{31} + \frac{1}{105}x_{41} + \frac{1}{342}x_{51} \right) \\&= 5.09032 + 1.46206 + 5 - \left(\frac{35}{625} + \frac{100}{105} \right) = 5\end{aligned}$$

Sensitivity Analysis

- Sensitivity Analysis (restricted to graphical solutions)
 - Changes in right hand side
 - One or more changes is possible
 - Changes in objective coefficients
 - One or more changes possible
 - Complicated changes require concepts in Duality
 - Not in the present scope

Changes in RHS(Example 3.6-1)

JOBCO produces two products on two machines. A unit of product 1 requires 2 hours on machine 1 and 1 hour on machine 2. For product 2, a unit requires 1 hour on machine 1 and 3 hours on machine 2. The revenues per unit of products 1 and 2 are \$30 and \$20 respectively. The total daily processing time available for each machine is 8 hours.

LPP

Maximize $Z = 30x_1 + 20x_2$

Subject to $2x_1 + x_2 \leq 8$ (Machine 1)

$x_1 + 3x_2 \leq 8$ (Machine 2)

$x_1, x_2 \geq 0$

Sensitivity Analysis 1 : Changes in Objective Coefficients

1. Suppose that the unit revenues for products 1 and 2 are changed to \$35 and \$25 respectively, will the current optimum remain the same?

Ans. Yes, however optimal value changes to \$152.

2. Suppose that the unit revenue of product 2 is fixed at its current value of $c_2 = \$20$, what is the associated range for c_1 , that will keep the optimum unchanged.

Ans. $6.67 \leq c_1 \leq 40$

Changes in RHS

	Machine 1	Machine 2
Shadow prce	\$ 14/ hour	\$ 2 /hour
Minimum capacity	2.67 hours	4 hrs
Maximum capacity	16 hours	24 hours

Questions

1. If JOBCO can increase the capacity of both machines, which machine should receive higher priority?

Ans Machine 1

2. A suggestion is made to increase the capacities of machines 1 and 2 at the additional cost of \$10 / hour. Is this advisable?

Ans Only machine 1 should be increased.

3. If the capacity of machine 1 is increased from the present 8 hours to 13 hours, how will the increase impact the optimum revenue?

Ans Increased to \$198.

198

4. Suppose the capacity of machine 1 is increased to 20 hours, how will this increase impact the optimum revenue?

Ans We do not have any conclusion

$$2.67 \text{ hrs} \leq \text{rhs}_{M_1} \leq 16 \text{ hrs}$$

Standard LPP

Max z = $c^T x$
 subject to $Ax = b$, (with $b \geq 0$)
 $x \geq 0$

$$\begin{aligned}
 & \max \quad 2x_1 + 3x_2 \\
 \text{st} \quad & 3x_1 + 4x_2 = 3 \\
 & 5x_1 + 7x_2 = 5 \\
 & x_1 \geq 0 \\
 & x_2 \geq 0
 \end{aligned}$$

To be observed:

- Constraints have the equality
- Non-negativity of b
- Non-negativity of variables

Transformations

1. Standardize the Minimization to Maximization
 2. Standardize the Negative components of \mathbf{b}
 3. Standardize the Constraints with \leq
.
 4. Standardize the Constraints with \geq
 5. Standardize the Variables unrestricted in sign
-



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 8

Agenda

- LPP in Standard Form
- FS, BS, BFS, OBFS
- Motivation for the Simplex Method
- Simplex tables(using excel)
- Sensitivity

Standard LPP

$\text{Max } Z = \mathbf{c}\mathbf{x}$

subject to

$\mathbf{A}\mathbf{x} = \mathbf{b}, \text{ with } \mathbf{b} \geq \mathbf{0};$

$\mathbf{x} \geq \mathbf{0};$

To be observed:

- Constraints have the equality
- Non-negativity of b
- Non-negativity of variables

Transformations

- Minimization to Maximization
- Negative components of \mathbf{b}
- Constraints with \leq
- Constraints with \geq
- Variables unrestricted in sign

Terminologies

Given m equations in n unknowns ($n \geq m$) in standard form

1. Set $(n-m)$ variables to zero and determine m unique values
 2. the $(n-m)$ are called non-basic variables
 3. the m variables are called basic variables
 4. a solution that satisfies the constraints and non-negativity –feasible
 5. a solution that is basic and satisfies (4) is – basic feasible solution (bfs)
 6. All basic variables are > 0 – non-degenerate bfs
 7. A bfs that maximizes the objective function is optimal solution
-

Naive LPP Solvers vs Simplex

- The optimal solution, if it exists, is a corner point
 - Fundamental theorem of linear programming
- Have nC_m ways of finding the corner point ($n \geq m$)
- Simplex takes just a fraction of the above
- Simplex is iterative and simple to interpret

Principles of Simplex Method

- Start with an initial *basic* feasible solution
- Improve the initial solution, if possible
- Stop, when the bfs cannot be improved

Example: Maximize $Z = 5x_1 + 2x_2 + 3x_3 - x_4 + x_5$

Subject to $x_1 + 2x_2 + 2x_3 + x_4 = 8$

$3x_1 + 4x_2 + x_3 + x_5 = 7$

$x_i \geq 0$ for all $i=1,2,3,4,5$

Reddy Mikks Problem

Maximize $Z = 5x_1 + 4x_2 + 0s_1 + 0s_2 + 0s_3 + 0s_4$

Subject to $6x_1 + 4x_2 + s_1 = 24$

$x_1 + 2x_2 + s_2 = 6$

$-x_1 + x_2 + s_3 = 1$

$x_2 + s_4 = 2$

$x_1, x_2, s_1, s_2, s_3, s_4 \geq 0$

(Refer to excel sheet for computational aspects)

Special cases

1. Unique optimal value
 - $c_j - z_j < 0$ for all non-basic variables
2. Alternative optima
 - At least one non-basic variable has $c_j - z_j = 0$
3. Unboundedness
 - The pivot column has entries which are all ≤ 0
4. Infeasibility (not in our scope)
 - Artificial variable is there in the final table and is > 0

Sensitivity Analysis

- Sensitivity Analysis (restricted to graphical solutions)
 - Changes in right hand side
 - One or more changes is possible
 - Changes in objective coefficients
 - One or more changes possible
 - Complicated changes require concepts in Duality
 - Not in the present scope

Changes in RHS(Example 3.6-1)

JOBCO produces two products on two machines. A unit of product 1 requires 2 hours on machine 1 and 1 hour on machine 2. For product 2, a unit requires 1 hour on machine 1 and 3 hours on machine 2. The revenues per unit of products 1 and 2 are \$30 and \$20 respectively. The total daily processing time available for each machine is 8 hours.

LPP

Maximize $Z = 30x_1 + 20x_2$

Subject to $2x_1 + x_2 \leq 8$ (Machine 1)

$x_1 + 3x_2 \leq 8$ (Machine 2)

$x_1, x_2 \geq 0$

Changes in RHS

	Machine 1	Machine 2
Shadow prce	\$ 14/ hour	\$ 2 /hour
Minimum capacity	2.67 hours	4 hrs
Maximum capacity	16 hours	24 hours

Questions

1. If JOBCO can increase the capacity of both machines, which machine should receive higher priority?

Ans Machine 1

2. A suggestion is made to increase the capacities of machines 1 and 2 at the additional cost of \$10 / hour. Is this advisable?

Ans Only machine 1 should be increased.

3. If the capacity of machine 1 is increased from the present 8 hours to 13 hours, how will the increase impact the optimum revenue?

Ans Increased to \$198.

4. Suppose the capacity of machine 1 is increased to 20 hours, how will this increase impact the optimum revenue?

Ans We do not have any conclusion

Changes in Objective Coefficients

1. Suppose that the unit revenues for products 1 and 2 are changed to \$35 and \$25 respectively, will the current optimum remain the same?

Ans. Yes, however optimal value changes to \$152.

2. Suppose that the unit revenue of product 2 is fixed at its current value of $c_2 = \$20$, what is the associated range for c_1 , that will keep the optimum unchanged.

Ans. $6.67 \leq c_1 \leq 40$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 9

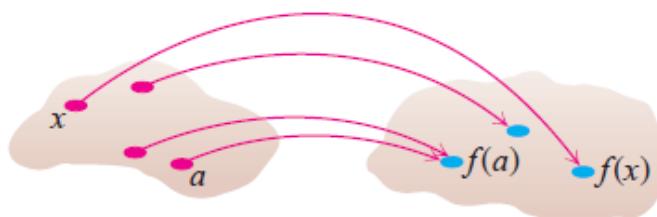
Agenda

- Functions
- Limits
- Continuity
- Intermediate Value Theorem
- Differentiability
- Taylor series expansion
- Maxima and minima

Functions

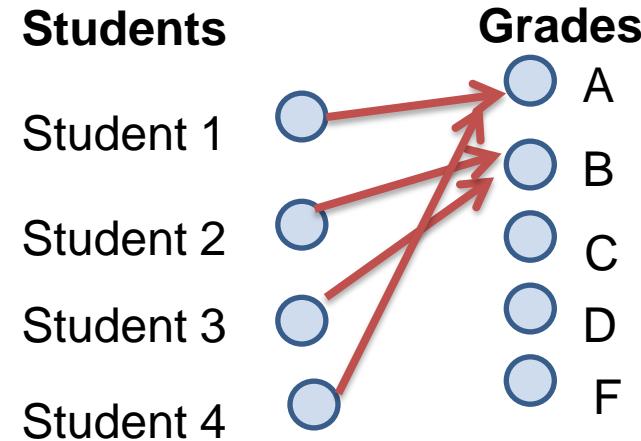
Definition: Let A and B be nonempty sets. A *function* f from A to B , denoted $f: A \rightarrow B$ is an assignment of each element of A to exactly one element of B . We write $f(a) = b$ if b is the unique element of B assigned by the function f to the element a of A .

Functions are sometimes called *mappings* or *transformations*.



D = domain set

Y = set containing the range



Functions

Given a function $f: A \rightarrow B$:

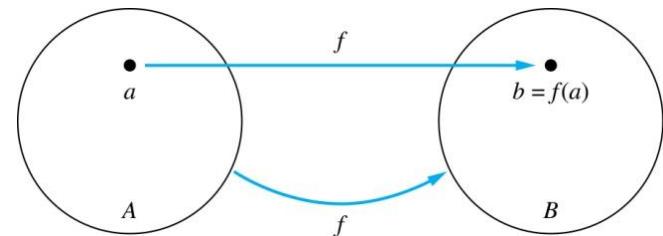
- We say f maps A to B or f is a *mapping* from A to B .
- A is called the *domain* of f .
- B is called the *co-domain* of f .

If $f(a) = b$,

- then b is called the *image* of a under f .
- a is called the *preimage* of b .

The range of f is the set of all images of points in A under f . We denote it by $f(A)$.

Two functions are *equal* when they have the same domain, the same co-domain and map each element of the domain to the same element of the co-domain.



Concept of Limits

Limit of the function – Value that $f(x)$ gets closer to as x approaches some number

$$f(x) = \frac{4}{3}x - 4,$$

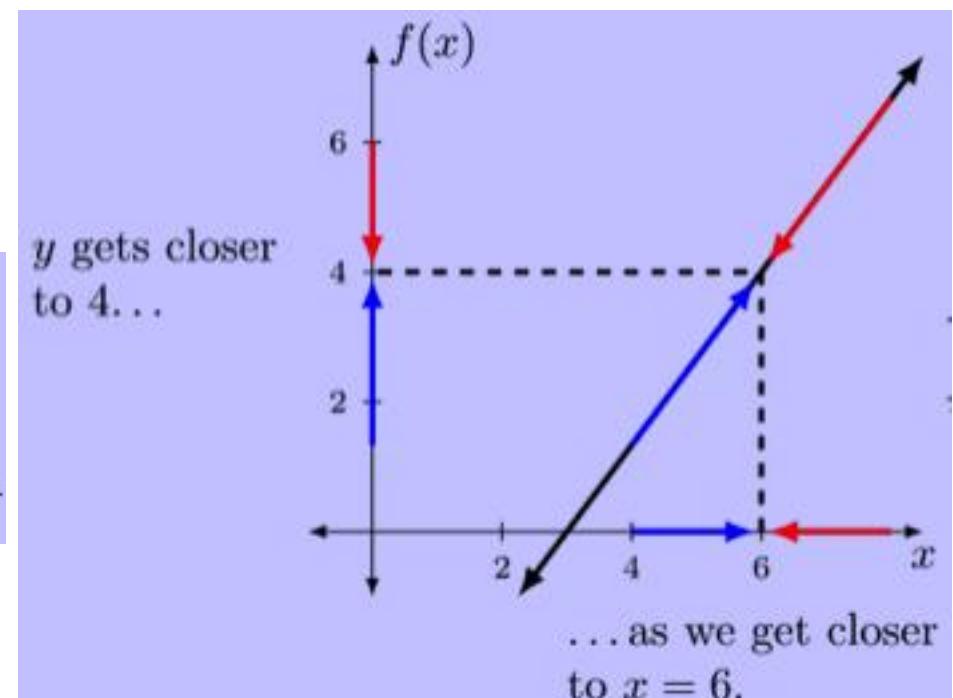
The “lim” tells us we’re looking for a limit value, not a function value.

$$\lim_{x \rightarrow 6} f(x) = 4$$

This tells us what the variable is, and what it is approaching.

This tells us which function we’re working with.

This is the value the function is approaching.



One sided Limits

Right Handed Limit

$$\lim_{x \rightarrow a^+} f(x) = L$$

Can make $f(x)$ as close to L as we want for all x sufficiently close to a with $x > a$ without actually letting x be a

Left Handed Limit

$$\lim_{x \rightarrow a^-} f(x) = L$$

Can make $f(x)$ as close to L as we want for all x sufficiently close to a with $x < a$ without actually letting x be a

NOTE: If two one sided limits **have different value**, then
normal limit will not exist

$$\lim_{x \rightarrow a^+} f(x) \neq \lim_{x \rightarrow a^-} f(x)$$

Limits

Check the existence of the limit for function $f(x) = |x| - |x - 1|$ at $x = 0$.

The limit of a function $f(x)$ at $x = a$ exists only when its left hand limit (*LHL*) and right hand limit (*RHL*) exist and are equal ,

$$\lim_{x \rightarrow a^-} f(x) = \lim_{x \rightarrow a^+} f(x).$$

Given, $\lim_{x \rightarrow 0} f(x) = \lim_{x \rightarrow 0} (|x| - |x - 1|)$

LHL : $\lim_{x \rightarrow 0^-} (|x| - |x - 1|)$

x is negative when $x \rightarrow 0^-$. Therefore $|x| = -x$

$$= \lim_{x \rightarrow 0^-} (-x - |x - 1|)$$

$x - 1$ is negative when $x \rightarrow 0^-$. Therefore $|x - 1| = -x + 1$

$$= \lim_{x \rightarrow 0^-} (-x - (-x + 1)) = \lim_{x \rightarrow 0^-} (-1)$$

$$= -1$$

Limits

$$\text{RHL : } \lim_{x \rightarrow 0^+} (|x| - |x - 1|)$$

x is positive when $x \rightarrow 0^+$. Therefore $|x| = x$

$$= \lim_{x \rightarrow 0^+} (x - |x - 1|)$$

$x - 1$ is negative when $x \rightarrow 0^+$. Therefore $|x - 1| = -x + 1$

$$= \lim_{x \rightarrow 0^+} (x - (-x + 1)) = \lim_{x \rightarrow 0^+} (2x - 1)$$

$$= -1.$$

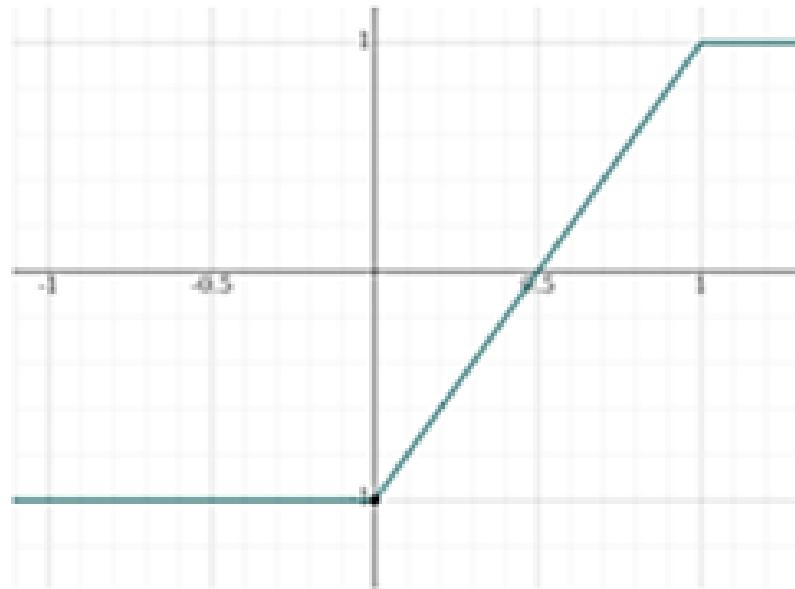


Figure: $f(x) = |x| - |x - 1|$

Continuity

Let f be a function. The f is continuous at $x = a$ if

$f(a)$ is defined This means a is in the domain of f .

$\lim_{x \rightarrow a} f(x)$ exists This means there exists a finite limit at $x = a$.

$\lim_{x \rightarrow a} f(x) = f(a)$ This means at $x = a$ the limit is equal to the functional value.

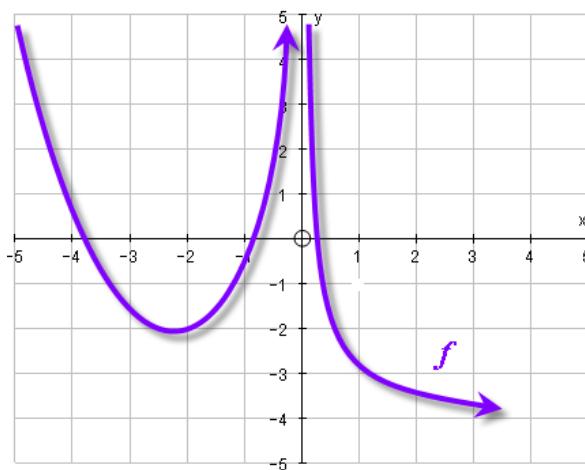
If a function is not continuous at a we say f is discontinuous at a .

A function is continuous on its domain if it is continuous at each point in the domain of f .

Continuity

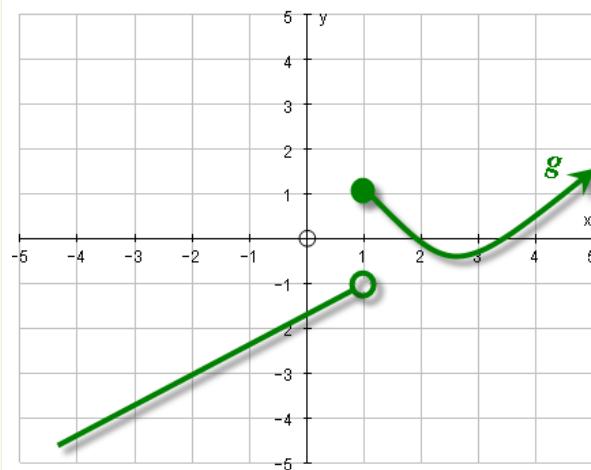
$f(0)$ is not defined.

The function is continuous on its domain, but not continuous at every real number.



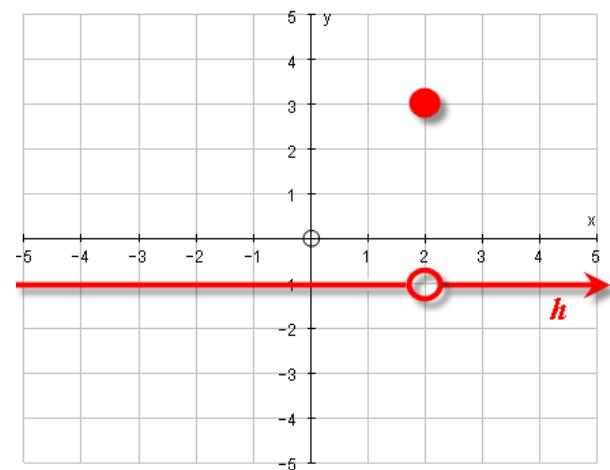
$\lim_{x \rightarrow 1} g(x)$ does not exist

The function does not have a limit at $x = a$.



$\lim_{x \rightarrow 2} h(x) \neq h(2)$

The limit and the functional value are not equal $x = a$.



Properties of Continuous Function

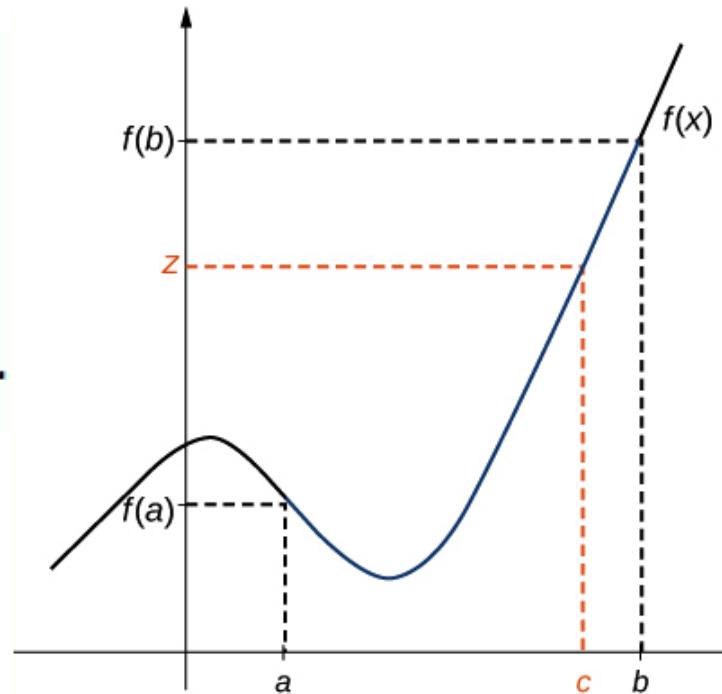
Let f and g be continuous functions. Then

1. $f + g$ is a continuous function.
2. fg is a continuous function.
3. $\frac{f}{g}$ is a continuous function, whenever $g(x) \neq 0$.

INTERMEDIATE VALUE THEOREM

Let f be continuous over a closed, bounded interval $[a, b]$. If z is any real number between $f(a)$ and $f(b)$, then there is a number c in $[a, b]$ satisfying $f(c) = z$ in Figure.

Note: If a function is continuous over a closed interval, then function takes on every value between the values at its end points



Intermediate Value Theorem

Show that there is a root of equation $4x^3 - 6x^2 + 3x - 2 = 0$ between 1 and 2

We are looking for a solution of given equation that is number c between 1 and 2 such that $f(c) = 0$

Given : $a = 1$, $b = 2$

$$f(1) = 4 - 6 + 3 - 2 = -1 < 0$$

$$f(2) = 32 - 24 + 6 - 2 = 12 > 0$$

$f(1) < 0 < f(2) \rightarrow$ d = 0 is number between f(1) and f(2)

f is a continuous since it is a polynomial. Intermediate value theorem says that there is a number c between 1 and 2 such that $f(c) = 0$

Equation has atleast one root c in the interval (1,2)

NOTE : We can precisely **locate root** using Intermediate Value Theorem

Intermediate Value Theorem **fails for Discontinuous Functions**

Derivative

The derivative of f at x is $f'(x) \equiv \frac{df}{dx}(x) = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$.

- The derivative is the slope of the tangent
- When the limit exists the function is called **DIFFERENTIABLE** at x .
- The term differentiable function is used to denote functions differentiable at every point in the domain
- If a function is differentiable at a point, it is continuous at that point

Functions which are not differentiable

Cube Root Function

$1/x$ in domain $[0, \infty]$

Floor and Ceiling Functions at integer values

Differentiability

$$\begin{aligned} \lim_{h \rightarrow 0_+} \frac{f(h) - f(0)}{h} &= \lim_{h \rightarrow 0_+} \frac{h - 0}{h} && \text{cancellation of } h \text{ okay, since } h \neq 0 \text{ for limit at 0} \\ &= \lim_{h \rightarrow 0_+} 1 \\ &= 1 \end{aligned}$$

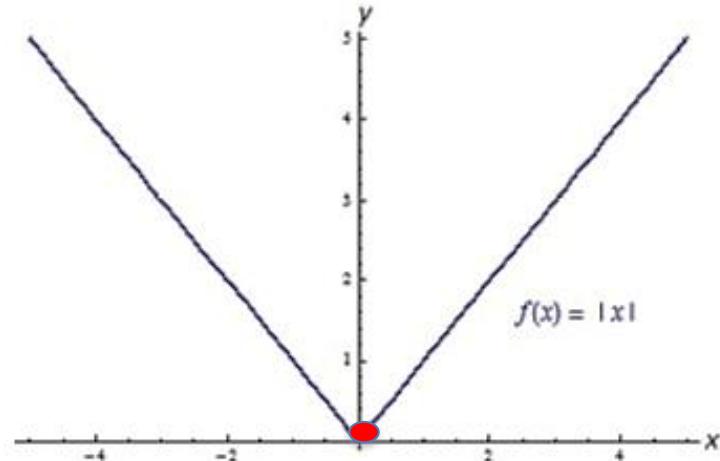
But

$$\begin{aligned} \lim_{h \rightarrow 0_-} \frac{f(h) - f(0)}{h} &= \lim_{h \rightarrow 0_-} \frac{-h - 0}{h} && \text{cancellation of } h \text{ okay, since } h \neq 0 \text{ for limit at 0} \\ &= \lim_{h \rightarrow 0_-} -1 \\ &= -1 \end{aligned}$$

Since the left- and right-hand limits do not agree,

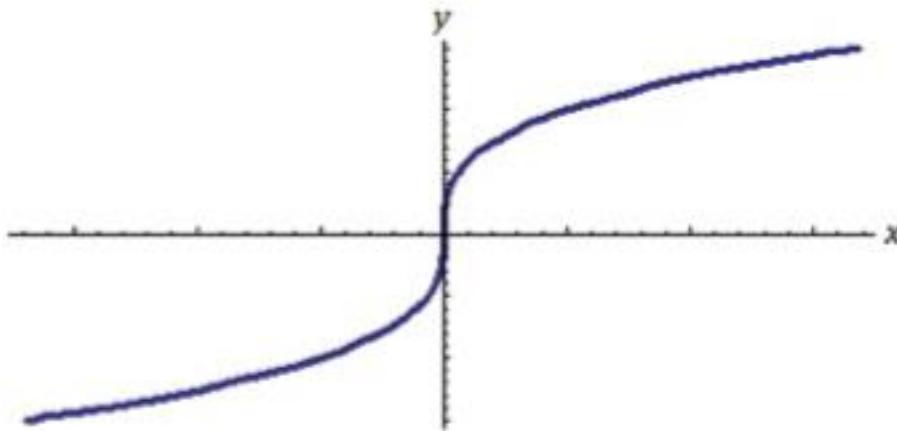
$$\lim_{h \rightarrow 0} \frac{f(h) - f(0)}{h}$$

Does not exist, and so $|x|$ is not differentiable at $x = 0$.



$$f(x) = |x| = \begin{cases} x & x \geq 0 \\ -x & x < 0 \end{cases}$$

Differentiability



$$g(x) = x^{1/3}$$

The graph is smooth at $x = 0$, but does appear to have a vertical tangent.

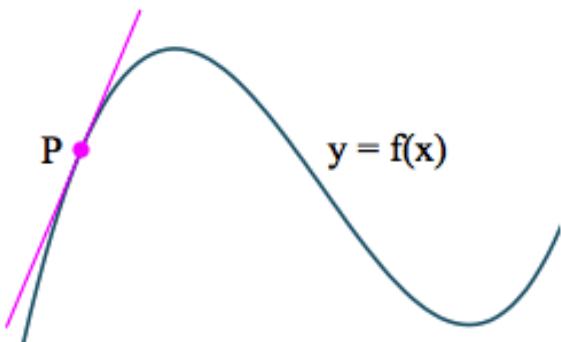
$$\lim_{h \rightarrow 0} \frac{(0+h)^{1/3} - 0^{1/3}}{h} = \lim_{h \rightarrow 0} \frac{(h)^{1/3}}{h} = \lim_{h \rightarrow 0} \frac{1}{h^{2/3}}$$

As $h \rightarrow 0$, the denominator becomes small, so the fraction grows without bound. Hence g is not differentiable at $x = 0$.

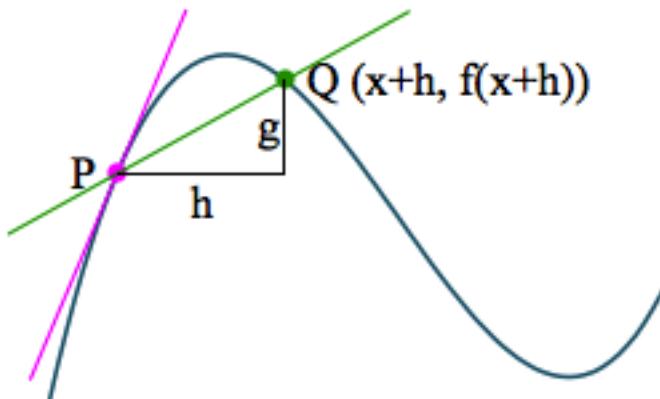
Differentiation from First Principle

$\frac{dy}{dx}$ or $f'(x)$ or y' .

$$\frac{dy}{dx} = \lim_{h \rightarrow 0} \frac{f(x + h) - f(x)}{h}$$



Slope of the tangent at P .



Slope of the line PQ .

The value $\frac{g}{h}$ is an approximation to the slope of the tangent which we require.

Taylor series

Let f be a function with derivatives of all orders throughout some interval containing a as an interior point. Then the **Taylor series generated by f at $x = a$** is

$$\sum_{k=0}^{\infty} \frac{f^{(k)}(a)}{k!} (x - a)^k = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!} (x - a)^2 + \dots + \frac{f^{(n)}(a)}{n!} (x - a)^n + \dots.$$

DEFINITION Taylor Polynomial of Order n

Let f be a function with derivatives of order k for $k = 1, 2, \dots, N$ in some interval containing a as an interior point. Then for any integer n from 0 through N , the **Taylor polynomial of order n** generated by f at $x = a$ is the polynomial

$$P_n(x) = f(a) + f'(a)(x - a) + \frac{f''(a)}{2!} (x - a)^2 + \dots + \frac{f^{(k)}(a)}{k!} (x - a)^k + \dots + \frac{f^{(n)}(a)}{n!} (x - a)^n.$$

Example

Find the Taylor series and the Taylor polynomials generated by $f(x) = e^x$ at $x = 0$.

Solution Since

$$f(x) = e^x, \quad f'(x) = e^x, \quad \dots, \quad f^{(n)}(x) = e^x, \quad \dots,$$

we have

$$f(0) = e^0 = 1, \quad f'(0) = 1, \quad \dots, \quad f^{(n)}(0) = 1, \quad \dots.$$

The Taylor series generated by f at $x = 0$ is

$$\begin{aligned} f(0) + f'(0)x + \frac{f''(0)}{2!}x^2 + \dots + \frac{f^{(n)}(0)}{n!}x^n + \dots \\ = 1 + x + \frac{x^2}{2} + \dots + \frac{x^n}{n!} + \dots \\ = \sum_{k=0}^{\infty} \frac{x^k}{k!}. \end{aligned}$$

Example

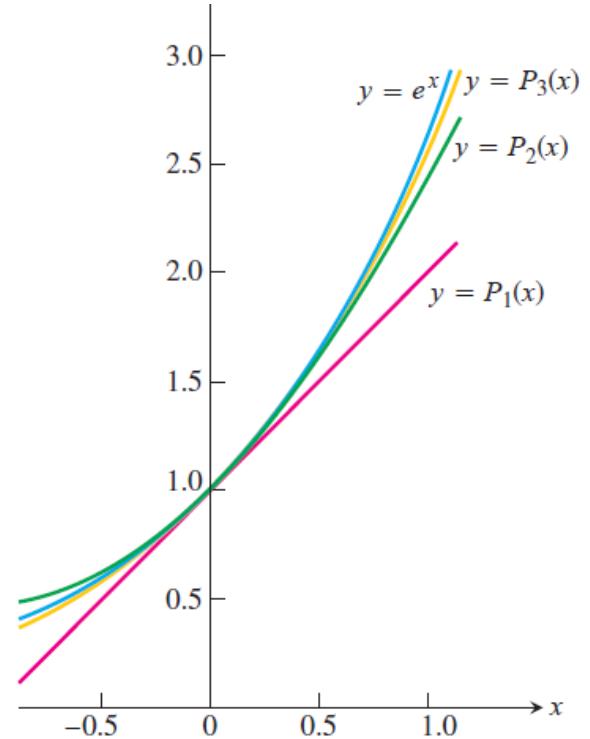
The Taylor polynomial of order n at $x = 0$ is

$$P_n(x) = 1 + x + \frac{x^2}{2} + \cdots + \frac{x^n}{n!}.$$

$$P_1(x) = 1 + x$$

$$P_2(x) = 1 + x + (x^2/2!)$$

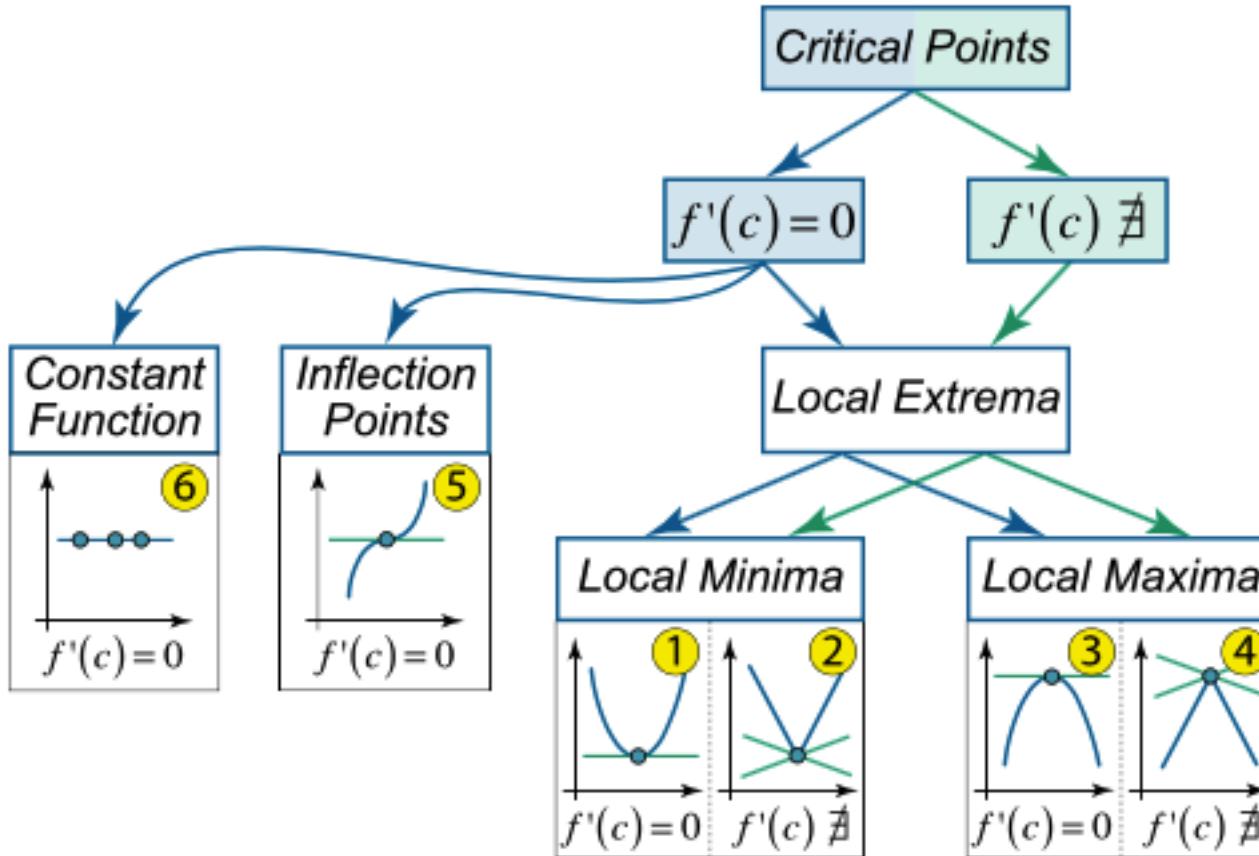
$$P_3(x) = 1 + x + (x^2/2!) + (x^3/3!).$$



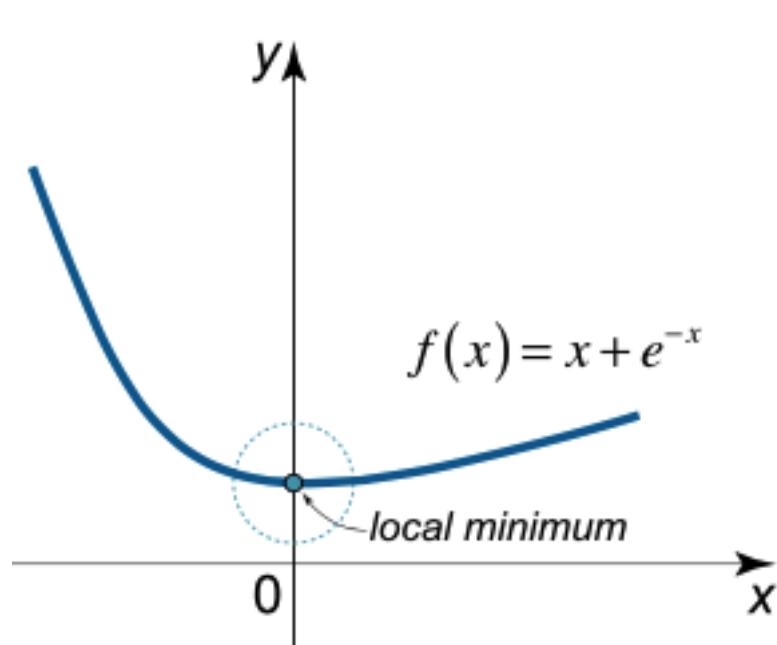
Critical Points

Let $f(x)$ be a function and let c be a point in the domain of the function.

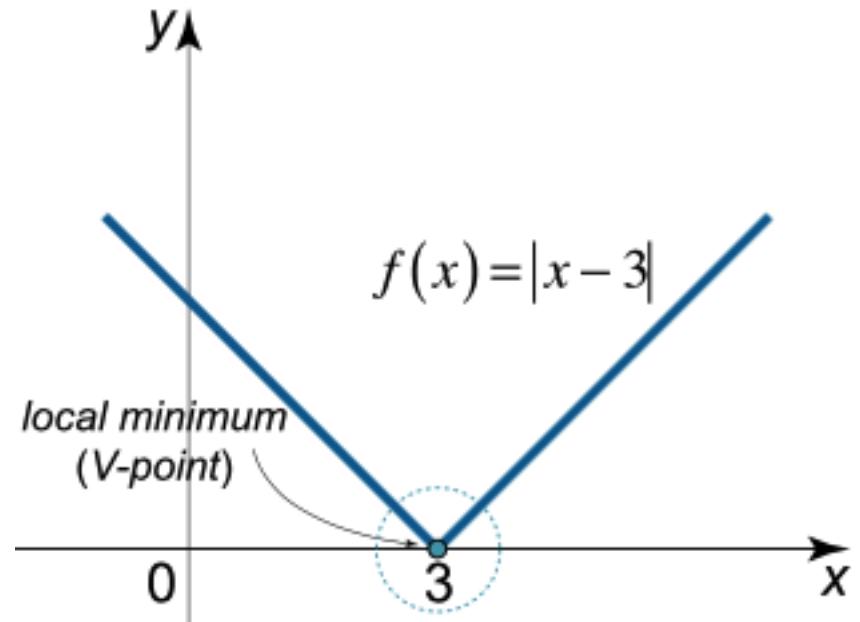
The point c is called a critical point of f if either $f'(c) = 0$ or $f'(c)$ does not exist.



Critical Point - Local Minimum

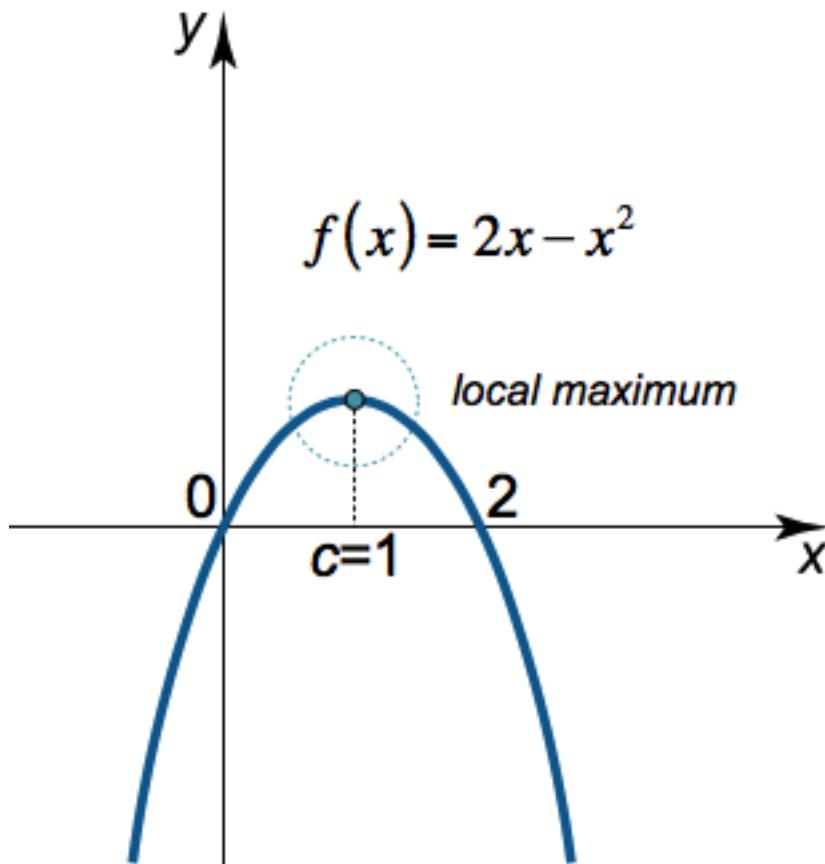


$x = c$ is **local minimum** if function changes from decreasing to increasing at that point

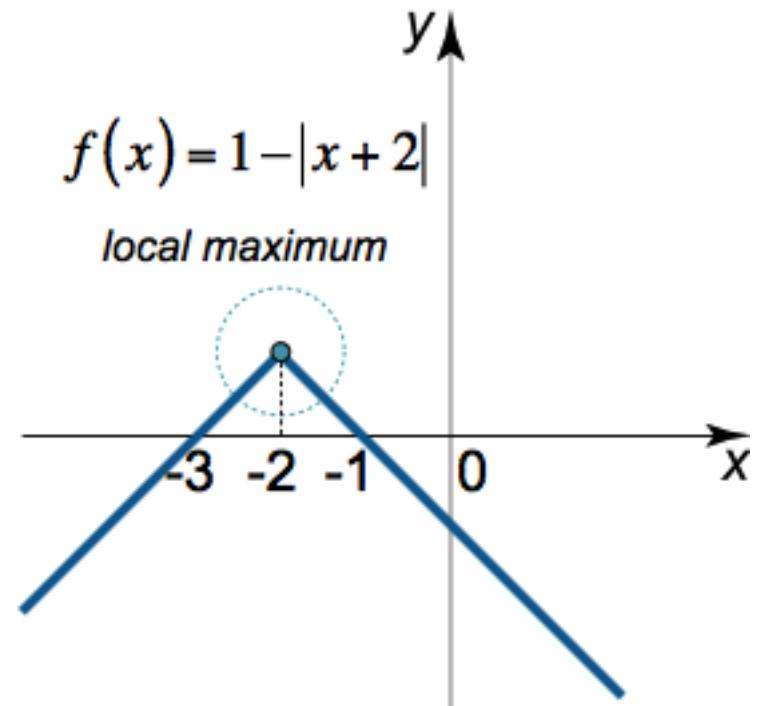


Derivative does not exist at this point

Critical Point – Local Maximum

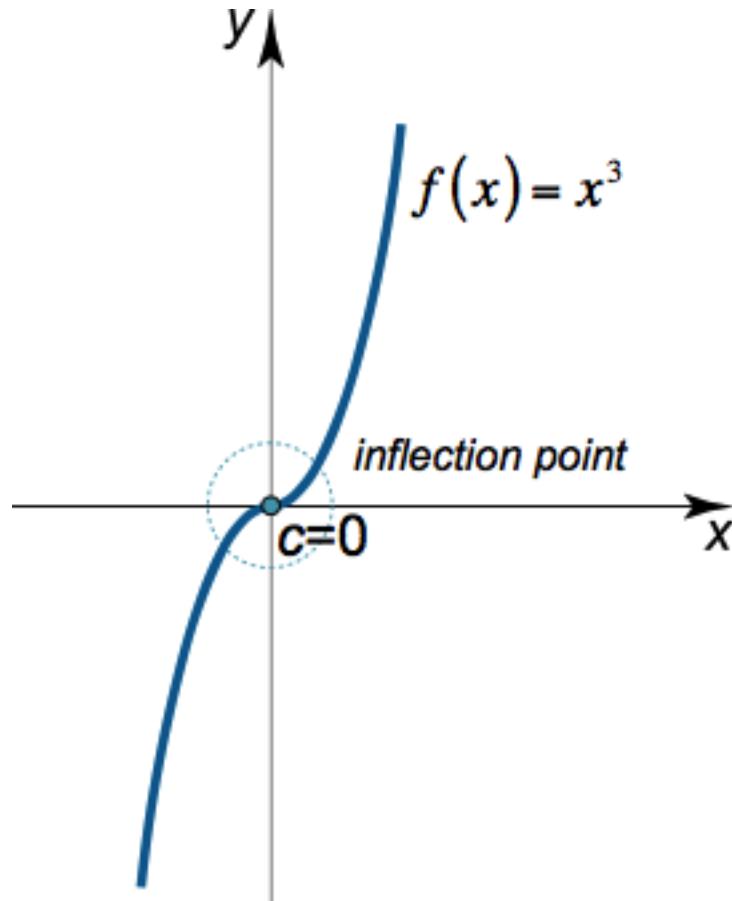


Critical Point $x = c$ is **local maximum** if the function changes from increasing to decreasing at that point

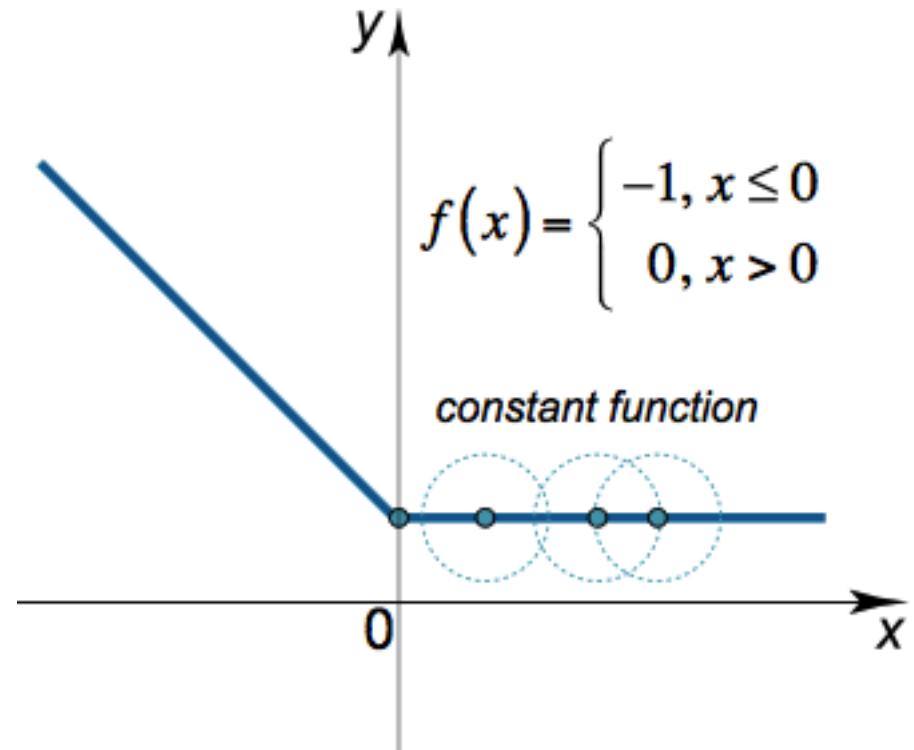


Derivative does not exist at this point

Critical Point



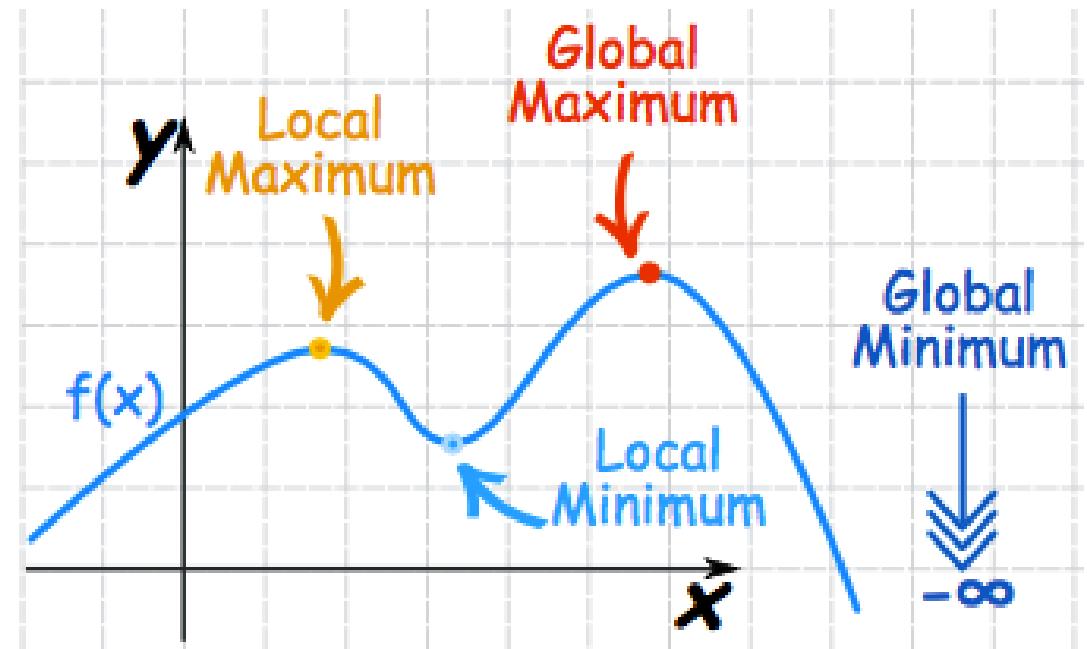
A critical point $x = c$ is an inflection point if function changes concavity at that point



Each point of a constant function is critical

Global Maximum and Minimum

- The maximum or minimum over the entire function is called Global maximum or minimum
- There is only one global maximum and global minimum but there can be more than one local maximum or minimum



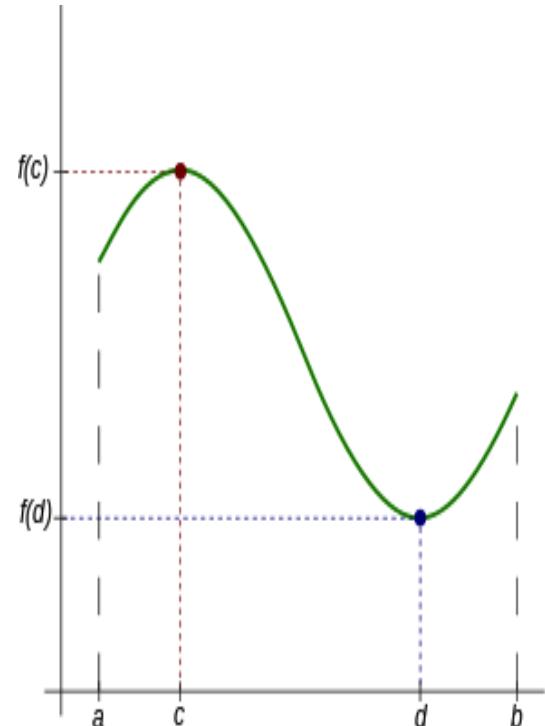
Extreme Value Theorem

Extreme Value Theorem guarantees that every function must have absolute(global) maximum and minimum

Assumptions:

1. f is continuous on an interval
2. Interval is a closed interval $[a,b]$

If either assumption fails, we are not allowed to draw conclusion that f hits minimum and maximum value on that closed interval



A continuous function $f(x)$ on the closed interval $[a, b]$ showing the absolute max (red) and the absolute min (blue).

Second Derivative Test

First Derivative Test is used in conjunction with extreme value theorem to find the absolute maximum and minimum of a real valued function defined on a closed, bounded interval

After establishing the critical points of a function, **second derivative test** uses the value of second derivative at those points to determine whether such points are a local maximum or local minimum.

If a function f is twice differentiable at a critical point x ($f'(x) = 0$ then

- If $f''(x) < 0$ then f has a local maximum at x .
- If $f''(x) > 0$ then f has a local minimum at x .
- If $f''(x) = 0$, the test is inconclusive.

Second Derivative Test

Example : $f(x) = 2x^3 - 3x^2 - 36x + 2$

$$f'(x) = 0 \rightarrow 6x^2 - 6x - 36 = 0$$

$$x = 3, -2$$

Compute $f''(x) = 12x - 6$

At $x = 3 \rightarrow f''(3) = 12(3) - 6 = 30 > 0$

At $x = -2 \rightarrow f''(-2) = 12(-2) - 6 = -30 < 0$

$X = -2$ Point is Maxima

$X = 3$ Point is Minima

- If $f''(x) < 0$ then f has a local maximum at x .
- If $f''(x) > 0$ then f has a local minimum at x .
- If $f''(x) = 0$, the test is inconclusive.



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 10

Agenda

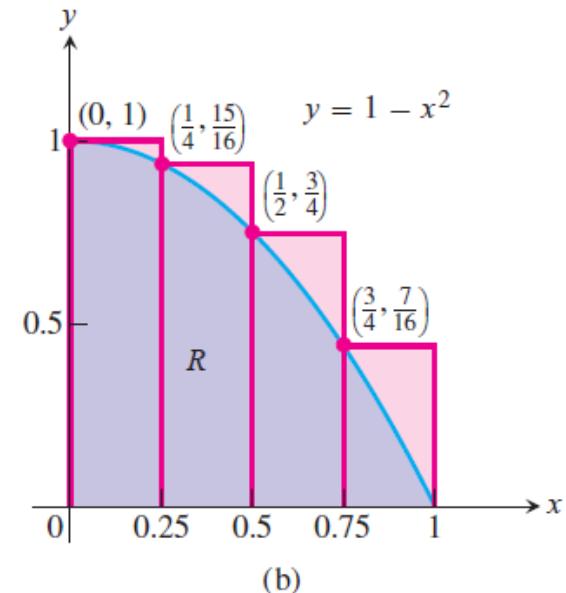
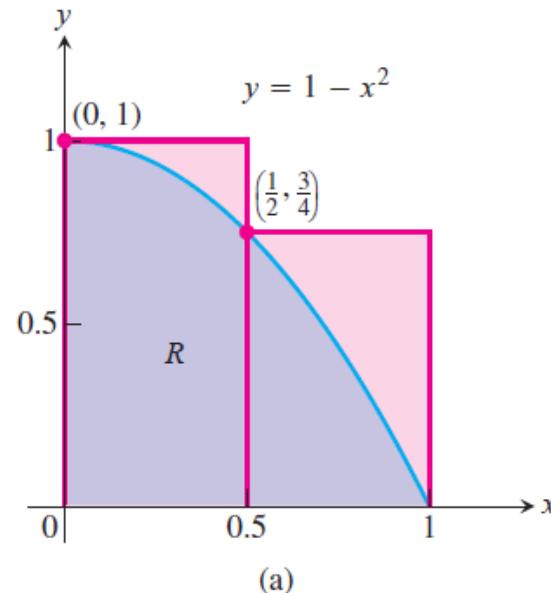
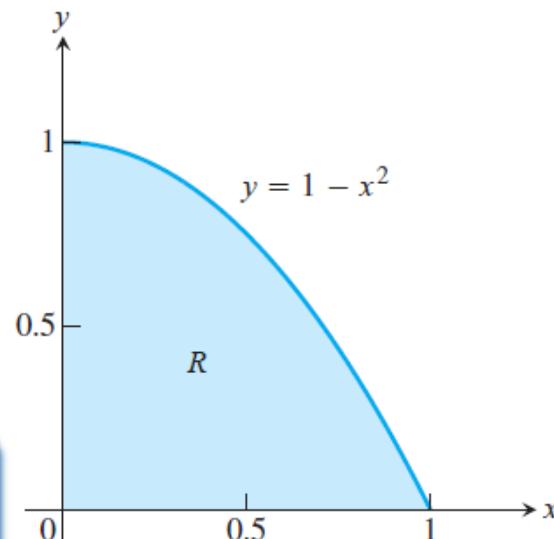
- Integral properties
 - cdf and pdf
 - Even and odd integrands
 - Integration by parts
- Partial derivatives and Directional derivatives
- Extremum in several variables
 - Lagrange multipliers
 - Method of steepest descent.

Properties of integrals

- The idea behind integration is that we can effectively compute many quantities by breaking them into small pieces, and then summing the contributions from each small part.
- Estimating with Finite Sums

$$A \approx 1 \cdot \frac{1}{2} + \frac{3}{4} \cdot \frac{1}{2} = \frac{7}{8} = 0.875.$$

$$A \approx 1 \cdot \frac{1}{4} + \frac{15}{16} \cdot \frac{1}{4} + \frac{3}{4} \cdot \frac{1}{4} + \frac{7}{16} \cdot \frac{1}{4} = \frac{25}{32} = 0.78125,$$



The Definite Integral

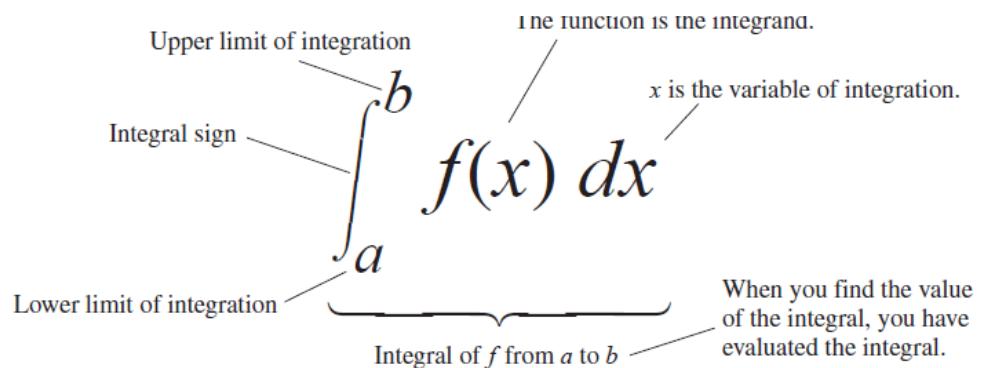
DEFINITION The Definite Integral as a Limit of Riemann Sums

Let $f(x)$ be a function defined on a closed interval $[a, b]$. We say that a number I is the **definite integral of f over $[a, b]$** and that I is the limit of the Riemann sums $\sum_{k=1}^n f(c_k) \Delta x_k$ if the following condition is satisfied:

Given any number $\epsilon > 0$ there is a corresponding number $\delta > 0$ such that for every partition $P = \{x_0, x_1, \dots, x_n\}$ of $[a, b]$ with $\|P\| < \delta$ and any choice of c_k in $[x_{k-1}, x_k]$, we have

$$\left| \sum_{k=1}^n f(c_k) \Delta x_k - I \right| < \epsilon.$$

$$\lim_{n \rightarrow \infty} \sum_{k=1}^n f(c_k) \Delta x = I = \int_a^b f(x) dx.$$

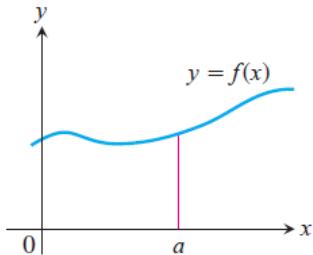


Properties of definite integrals

1. *Order of Integration:* $\int_b^a f(x) dx = - \int_a^b f(x) dx$ A Definition
2. *Zero Width Interval:* $\int_a^a f(x) dx = 0$ Also a Definition
3. *Constant Multiple:* $\int_a^b kf(x) dx = k \int_a^b f(x) dx$ Any Number k
 $\int_a^b -f(x) dx = - \int_a^b f(x) dx$ $k = -1$
4. *Sum and Difference:* $\int_a^b (f(x) \pm g(x)) dx = \int_a^b f(x) dx \pm \int_a^b g(x) dx$
5. *Additivity:* $\int_a^b f(x) dx + \int_b^c f(x) dx = \int_a^c f(x) dx$
6. *Max-Min Inequality:* If f has maximum value $\max f$ and minimum value $\min f$ on $[a, b]$, then

$$\min f \cdot (b - a) \leq \int_a^b f(x) dx \leq \max f \cdot (b - a).$$
7. *Domination:* $f(x) \geq g(x)$ on $[a, b] \Rightarrow \int_a^b f(x) dx \geq \int_a^b g(x) dx$
 $f(x) \geq 0$ on $[a, b] \Rightarrow \int_a^b f(x) dx \geq 0$ (Special Case)

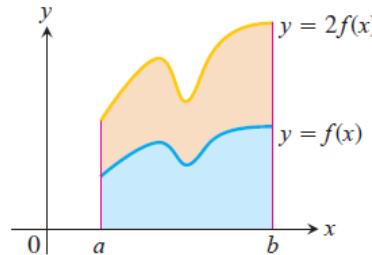
Properties



(a) Zero Width Interval:

$$\int_a^a f(x) dx = 0.$$

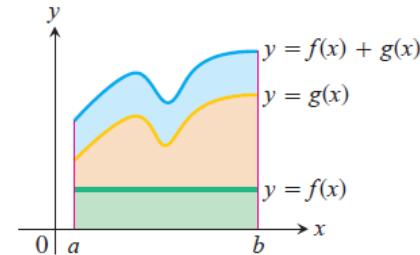
(The area over a point is 0.)



(b) Constant Multiple:

$$\int_a^b kf(x) dx = k \int_a^b f(x) dx.$$

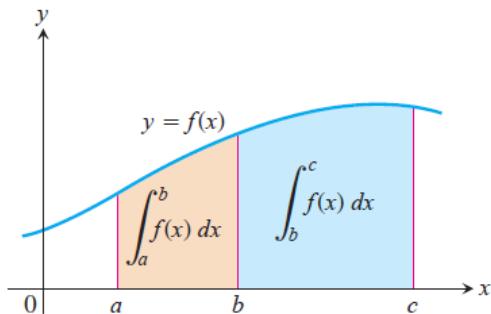
(Shown for $k = 2$.)



(c) Sum:

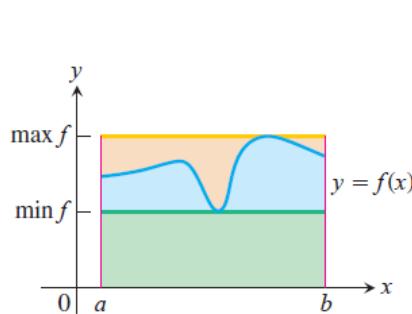
$$\int_a^b (f(x) + g(x)) dx = \int_a^b f(x) dx + \int_a^b g(x) dx$$

(Areas add)



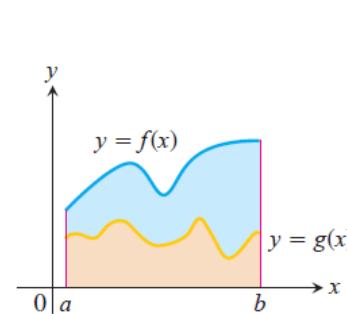
(d) Additivity for definite integrals:

$$\int_a^b f(x) dx + \int_b^c f(x) dx = \int_a^c f(x) dx$$



(e) Max-Min Inequality:

$$\begin{aligned} \min f \cdot (b - a) &\leq \int_a^b f(x) dx \\ &\leq \max f \cdot (b - a) \end{aligned}$$



(f) Domination:

$$\begin{aligned} f(x) &\geq g(x) \text{ on } [a, b] \\ \Rightarrow \int_a^b f(x) dx &\geq \int_a^b g(x) dx \end{aligned}$$

FIGURE 5.11

Properties

DEFINITION Area Under a Curve as a Definite Integral

If $y = f(x)$ is nonnegative and integrable over a closed interval $[a, b]$, then the **area under the curve $y = f(x)$ over $[a, b]$** is the integral of f from a to b ,

$$A = \int_a^b f(x) dx.$$

DEFINITION The Average or Mean Value of a Function

If f is integrable on $[a, b]$, then its **average value on $[a, b]$** , also called its **mean value**, is

$$\text{av}(f) = \frac{1}{b-a} \int_a^b f(x) dx.$$

CDF and PDF

- If X is a continuous random variable with density f , then **cumulative distribution function** (cdf) is defined by

$$F_X(x) := \mathbb{P}(X \leq x) = \int_{-\infty}^x f(t)dt.$$

- Pictorially, $F(x)$ is the area under the density $f(t)$ from $-\infty < t \leq x$.
- For $a < b$, we can use the cdf to compute probabilities of the form

$$\begin{aligned}\mathbb{P}(a \leq X \leq b) &= \int_a^b f(t)dt \\ &= \int_{-\infty}^b f(t)dt - \int_{-\infty}^a f(t)dt \\ &= F(b) - F(a).\end{aligned}$$

CDF and PDF

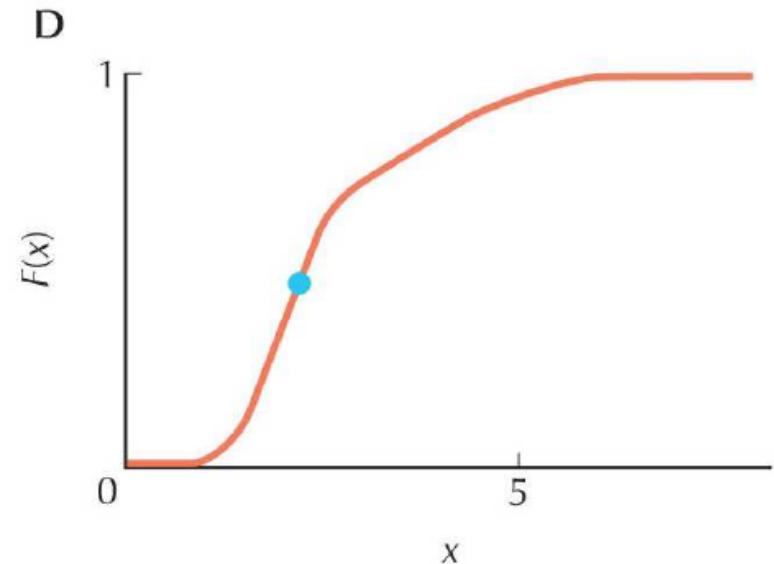
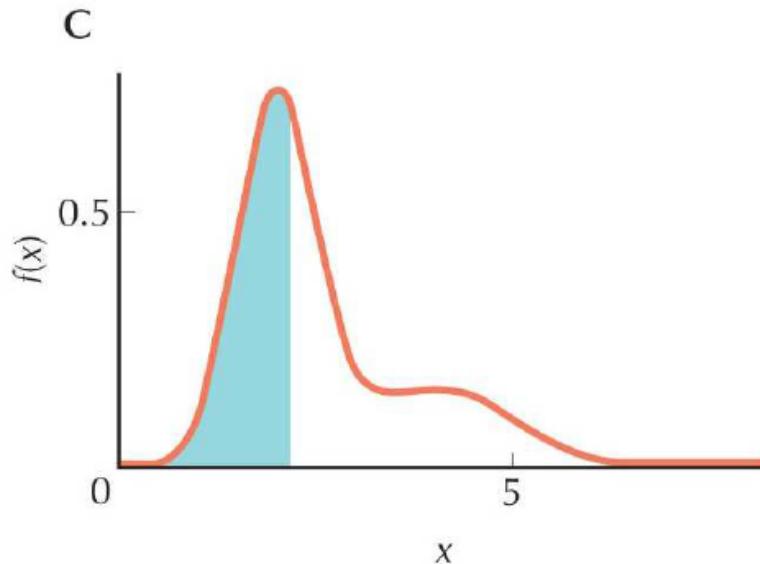
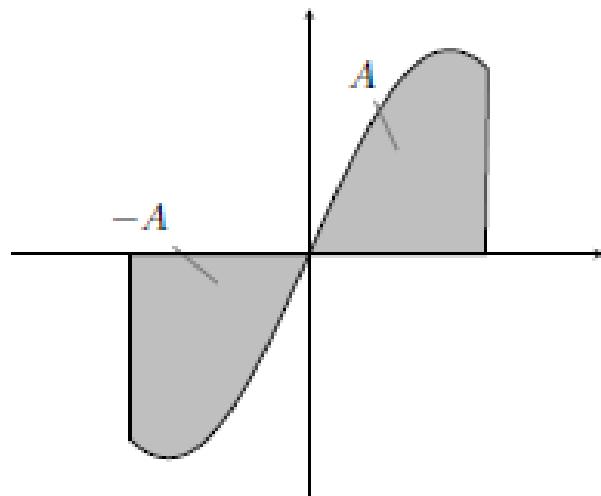


Figure 1: The relationship between pdf $f(x)$ and cdf $F(x)$.

Even and odd integrands

Principle 1: The integral of an odd function from $-p$ to p is zero.

If one looks at the graph of an odd function, the area under the curve and to the left of $x = 0$ is *exactly the same* as the area under the curve and to the right of $x = 0$, but *opposite* in sign. Therefore,



$$\left[\begin{array}{l} \text{area under } f \text{ and} \\ \text{to the left of } x = 0 \end{array} \right] + \left[\begin{array}{l} \text{area under } f \text{ and} \\ \text{to the right of } x = 0 \end{array} \right] = 0,$$

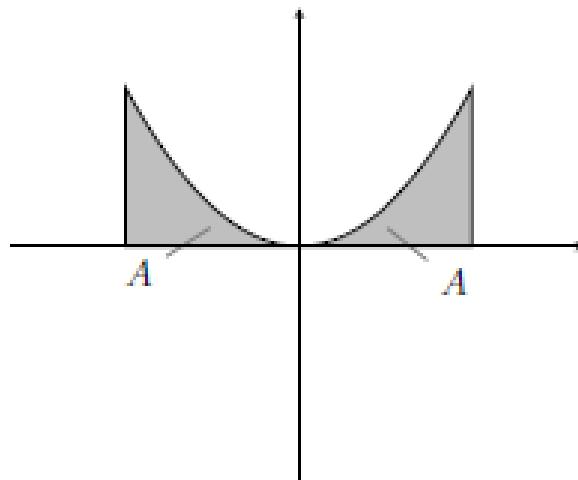
or, in terms of integrals,

$$\int_{-P}^P f(t) dt = 0 \text{ if } f \text{ is an odd function.}$$

Even and odd integrands

Principle 2: The integral of an even function from $-p$ to p is twice the integral from 0 to p .

Again looking at the graph of a typical even function, the area under the curve and to the left of $x = 0$ is again *exactly the same* as the area under the curve and to the right of $x = 0$, but unlike the case for odd functions, it has the *same sign*. Thus we have



$$\left[\begin{array}{l} \text{area under } f \text{ and} \\ \text{to the left of } x = 0 \end{array} \right] + \left[\begin{array}{l} \text{area under } f \text{ and} \\ \text{to the right of } x = 0 \end{array} \right] = 2 \left[\begin{array}{l} \text{area under } f \text{ and} \\ \text{to the right of } x = 0 \end{array} \right].$$

Translating this into integrals, we have the mathematical statement

$$\int_{-p}^p f(t) dt = 2 \int_0^p f(t) dt \text{ if } f \text{ is an even function.}$$

Integration by parts

We already know how to differentiate a product: if

$$y = u v$$

then

$$\frac{dy}{dx} = \frac{d(uv)}{dx} = u \frac{dv}{dx} + v \frac{du}{dx}.$$

Rearranging this rule:

$$u \frac{dv}{dx} = \frac{d(uv)}{dx} - v \frac{du}{dx}.$$

Now integrate both sides:

$$\int u \frac{dv}{dx} dx = \int \frac{d(uv)}{dx} dx - \int v \frac{du}{dx} dx.$$

The first term on the right simplifies since we are simply integrating what has been differentiated.

$$\int u \frac{dv}{dx} dx = u v - \int v \frac{du}{dx} dx.$$

This is the formula known as **integration by parts**.

Partial Derivative

When input of function is made up of multiple variables, we want to see how function changes as we let just one of those variable change while holding all the others constant

$$\frac{\partial f}{\partial x}$$

f is a multivariable function

$\frac{\partial f}{\partial x}$ Tiny change in functions output
 $\frac{\partial x}{\partial x}$ Tiny change in x

$$\frac{\partial f}{\partial x} = \underbrace{\frac{\partial}{\partial x} x^2 y}_{\text{Treat } y \text{ as constant;}} = 2xy$$

take derivative.

$$\frac{\partial f}{\partial y} = \underbrace{\frac{\partial}{\partial y} x^2 y}_{\text{Treat } x \text{ as constant;}} = x^2 \cdot 1$$

take derivative.

Gradient

The gradient of a scalar-valued multivariable function $f(x, y, \dots)$, denoted ∇f , packages all its partial derivative information into a vector:

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \\ \vdots \end{bmatrix}$$

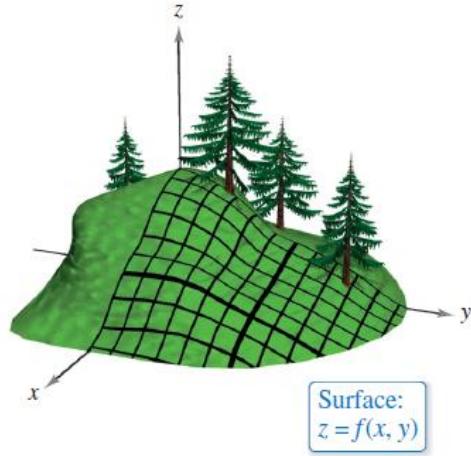
In particular, this means ∇f is a vector-valued function.

If you imagine standing at a point (x_0, y_0, \dots) in the input space of f , the vector $\nabla f(x_0, y_0, \dots)$ tells you which direction you should travel to increase the value of f most rapidly.

These gradient vectors— $\nabla f(x_0, y_0, \dots)$ —are also perpendicular to the contour lines of f .

Directional Derivatives

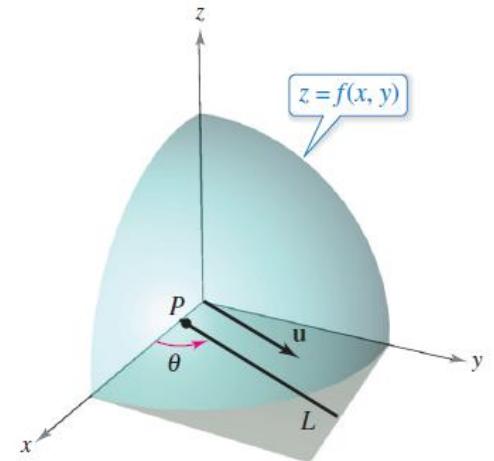
- You are standing on the hillside represented by $z = f(x,y)$ in Figure and want to determine the hill's incline toward the z - axis.



- The slope in the y -direction would be given by the partial derivative $f_y(x, y)$, and the slope in the x -direction would be given by the partial derivative $f_x(x, y)$.
- These two partial derivatives can be used to find the slope in *any* direction.

Directional derivatives

- To determine the slope at a point on a surface, you will define a new type of derivative called a **directional derivative**.
- Let $z = f(x, y)$ be a *surface* and $P(x_0, y_0)$ be a *point* in the domain of f , as shown in Figure.
- The “direction” of the directional derivative is given by a unit vector $\mathbf{u} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}$ where θ the vector makes with the positive x -axis.



THEOREM 13.9 Directional Derivative

If f is a differentiable function of x and y , then the directional derivative of f in the direction of the unit vector $\mathbf{u} = \cos \theta \mathbf{i} + \sin \theta \mathbf{j}$ is

$$D_{\mathbf{u}}f(x, y) = f_x(x, y) \cos \theta + f_y(x, y) \sin \theta.$$

Directional Derivative

Because the gradient of f is a vector, you can write the directional derivative of f in the direction of \mathbf{u} as

$$D_{\mathbf{u}}f(x, y) = [f_x(x, y)\mathbf{i} + f_y(x, y)\mathbf{j}] \cdot [\cos \theta \mathbf{i} + \sin \theta \mathbf{j}].$$

In other words, the directional derivative is the dot product of the gradient and the direction vector.

THEOREM 13.10 Alternative Form of the Directional Derivative

If f is a differentiable function of x and y , then the directional derivative of f in the direction of the unit vector \mathbf{u} is

$$D_{\mathbf{u}}f(x, y) = \nabla f(x, y) \cdot \mathbf{u}.$$

Hessian Matrix

Square matrix which is a way of organizing all second order partial derivative information of a multivariable function

Hessian matrix is always symmetric matrix → entries of the matrix are symmetric across its main diagonal

$$Hf(x, y) = \begin{bmatrix} f_{xx}(x, y) & \underline{f_{xy}(x, y)} \\ \underline{f_{yx}(x, y)} & f_{yy}(x, y) \end{bmatrix}$$

$$H f (x, y) \equiv \begin{vmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{vmatrix}.$$

Define $D(x,y)$ to be the **determinant**

Note : We expect the eigenvalues of the Hessian to be **positive at local minimum** and **negative at local maximum**

If the Hessian has both **positive and negative eigen values** the corresponding point **must be saddle point**

Hessian Matrix

The function $f(x, y) = x^3 + 2(x - y)^2 - 3x$ has a critical point at $(1, 1)$. Classify this critical point as a local maximum, a local minimum, or a saddle point.

SOLUTION The Hessian of f is

$$Hf(x, y) = \begin{bmatrix} 6x + 4 & -4 \\ -4 & 4 \end{bmatrix}$$

and in particular

$$Hf(1, 1) = \begin{bmatrix} 10 & -4 \\ -4 & 4 \end{bmatrix}$$

The eigenvalues of this matrix are 2 and 12, so $(1, 1)$ is a local minimum.

The function $f(x, y) = 6 \cos x + 4x \sin y$ has a critical point at $(0, 0)$. Classify this critical point as a local maximum, a local minimum, or a saddle point.

SOLUTION The Hessian of f is

$$Hf(x, y) = \begin{bmatrix} -6 \cos x & 4 \cos y \\ 4 \cos y & -4x \sin y \end{bmatrix}$$

and in particular

$$Hf(0, 0) = \begin{bmatrix} -6 & 4 \\ 4 & 0 \end{bmatrix}$$

The eigenvalues of this matrix are -8 and 2 , so $(0, 0)$ is a saddle point.

Lagrange Multiplier

-
1. Constrained Optimization problem
 2. Lagrange Multiplier is developed to figure out the maxima/minima of an objective function f under a constraint function g

Core Idea : Look for points where the **contour lines of f and g are tangent to each other**

Suppose you want to maximize this function:

$$f(x, y) = 2x + y$$

Function that needs to be optimized

But let's also say you limited yourself to inputs (x, y) which satisfy the following equation:

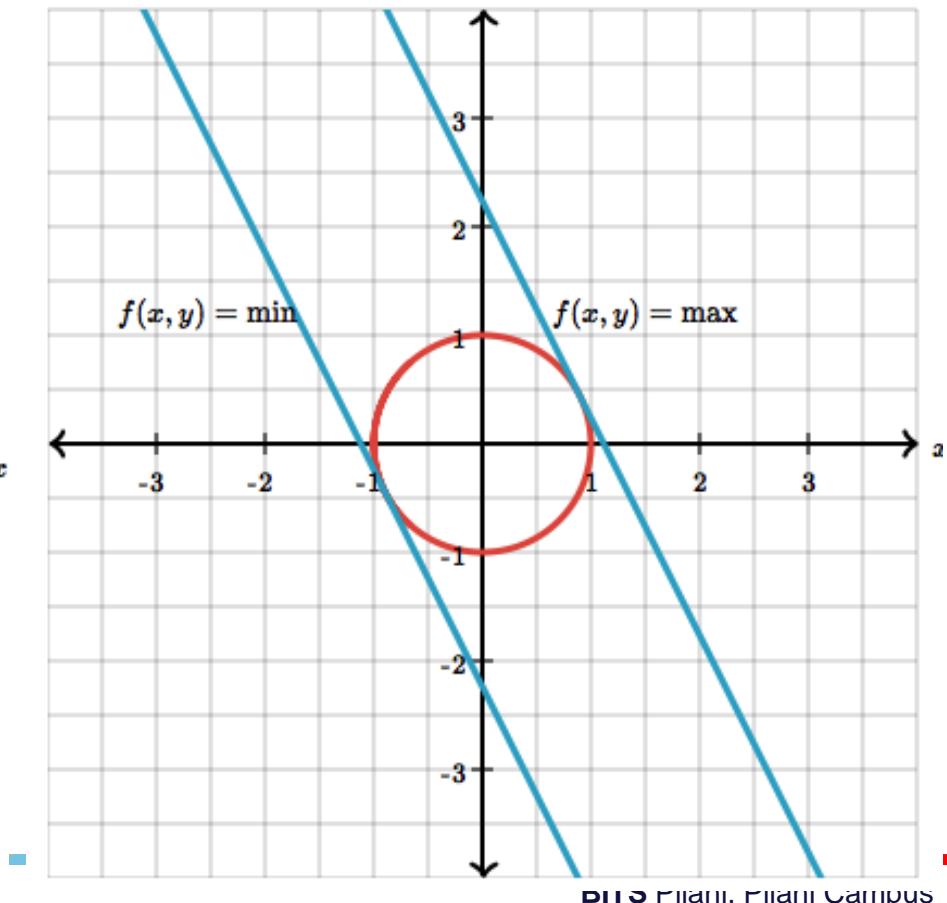
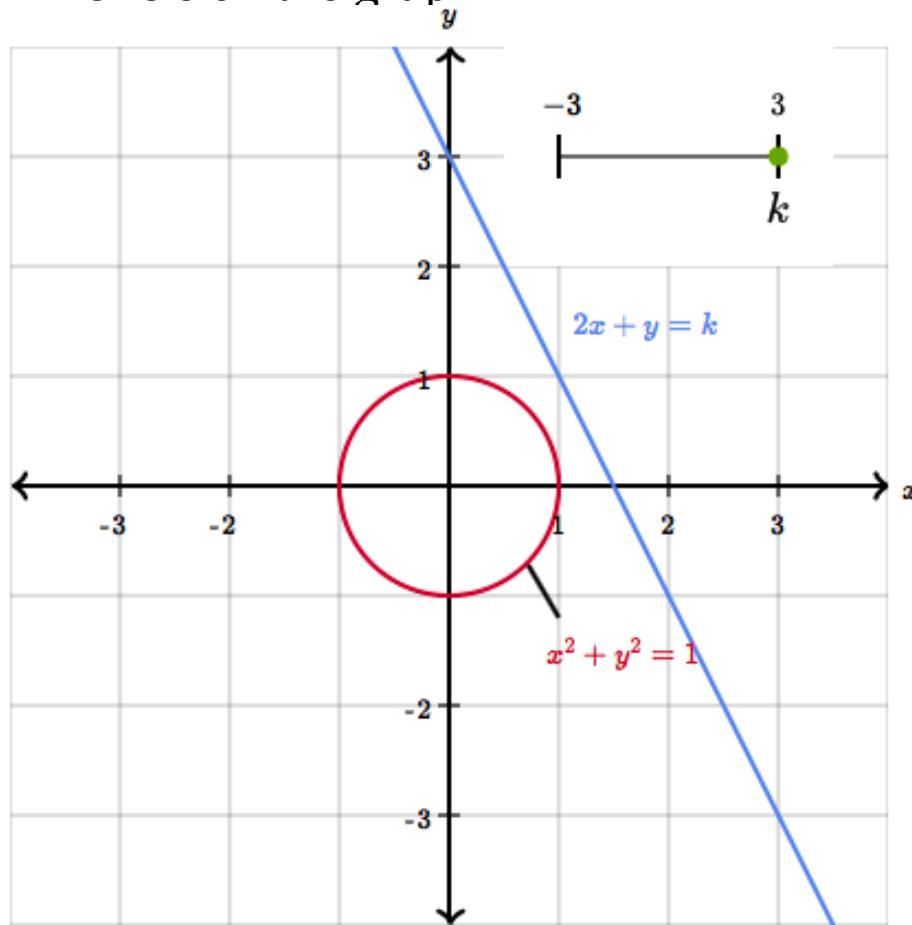
$$x^2 + y^2 = 1$$

Constraint

For which point (x, y) on the unit circle is the value $2x + y$ biggest?

Lagrange Multiplier

1. First draw the graph of $f(x,y)$ which looks like slanted plane since f is linear.
2. Then project the circle $x^2 + y^2 = 1$ from xy plane vertically onto the graph of f
3. Maximum we are seeking corresponds with highest point of this projected circle on the graph



Lagrange Multiplier

$$\nabla f(x_0, y_0) = \lambda_0 \nabla g(x_0, y_0)$$

Here, λ_0 represents some constant. Some authors use a negative constant, $-\lambda_0$, but I personally prefer a positive constant, as it gives a cleaner interpretation of λ_0 down the road.

Let's see what this looks like in our example where $f(x, y) = 2x + y$ and $g(x, y) = x^2 + y^2$. The gradient of f is

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial}{\partial x}(2x + y) \\ \frac{\partial}{\partial y}(2x + y) \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

and the gradient of g is

$$\nabla g(x, y) = \begin{bmatrix} \frac{\partial}{\partial x}(x^2 + y^2 - 1) \\ \frac{\partial}{\partial y}(x^2 + y^2 - 1) \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

Lagrange Multiplier

Therefore, the tangency condition ends up looking like this:

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} = \lambda_0 \begin{bmatrix} 2x_0 \\ 2y_0 \end{bmatrix}$$

$$x_0^2 + y_0^2 = 1$$

$$2 = 2\lambda_0 x_0$$

$$1 = 2\lambda_0 y_0$$



Three Equations and
three unknowns

$$x_0 = \frac{1}{\lambda_0}$$

$$y_0 = \frac{1}{2\lambda_0}$$



$$\left(\frac{1}{\lambda_0}\right)^2 + \left(\frac{1}{2\lambda_0}\right)^2 = 1$$

$$\frac{1}{\lambda_0^2} + \frac{1}{4\lambda_0^2} = 1$$



$$\pm\sqrt{\frac{5}{4}} = \lambda_0$$

$$\frac{\pm\sqrt{5}}{2} = \lambda_0$$

Lagrange Multiplier

$$\begin{aligned}(x_0, y_0) &= \left(\frac{1}{\lambda_0}, \frac{1}{2\lambda_0} \right) \\ &= \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right) \quad \text{or} \quad \left(\frac{-2}{\sqrt{5}}, \frac{-1}{\sqrt{5}} \right)\end{aligned}$$

We can see which of these is a maximum point and which is a minimum point by plugging these solutions into $f(x, y)$ and seeing which is bigger.

$$f \left(\frac{2}{\sqrt{5}}, \frac{1}{\sqrt{5}} \right) = 2 \frac{2}{\sqrt{5}} + \frac{1}{\sqrt{5}}$$

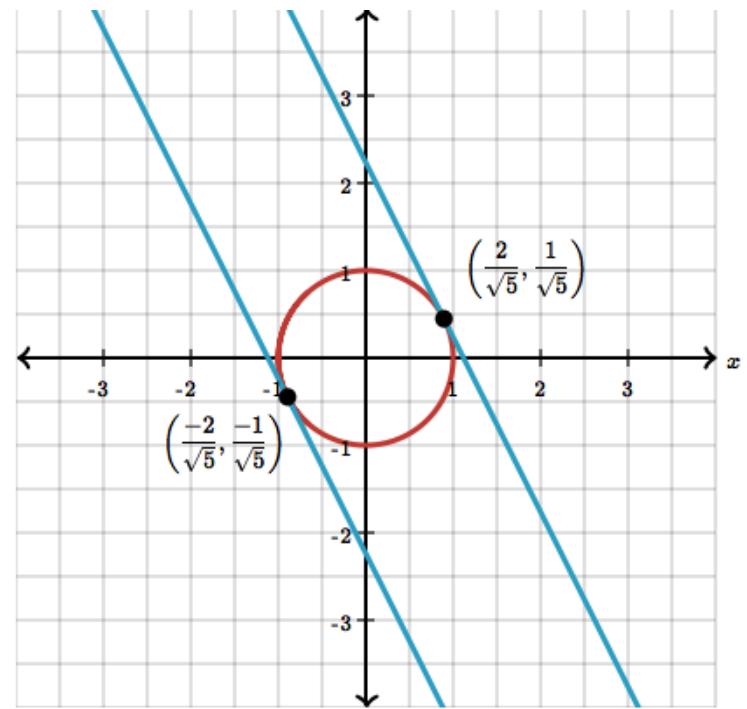
$$= \frac{5}{\sqrt{5}}$$

$$= \sqrt{5} \quad \leftarrow \text{Maximum}$$

$$f \left(-\frac{2}{\sqrt{5}}, -\frac{1}{\sqrt{5}} \right) = 2 \frac{-2}{\sqrt{5}} + \frac{-1}{\sqrt{5}}$$

$$= \frac{-5}{\sqrt{5}}$$

$$= -\sqrt{5} \quad \leftarrow \text{Minimum}$$



Method of Steepest Descent

- Steepest Descent is an iterative optimization method used to minimize an objective function by moving in the direction of steepest descent.

Step 1. Start with an arbitrary initial point X_1

Step 2. The direction of steepest descent is $-\nabla f(X_i)$

Step 3. Determine the optimal step length $t(\tau)$

$$X_{i+1} = X_i - \tau \nabla f(X_i)$$

Step 4. Test the new point X_{i+1} , for optimality.

If X_{i+1} is optimum, stop the process.

Otherwise set $i = i + 1$, go to step 2

Method of Steepest Descent

Minimize $Z = x^2 + 3y^2$ starting from (2,2)

In this case, $t = \frac{x^2 + 9y^2}{2x^2 + 54y^2}$

j	x(j)	y(j)	f(x(j),y(j))
1	2.000000	2.000000	16.000000
2	1.285714	-0.142857	1.714286
3	0.214286	0.214286	0.183673
4	0.137755	-0.015306	0.019679
5	0.022959	0.022959	0.002108
6	0.014759	-0.001640	0.000226
7	0.002460	0.002460	0.000024
8	0.001581	-0.000176	0.000003

The value of j is 8 and x(8) and y(8) are 0.001581 and -0.000176



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 11

Agenda

- Mathematical Induction
 - Examples of Proof by Mathematical Induction
 - Mistaken Proofs by Mathematical Induction
 - Guidelines for Proofs by Mathematical Induction
 - Strong Induction
 - Well-Ordering Property
 - Example Proofs using Strong Induction
-

Mathematical Induction

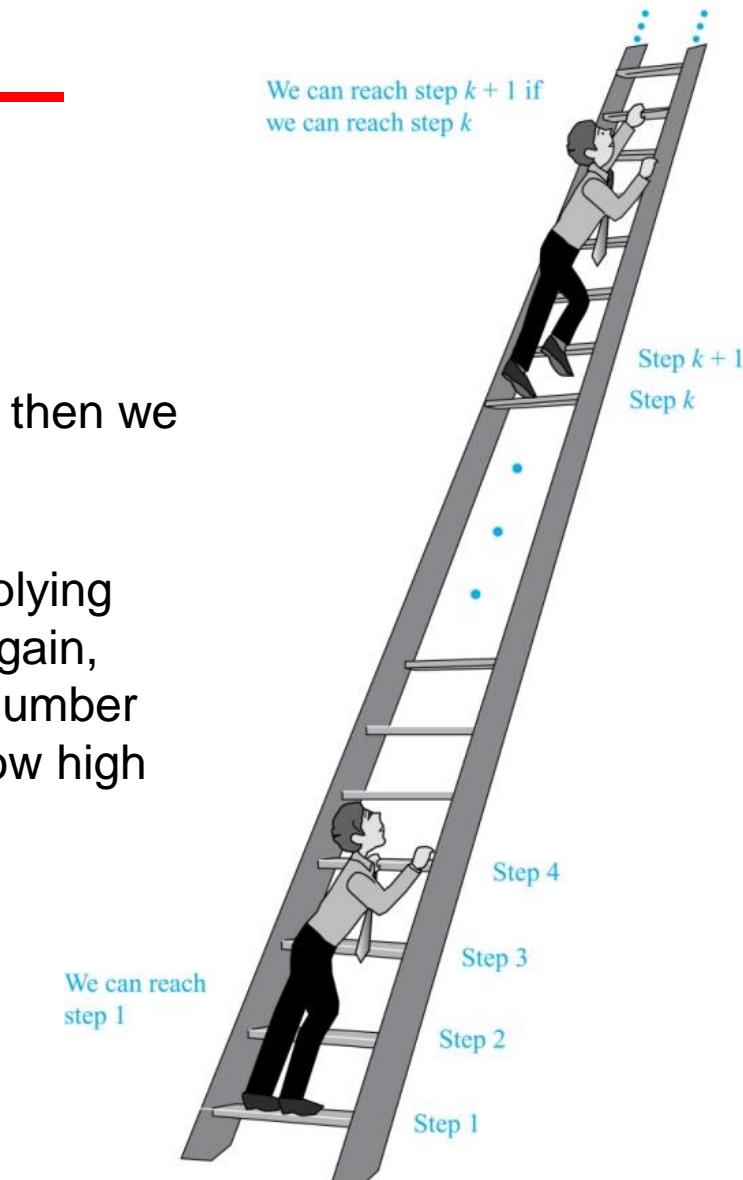
Climbing an Infinite Ladder

Suppose we have an infinite ladder:

1. We can reach the first rung of the ladder.
2. If we can reach a particular rung of the ladder, then we can reach the next rung.

From (1), we can reach the first rung. Then by applying (2), we can reach the second rung. Applying (2) again, the third rung. And so on. We can apply (2) any number of times to reach any particular rung, no matter how high up.

This example motivates proof by mathematical induction.



Principle of Mathematical Induction

Principle of Mathematical Induction: To prove that $P(n)$ is true for all positive integers n , we complete these steps:

- *Basis Step:* Show that $P(1)$ is true.
- *Inductive Step:* Show that $P(k) \rightarrow P(k + 1)$ is true for all positive integers k .

To complete the inductive step, assuming the *inductive hypothesis* that $P(k)$ holds for an arbitrary integer k , show that $P(k + 1)$ must be true.

Climbing an Infinite Ladder Example:

- BASIS STEP: By (1), we can reach rung 1.
- INDUCTIVE STEP: Assume the inductive hypothesis that we can reach rung k . Then by (2), we can reach rung $k + 1$.

Hence, $P(k) \rightarrow P(k + 1)$ is true for all positive integers k . We can reach every rung on the ladder.

Important Points About Using Mathematical Induction

- Mathematical induction can be expressed as the rule of inference

$$(P(1) \wedge \forall k (P(k) \rightarrow P(k + 1))) \rightarrow \forall n P(n),$$

where the domain is the set of positive integers.

- In a proof by mathematical induction, we don't assume that $P(k)$ is true for all positive integers! We show that if we assume that $P(k)$ is true, then $P(k + 1)$ must also be true.
- Proofs by mathematical induction do not always start at the integer 1. In such a case, the basis step begins at a starting point b where b is an integer. We will see examples of this soon.

Validity of Mathematical Induction

- Mathematical induction is valid because of the well ordering property, which states that every nonempty subset of the set of positive integers has a least element. Here is the proof:
 - Suppose that $P(1)$ holds and $P(k) \rightarrow P(k + 1)$ is true for all positive integers k .
 - Assume there is at least one positive integer n for which $P(n)$ is false. Then the set S of positive integers for which $P(n)$ is false is nonempty.
 - By the well-ordering property, S has a least element, say m .
 - We know that m can not be 1 since $P(1)$ holds.
 - Since m is positive and greater than 1, $m - 1$ must be a positive integer. Since $m - 1 < m$, it is not in S , so $P(m - 1)$ must be true.
 - But then, since the conditional $P(k) \rightarrow P(k + 1)$ for every positive integer k holds, $P(m)$ must also be true. This contradicts $P(m)$ being false.
 - Hence, $P(n)$ must be true for every positive integer n .

Proving a Summation Formula by Mathematical Induction

Example: Show that: $\sum_{i=1}^n i = \frac{n(n+1)}{2}$

Solution:

- BASIS STEP: $P(1)$ is true since $1(1 + 1)/2 = 1$.
- INDUCTIVE STEP: Assume true for $P(k)$.

The inductive hypothesis is $\sum_{i=1}^k i = \frac{k(k+1)}{2}$

Under this assumption,

$$\begin{aligned}1 + 2 + \dots + k + (k + 1) &= \frac{k(k + 1)}{2} + (k + 1) \\&= \frac{k(k + 1) + 2(k + 1)}{2} \\&= \frac{(k + 1)(k + 2)}{2}\end{aligned}$$

Note: Once we have this conjecture, mathematical induction can be used to prove it correct.

Conjecturing and Proving Correct a Summation Formula



Example: Conjecture and prove correct a formula for the sum of the first n positive odd integers. Then prove your conjecture.

Solution: We have: $1 = 1$, $1 + 3 = 4$, $1 + 3 + 5 = 9$, $1 + 3 + 5 + 7 = 16$, $1 + 3 + 5 + 7 + 9 = 25$.

- We can conjecture that the sum of the first n positive odd integers is n^2 ,

$$1 + 3 + 5 + \dots + (2n - 1) + (2n + 1) = n^2.$$

- We prove the conjecture is proved correct with mathematical induction.
- BASIS STEP: $P(1)$ is true since $1^2 = 1$.
- INDUCTIVE STEP: $P(k) \rightarrow P(k + 1)$ for every positive integer k .

Assume the inductive hypothesis holds and then show that $P(k + 1)$ holds has well.

Inductive Hypothesis: $1 + 3 + 5 + \dots + (2k - 1) = k^2$

- So, assuming $P(k)$, it follows that:

$$\begin{aligned}1 + 3 + 5 + \dots + (2k - 1) + (2k + 1) &= [1 + 3 + 5 + \dots + (2k - 1)] + (2k + 1) \\&= k^2 + (2k + 1) \quad (\text{by the inductive hypothesis}) \\&= k^2 + 2k + 1 \\&= (k + 1)^2\end{aligned}$$

- Hence, we have shown that $P(k + 1)$ follows from $P(k)$. Therefore the sum of the first n positive odd integers is n^2 .

Proving Inequalities

Example: Use mathematical induction to prove that $n < 2^n$ for all positive integers n .

Solution: Let $P(n)$ be the proposition that $n < 2^n$.

- BASIS STEP: $P(1)$ is true since $1 < 2^1 = 2$.
- INDUCTIVE STEP: Assume $P(k)$ holds, i.e., $k < 2^k$, for an arbitrary positive integer k .
- Must show that $P(k + 1)$ holds. Since by the inductive hypothesis, $k < 2^k$, it follows that:

$$k + 1 < 2^k + 1 \leq 2^k + 2^k = 2 \cdot 2^k = 2^{k+1}$$

Therefore $n < 2^n$ holds for all positive integers n .

Number of Subsets of a Finite Set

Example: Use mathematical induction to show that if S is a finite set with n elements, where n is a nonnegative integer, then S has 2^n subsets.

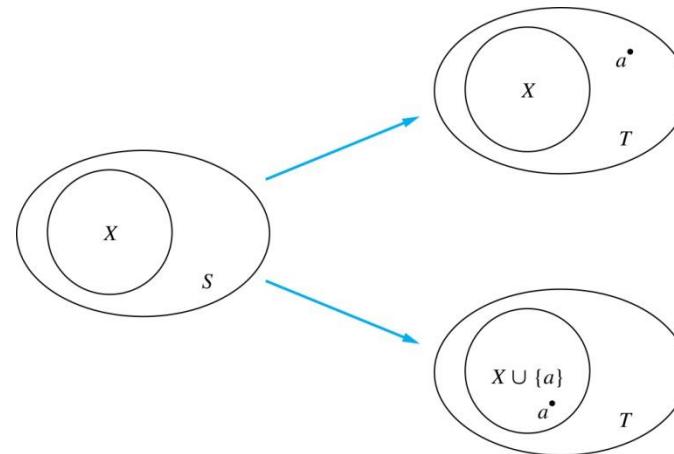
Solution: Let $P(n)$ be the proposition that a set with n elements has 2^n subsets.

- Basis Step: $P(0)$ is true, because the empty set has only itself as a subset and $2^0 = 1$.
- Inductive Step: Assume $P(k)$ is true for an arbitrary nonnegative integer k .

Number of Subsets of a Finite Set

Inductive Hypothesis: For an arbitrary nonnegative integer k , every set with k elements has 2^k subsets.

- Let T be a set with $k + 1$ elements. Then $T = S \cup \{a\}$, where $a \in T$ and $S = T - \{a\}$. Hence $|S| = k$.
- For each subset X of S , there are exactly two subsets of T , i.e., X and $X \cup \{a\}$.



- By the inductive hypothesis S has 2^k subsets. Since there are two subsets of T for each subset of S , the number of subsets of T is $2 \cdot 2^k = 2^{k+1}$.

An Incorrect “Proof” by Mathematical Induction



Example: Let $P(n)$ be the statement that every set of n lines in the plane, no two of which are parallel, meet in a common point.

Here is a “proof” that $P(n)$ is true for all positive integers $n \geq 2$.

- BASIS STEP: The statement $P(2)$ is true because any two lines in the plane that are not parallel meet in a common point.
- INDUCTIVE STEP: The inductive hypothesis is the statement that $P(k)$ is true for the positive integer $k \geq 2$, i.e., every set of k lines in the plane, no two of which are parallel, meet in a common point.
- We must show that if $P(k)$ holds, then $P(k + 1)$ holds, i.e., if every set of k lines in the plane, no two of which are parallel, $k \geq 2$, meet in a common point, then every set of $k + 1$ lines in the plane, no two of which are parallel, meet in a common point.

An Incorrect “Proof” by Mathematical Induction



Inductive Hypothesis: Every set of k lines in the plane, where $k \geq 2$, no two of which are parallel, meet in a common point.

- Consider a set of $k + 1$ distinct lines in the plane, no two parallel. By the inductive hypothesis, the first k of these lines must meet in a common point p_1 . By the inductive hypothesis, the last k of these lines meet in a common point p_2 .
- If p_1 and p_2 are different points, all lines containing both of them must be the same line since two points determine a line. This contradicts the assumption that the lines are distinct. Hence, $p_1 = p_2$ lies on all $k + 1$ distinct lines, and therefore $P(k + 1)$ holds. Assuming that $k \geq 2$, distinct lines meet in a common point, then every $k + 1$ lines meet in a common point.
- There must be an error in this proof since the conclusion is absurd. But where is the error?
 - **Answer:** $P(k) \rightarrow P(k + 1)$ only holds for $k \geq 3$. It is not the case that $P(2)$ implies $P(3)$. The first two lines must meet in a common point p_1 and the second two must meet in a common point p_2 . They do not have to be the same point since only the second line is common to both sets of lines.

Guidelines: Mathematical Induction Proofs

Template for Proofs by Mathematical Induction

1. Express the statement that is to be proved in the form “for all $n \geq b$, $P(n)$ ” for a fixed integer b .
2. Write out the words “Basis Step.” Then show that $P(b)$ is true, taking care that the correct value of b is used. This completes the first part of the proof.
3. Write out the words “Inductive Step.”
4. State, and clearly identify, the inductive hypothesis, in the form “assume that $P(k)$ is true for an arbitrary fixed integer $k \geq b$.”
5. State what needs to be proved under the assumption that the inductive hypothesis is true. That is, write out what $P(k + 1)$ says.
6. Prove the statement $P(k + 1)$ making use of the assumption $P(k)$. Be sure that your proof is valid for all integers k with $k \geq b$, taking care that the proof works for small values of k , including $k = b$.
7. Clearly identify the conclusion of the inductive step, such as by saying “this completes the inductive step.”
8. After completing the basis step and the inductive step, state the conclusion, namely that by mathematical induction, $P(n)$ is true for all integers n with $n \geq b$.

Strong Induction

- *Strong Induction:* To prove that $P(n)$ is true for all positive integers n , where $P(n)$ is a propositional function, complete two steps:
 - *Basis Step:* Verify that the proposition $P(1)$ is true.
 - *Inductive Step:* Show the conditional statement $[P(1) \wedge P(2) \wedge \dots \wedge P(k)] \rightarrow P(k + 1)$ holds for all positive integers k .

Strong Induction is sometimes called the *second principle of mathematical induction* or *complete induction*.



Strong Induction and the Infinite Ladder

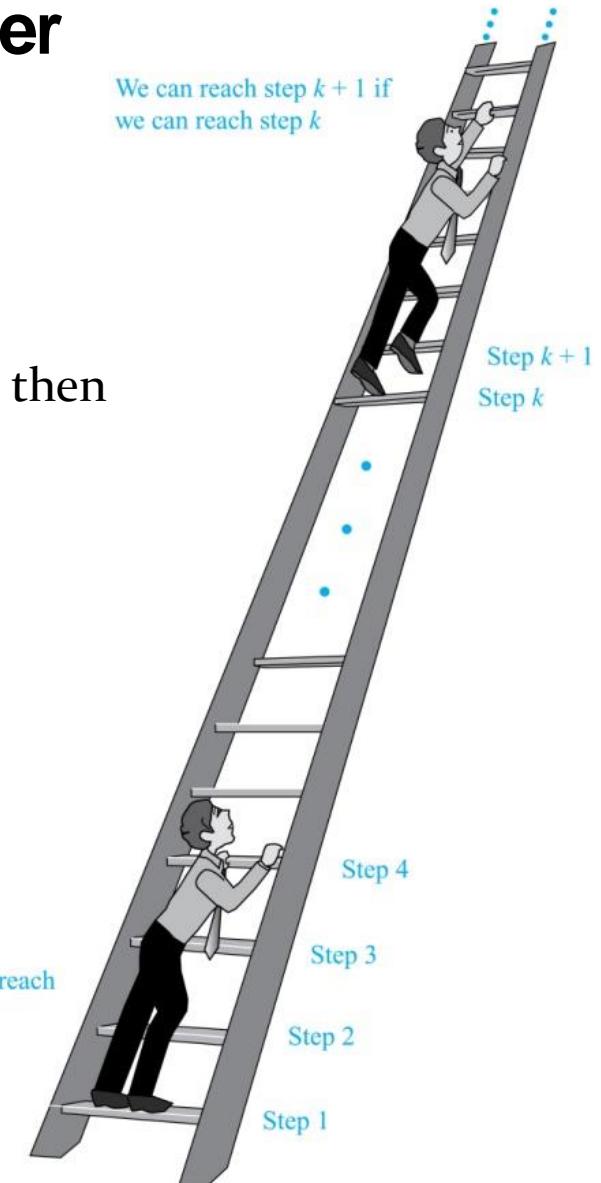
Strong induction tells us that we can reach all rungs if:

1. We can reach the first rung of the ladder.
2. For every integer k , if we can reach the first k rungs, then we can reach the $(k + 1)$ st rung.

To conclude that we can reach every rung by strong induction:

- BASIS STEP: $P(1)$ holds
- INDUCTIVE STEP: Assume $P(1) \wedge P(2) \wedge \dots \wedge P(k)$ holds for an arbitrary integer k , and show that $P(k + 1)$ must also hold.

We will have then shown by strong induction that for every positive integer n , $P(n)$ holds, i.e., we can reach the n th rung of the ladder.



Proof using Strong Induction

Example: Suppose we can reach the first and second rungs of an infinite ladder, and we know that if we can reach a rung, then we can reach two rungs higher. Prove that we can reach every rung.

(Try this with mathematical induction.)

Solution: Prove the result using strong induction.

- **BASIS STEP:** We can reach the first step.
- **INDUCTIVE STEP:** The inductive hypothesis is that we can reach the first k rungs, for any $k \geq 2$. We can reach the $(k + 1)$ st rung since we can reach the $(k - 1)$ st rung by the inductive hypothesis.
- Hence, we can reach all rungs of the ladder.

Well-Ordering Property

- *Well-ordering property:* Every nonempty set of nonnegative integers has a least element.
- The well-ordering property is one of the axioms of the positive integers.
- The well-ordering property can be used directly in proofs, as the next example illustrates.
- The well-ordering property can be generalized.
 - **Definition:** A set is *well ordered* if every subset has a least element.
 - \mathbb{N} is well ordered under \leq .
 - The set of finite strings over an alphabet using lexicographic ordering is well ordered.

Well-Ordering Property

Example: Use the well-ordering property to prove the division algorithm, which states that if a is an integer and d is a positive integer, then there are unique integers q and r with $0 \leq r < d$, such that $a = dq + r$.

Solution: Let S be the set of nonnegative integers of the form $a - dq$, where q is an integer. The set is nonempty since $-dq$ can be made as large as needed.

- By the well-ordering property, S has a least element $r = a - dq_0$. The integer r is nonnegative. It also must be the case that $r < d$. If it were not, then there would be a smaller nonnegative element in S , namely,
$$a - d(q_0 + 1) = a - dq_0 - d = r - d > 0.$$
- Therefore, there are integers q and r with $0 \leq r < d$.

Which Form of Induction Should Be Used?



- We can always use strong induction instead of mathematical induction. But there is no reason to use it if it is simpler to use mathematical induction.
- In fact, the principles of mathematical induction, strong induction, and the well-ordering property are all equivalent.
- Sometimes it is clear how to proceed using one of the three methods, but not the other two.

Completion of the proof of the Fundamental Theorem of Arithmetic

Example: Show that if n is an integer greater than 1, then n can be written as the product of primes.

Solution: Let $P(n)$ be the proposition that n can be written as a product of primes.

- BASIS STEP: $P(2)$ is true since 2 itself is prime.
- INDUCTIVE STEP: The inductive hypothesis is $P(j)$ is true for all integers j with $2 \leq j \leq k$. To show that $P(k + 1)$ must be true under this assumption, two cases need to be considered:
 - If $k + 1$ is prime, then $P(k + 1)$ is true.
 - Otherwise, $k + 1$ is composite and can be written as the product of two positive integers a and b with $2 \leq a \leq b < k + 1$. By the inductive hypothesis a and b can be written as the product of primes and therefore $k + 1$ can also be written as the product of those primes.

Hence, it has been shown that every integer greater than 1 can be written as the product of primes.

Proof using Strong Induction

Example: Prove that every amount of postage of 12 cents or more can be formed using just 4-cent and 5-cent stamps.

Solution: Let $P(n)$ be the proposition that postage of n cents can be formed using 4-cent and 5-cent stamps.

- BASIS STEP: $P(12)$, $P(13)$, $P(14)$, and $P(15)$ hold.
 - $P(12)$ uses three 4-cent stamps.
 - $P(13)$ uses two 4-cent stamps and one 5-cent stamp.
 - $P(14)$ uses one 4-cent stamp and two 5-cent stamps.
 - $P(15)$ uses three 5-cent stamps.
- INDUCTIVE STEP: The inductive hypothesis states that $P(j)$ holds for $12 \leq j \leq k$, where $k \geq 15$. Assuming the inductive hypothesis, it can be shown that $P(k + 1)$ holds.
- Using the inductive hypothesis, $P(k - 3)$ holds since $k - 3 \geq 12$. To form postage of $k + 1$ cents, add a 4-cent stamp to the postage for $k - 3$ cents.

Hence, $P(n)$ holds for all $n \geq 12$.

Proof of Same Example using Mathematical Induction

Example: Prove that every amount of postage of 12 cents or more can be formed using just 4-cent and 5-cent stamps.

Solution: Let $P(n)$ be the proposition that postage of n cents can be formed using 4-cent and 5-cent stamps.

- **BASIS STEP:** Postage of 12 cents can be formed using three 4-cent stamps.
- **INDUCTIVE STEP:** The inductive hypothesis $P(k)$ for any positive integer k is that postage of k cents can be formed using 4-cent and 5-cent stamps. To show $P(k + 1)$ where $k \geq 12$, we consider two cases:
 - If at least one 4-cent stamp has been used, then a 4-cent stamp can be replaced with a 5-cent stamp to yield a total of $k + 1$ cents.
 - Otherwise, no 4-cent stamp have been used and at least three 5-cent stamps were used. Three 5-cent stamps can be replaced by four 4-cent stamps to yield a total of $k + 1$ cents.

Hence, $P(n)$ holds for all $n \geq 12$.



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 12

Agenda

- Recursively Defined Functions
- Recursively Defined Sets and Structures
- Structural Induction
- Generalized Induction

Recursively Defined Functions

Definition: A *recursive or inductive definition* of a function consists of two steps.

- BASIS STEP: Specify the value of the function at zero.
- RECURSIVE STEP: Give a rule for finding its value at an integer from its values at smaller integers.
- A function $f(n)$ is the same as a sequence a_0, a_1, \dots , where a_i , where $f(i) = a_i$.

Recursively Defined Functions

Example: Suppose f is defined by:

$$f(0) = 3,$$

$$f(n + 1) = 2f(n) + 3$$

Find $f(1), f(2), f(3), f(4)$

Solution:

- $f(1) = 2f(0) + 3 = 2 \cdot 3 + 3 = 9$
- $f(2) = 2f(1) + 3 = 2 \cdot 9 + 3 = 21$
- $f(3) = 2f(2) + 3 = 2 \cdot 21 + 3 = 45$
- $f(4) = 2f(3) + 3 = 2 \cdot 45 + 3 = 93$

Example: Give a recursive definition of the factorial function $n!$:

Solution:

$$f(0) = 1$$

$$f(n + 1) = (n + 1) \cdot f(n)$$

Recursively Defined Functions

Example: Give a recursive definition of:

$$\sum_{k=0}^n a_k.$$

Solution: The first part of the definition is

$$\sum_{k=0}^0 a_k = a_0.$$

The second part is

$$\sum_{k=0}^{n+1} a_k = \left(\sum_{k=0}^n a_k \right) + a_{n+1}.$$

Fibonacci Numbers

Example : The Fibonacci numbers are defined as follows:

$$f_0 = 0$$

$$f_1 = 1$$

$$f_n = f_{n-1} + f_{n-2}$$

Find f_2, f_3, f_4, f_5 .

- $f_2 = f_1 + f_0 = 1 + 0 = 1$
- $f_3 = f_2 + f_1 = 1 + 1 = 2$
- $f_4 = f_3 + f_2 = 2 + 1 = 3$
- $f_5 = f_4 + f_3 = 3 + 2 = 5$

Fibonacci
(1170- 1250)



Fibonacci Numbers

Example 4: Show that whenever $n \geq 3$, $f_n > \alpha^{n-2}$, where $\alpha = (1 + \sqrt{5})/2$.

Solution: Let $P(n)$ be the statement $f_n > \alpha^{n-2}$. Use strong induction to show that $P(n)$ is true whenever $n \geq 3$.

- BASIS STEP: $P(3)$ holds since $\alpha < 2 = f_3$
- $P(4)$ holds since $\alpha^2 = (3 + \sqrt{5})/2 < 3 = f_4$.
- INDUCTIVE STEP: Assume that $P(j)$ holds, i.e., $f_j > \alpha^{j-2}$ for all integers j with $3 \leq j \leq k$, where $k \geq 4$. Show that $P(k+1)$ holds, i.e., $f_{k+1} > \alpha^{k-1}$.
- Since $\alpha^2 = \alpha + 1$ (because α is a solution of $x^2 - x - 1 = 0$),

$$\alpha^{k-1} = \alpha^2 \cdot \alpha^{k-3} = (\alpha + 1) \cdot \alpha^{k-3} = \alpha \cdot \alpha^{k-3} + 1 \cdot \alpha^{k-3} = \alpha^{k-2} + \alpha^{k-3}$$

- By the inductive hypothesis, because $k \geq 4$ we have

$$f_{k-1} > \alpha^{k-3}, \quad f_k > \alpha^{k-2}.$$

- Therefore, it follows that

$$f_{k+1} = f_k + f_{k-1} > \alpha^{k-2} + \alpha^{k-3} = \alpha^{k-1}.$$

- Hence, $P(k+1)$ is true.

Why does this equality hold?

Lamé's Theorem

Gabriel Lamé
(1795-1870)



Lamé's Theorem: Let a and b be positive integers with $a \geq b$. Then the number of divisions used by the Euclidian algorithm to find $\gcd(a,b)$ is less than or equal to five times the number of decimal digits in b .

Proof: When we use the Euclidian algorithm to find $\gcd(a,b)$ with $a \geq b$,

- n divisions are used to obtain (with $a = r_0, b = r_1$):

$$\begin{aligned}r_0 &= r_1 q_1 + r_2 & 0 \leq r_2 < r_1, \\r_1 &= r_2 q_2 + r_3 & 0 \leq r_3 < r_2, \\&\vdots \\r_{n-2} &= r_{n-1} q_{n-1} + r_n & 0 \leq r_n < r_{n-1}, \\r_{n-1} &= r_n q_n.\end{aligned}$$

- Since each quotient q_1, q_2, \dots, q_{n-1} is at least 1 and $q_n \geq 2$:

$$\begin{aligned}r_n &\geq 1 = f_2, \\r_{n-1} &\geq 2 \quad r_n \geq 2 \quad f_2 = f_3, \\r_{n-2} &\geq r_{n-1} + r_n \geq f_3 + f_2 = f_4, \\&\vdots \\r_2 &\geq r_3 + r_4 \geq f_{n-1} + f_{n-2} = f_n, \\b = r_1 &\geq r_2 + r_3 \geq f_n + f_{n-1} = f_{n+1}.\end{aligned}$$

Lamé's Theorem

- It follows that if n divisions are used by the Euclidian algorithm to find $\gcd(a,b)$ with $a \geq b$, then $b \geq f_{n+1}$. By Example 4, $f_{n+1} > \alpha^{n-1}$, for $n > 2$, where $\alpha = (1 + \sqrt{5})/2$. Therefore, $b > \alpha^{n-1}$.
- Because $\log_{10} \alpha \approx 0.208 > 1/5$, $\log_{10} b > (n-1) \log_{10} \alpha > (n-1)/5$. Hence,

$$n-1 < 5 \cdot \log_{10} b.$$

- Suppose that b has k decimal digits. Then $b < 10^k$ and $\log_{10} b < k$. It follows that $n - 1 < 5k$ and since k is an integer, $n \leq 5k$.
- As a consequence of Lamé's Theorem, $O(\log b)$ divisions are used by the Euclidian algorithm to find $\gcd(a,b)$ whenever $a > b$.
 - By Lamé's Theorem, the number of divisions needed to find $\gcd(a,b)$ with $a > b$ is less than or equal to $5 (\log_{10} b + 1)$ since the number of decimal digits in b (which equals $\lfloor \log_{10} b \rfloor + 1$) is less than or equal to $\log_{10} b + 1$.

Lamé's Theorem was the first result in computational complexity

Recursively Defined Sets and Structures



Recursive definitions of sets have two parts:

- The *basis step* specifies an initial collection of elements.
- The *recursive step* gives the rules for forming new elements in the set from those already known to be in the set.
- Sometimes the recursive definition has an *exclusion rule*, which specifies that the set contains nothing other than those elements specified in the basis step and generated by applications of the rules in the recursive step.
- We will always assume that the exclusion rule holds, even if it is not explicitly mentioned.
- We will later develop a form of induction, called *structural induction*, to prove results about recursively defined sets.

Recursively Defined Sets and Structures



Example : Subset of Integers S :

BASIS STEP: $3 \in S$.

RECURSIVE STEP: If $x \in S$ and $y \in S$, then $x + y$ is in S .

- Initially 3 is in S , then $3 + 3 = 6$, then $3 + 6 = 9$, etc.

Example: The natural numbers \mathbf{N} .

BASIS STEP: $0 \in \mathbf{N}$.

RECURSIVE STEP: If n is in \mathbf{N} , then $n + 1$ is in \mathbf{N} .

- Initially 0 is in S , then $0 + 1 = 1$, then $1 + 1 = 2$, etc.

Strings

Definition: The set Σ^* of *strings* over the alphabet Σ :

BASIS STEP: $\lambda \in \Sigma^*$ (λ is the empty string)

RECURSIVE STEP: If w is in Σ^* and x is in Σ ,
then $wx \in \Sigma^*$.

Example: If $\Sigma = \{0,1\}$, the strings in Σ^* are the set of
all bit strings, $\lambda, 0, 1, 00, 01, 10, 11$, etc.

Example: If $\Sigma = \{a,b\}$, show that aab is in Σ^* .

- Since $\lambda \in \Sigma^*$ and $a \in \Sigma$, $a \in \Sigma^*$.
- Since $a \in \Sigma^*$ and $a \in \Sigma$, $aa \in \Sigma^*$.
- Since $aa \in \Sigma^*$ and $b \in \Sigma$, $aab \in \Sigma^*$.

String Concatenation

Definition: Two strings can be combined via the operation of *concatenation*. Let Σ be a set of symbols and Σ^* be the set of strings formed from the symbols in Σ . We can define the concatenation of two strings, denoted by \cdot , recursively as follows.

BASIS STEP: If $w \in \Sigma^*$, then $w \cdot \lambda = w$.

RECURSIVE STEP: If $w_1 \in \Sigma^*$ and $w_2 \in \Sigma^*$ and $x \in \Sigma$, then
 $w_1 \cdot (w_2 x) = (w_1 \cdot w_2)x$.

- Often $w_1 \cdot w_2$ is written as $w_1 w_2$.
- If $w_1 = abra$ and $w_2 = cadabra$, the concatenation $w_1 w_2 = abracadabra$.

Length of a String

Example: Give a recursive definition of $l(w)$, the length of the string w .

Solution: The length of a string can be recursively defined by:

$$l(\lambda) = 0;$$

$$l(wx) = l(w) + 1 \text{ if } w \in \Sigma^* \text{ and } x \in \Sigma.$$

Balanced Parentheses

Example: Give a recursive definition of the set of balanced parentheses P .

Solution:

BASIS STEP: $() \in P$

RECURSIVE STEP: If $w \in P$, then $(w) \in P$, $(w) \in P$ and $w() \in P$.

- Show that $((())()$ is in P .
- Why is $))((()$ not in P ?

Well-Formed Formulae in Propositional Logic



Definition: The set of *well-formed formulae* in propositional logic involving T, F, propositional variables, and operators from the set $\{\neg, \wedge, \vee, \rightarrow, \leftrightarrow\}$.

BASIS STEP: T, F, and s , where s is a propositional variable, are well-formed formulae.

RECURSIVE STEP: If E and F are well formed formulae, then $(\neg E)$, $(E \wedge F)$, $(E \vee F)$, $(E \rightarrow F)$, $(E \leftrightarrow F)$, are well-formed formulae.

Examples: $((p \vee q) \rightarrow (q \wedge F))$ is a well-formed formula.

$p q \wedge$ is not a well formed formula.

Rooted Trees

Definition: The set of *rooted trees*, where a rooted tree consists of a set of vertices containing a distinguished vertex called the *root*, and edges connecting these vertices, can be defined recursively by these steps:

BASIS STEP: A single vertex r is a rooted tree.

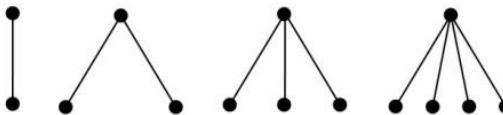
RECURSIVE STEP: Suppose that T_1, T_2, \dots, T_n are disjoint rooted trees with roots r_1, r_2, \dots, r_n , respectively. Then the graph formed by starting with a root r , which is not in any of the rooted trees T_1, T_2, \dots, T_n , and adding an edge from r to each of the vertices r_1, r_2, \dots, r_n , is also a rooted tree.

Building Up Rooted Trees

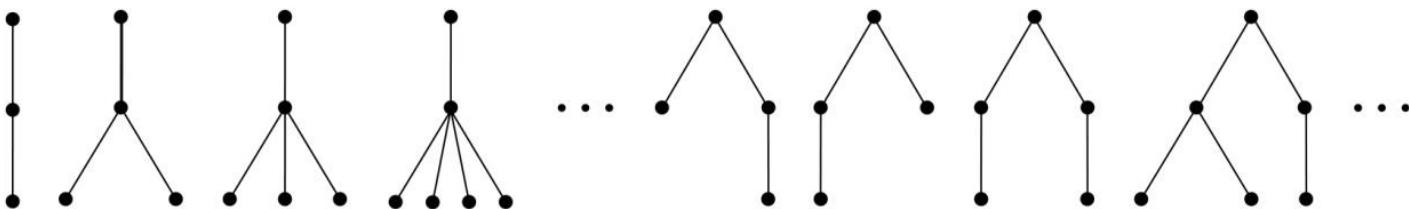
Basis step



Step 1



Step 2



Full Binary Trees

Definition: The set of *full binary trees* can be defined recursively by these steps.

BASIS STEP: There is a full binary tree consisting of only a single vertex r .

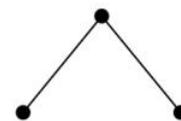
RECURSIVE STEP: If T_1 and T_2 are disjoint full binary trees, there is a full binary tree, denoted by $T_1 \cdot T_2$, consisting of a root r together with edges connecting the root to each of the roots of the left subtree T_1 and the right subtree T_2 .

Building Up Full Binary Trees

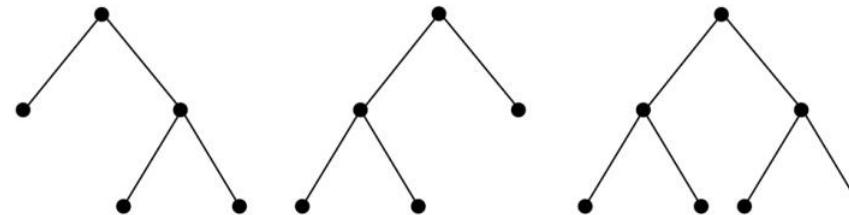
Basis step



Step 1



Step 2



Induction and Recursively Defined Sets



Example: Show that the set S defined by specifying that $3 \in S$ and that if $x \in S$ and $y \in S$, then $x + y$ is in S , is the set of all positive integers that are multiples of 3.

Solution: Let A be the set of all positive integers divisible by 3. To prove that $A = S$, show that A is a subset of S and S is a subset of A .

- $A \subset S$: Let $P(n)$ be the statement that $3n$ belongs to S .

BASIS STEP: $3 \cdot 1 = 3 \in S$, by the first part of recursive definition.

INDUCTIVE STEP: Assume $P(k)$ is true. By the second part of the recursive definition, if $3k \in S$, then since $3 \in S$, $3k + 3 = 3(k + 1) \in S$. Hence, $P(k + 1)$ is true.

- $S \subset A$:

BASIS STEP: $3 \in S$ by the first part of recursive definition, and $3 = 3 \cdot 1$.

INDUCTIVE STEP: The second part of the recursive definition adds $x + y$ to S , if both x and y are in S . If x and y are both in A , then both x and y are divisible by 3. By part (i) of Theorem 1 of Section 4.1, it follows that $x + y$ is divisible by 3.

- We used mathematical induction to prove a result about a recursively defined set. Next we study a more direct form induction for proving results about recursively defined sets.

Structural Induction

Definition: To prove a property of the elements of a recursively defined set, we use *structural induction*.

BASIS STEP: Show that the result holds for all elements specified in the basis step of the recursive definition.

RECURSIVE STEP: Show that if the statement is true for each of the elements used to construct new elements in the recursive step of the definition, the result holds for these new elements.

- The validity of structural induction can be shown to follow from the principle of mathematical induction.
-

Full Binary Trees

Definition: The *height* $h(T)$ of a full binary tree T is defined recursively as follows:

- **BASIS STEP:** The height of a full binary tree T consisting of only a root r is $h(T) = 0$.
- **RECURSIVE STEP:** If T_1 and T_2 are full binary trees, then the full binary tree $T = T_1 \cdot T_2$ has height $h(T) = 1 + \max(h(T_1), h(T_2))$.
- The number of vertices $n(T)$ of a full binary tree T satisfies the following recursive formula:
 - **BASIS STEP:** The number of vertices of a full binary tree T consisting of only a root r is $n(T) = 1$.
 - **RECURSIVE STEP:** If T_1 and T_2 are full binary trees, then the full binary tree $T = T_1 \cdot T_2$ has the number of vertices $n(T) = 1 + n(T_1) + n(T_2)$.

Structural Induction and Binary Trees

Theorem: If T is a full binary tree, then $n(T) \leq 2^{h(T)+1} - 1$.

Proof: Use structural induction.

- **BASIS STEP:** The result holds for a full binary tree consisting only of a root, $n(T) = 1$ and $h(T) = 0$. Hence, $n(T) = 1 \leq 2^{0+1} - 1 = 1$.
- **RECURSIVE STEP:** Assume $n(T_1) \leq 2^{h(T_1)+1} - 1$ and also $n(T_2) \leq 2^{h(T_2)+1} - 1$ whenever T_1 and T_2 are full binary trees.

$$\begin{aligned}
 n(T) &= 1 + n(T_1) + n(T_2) && (\text{by recursive formula of } n(T)) \\
 &\leq 1 + (2^{h(T_1)+1} - 1) + (2^{h(T_2)+1} - 1) && (\text{by inductive hypothesis}) \\
 &\leq 2 \cdot \max(2^{h(T_1)+1}, 2^{h(T_2)+1}) - 1 \\
 &= 2 \cdot 2^{\max(h(T_1), h(T_2))+1} - 1 && (\max(2^x, 2^y) = 2^{\max(x, y)}) \\
 &= 2 \cdot 2^{h(T)} - 1 && (\text{by recursive definition of } h(T)) \\
 &= 2^{h(T)+1} - 1
 \end{aligned}$$

Generalized Induction

- *Generalized induction* is used to prove results about sets other than the integers that have the well-ordering property.
- For example, consider an ordering on $\mathbb{N} \times \mathbb{N}$, ordered pairs of nonnegative integers. Specify that (x_1, y_1) is less than or equal to (x_2, y_2) if either $x_1 < x_2$, or $x_1 = x_2$ and $y_1 < y_2$. This is called the *lexicographic ordering*.
- Strings are also commonly ordered by a *lexicographic ordering*.
- The next example uses generalized induction to prove a result about ordered pairs from $\mathbb{N} \times \mathbb{N}$.

Generalized Induction

Example: Suppose that $a_{m,n}$ is defined for $(m,n) \in \mathbb{N} \times \mathbb{N}$ by $a_{0,0} = 0$ and

$$a_{m,n} = \begin{cases} a_{m-1,n} + 1 & \text{if } n = 0 \text{ and } m > 0 \\ a_{m,n-1} + n & \text{if } n > 0 \end{cases}.$$

Show that $a_{m,n} = m + n(n+1)/2$ is defined for all $(m,n) \in \mathbb{N} \times \mathbb{N}$.

Solution: Use generalized induction.

BASIS STEP: $a_{0,0} = 0 = 0 + (0 \cdot 1)/2$

INDUCTIVE STEP: Assume that $a_{m',n'} = m' + n'(n'+1)/2$ whenever (m',n') is less than (m,n) in the lexicographic ordering of $\mathbb{N} \times \mathbb{N}$.

- If $n = 0$, by the inductive hypothesis we can conclude

$$a_{m,n} = a_{m-1,n} + 1 = m - 1 + n(n+1)/2 + 1 = m + n(n+1)/2.$$

- If $n > 0$, by the inductive hypothesis we can conclude

$$a_{m,n} = a_{m,n-1} + n = m + n(n-1)/2 + n = m + n(n+1)/2.$$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 13

Agenda

- The Basics of Counting
 - The Pigeonhole Principle
 - The Generalized Pigeonhole Principle
 - Permutations
 - Combinations
-

Basic Counting Principles: The Product Rule

The Product Rule: A procedure can be broken down into a sequence of two tasks. There are n_1 ways to do the first task and n_2 ways to do the second task. Then there are $n_1 \cdot n_2$ ways to do the procedure.

Example: How many bit strings of length seven are there?

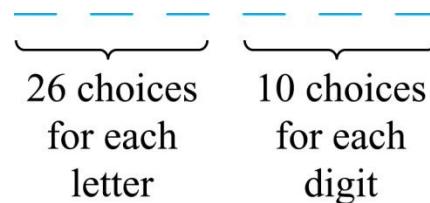
Solution: Since each of the seven bits is either a 0 or a 1, the answer is $2^7 = 128$.

The Product Rule

Example: How many different license plates can be made if each plate contains a sequence of three uppercase English letters followed by three digits?

Solution: By the product rule,

there are $26 \cdot 26 \cdot 26 \cdot 10 \cdot 10 \cdot 10 = 17,576,000$ different possible license plates.



Counting Functions

Counting Functions: How many functions are there from a set with m elements to a set with n elements?

Solution: Since a function represents a choice of one of the n elements of the codomain for each of the m elements in the domain, the product rule tells us that there are $n \cdot n \cdots n = n^m$ such functions.

Counting Functions

Counting One-to-One Functions: How many one-to-one functions are there from a set with m elements to one with n elements?

Solution: Suppose the elements in the domain are a_1, a_2, \dots, a_m . There are n ways to choose the value of a_1 and $n-1$ ways to choose a_2 , etc. The product rule tells us that there are $n(n-1)(n-2)\cdots(n-m+1)$ such functions.

Product Rule in Terms of Sets

- If A_1, A_2, \dots, A_m are finite sets, then the number of elements in the Cartesian product of these sets is the product of the number of elements of each set.
- The task of choosing an element in the Cartesian product $A_1 \times A_2 \times \dots \times A_m$ is done by choosing an element in A_1 , an element in A_2 , ..., and an element in A_m .
- By the product rule, it follows that:

$$|A_1 \times A_2 \times \dots \times A_m| = |A_1| \cdot |A_2| \cdot \dots \cdot |A_m|.$$

Basic Counting Principles: The Sum Rule

The Sum Rule: If a task can be done either in one of n_1 ways or in one of n_2 , where none of the set of n_1 ways is the same as any of the n_2 ways, then there are $n_1 + n_2$ ways to do the task.

Example: The mathematics department must choose either a student or a faculty member as a representative for a university committee. How many choices are there for this representative if there are 37 members of the mathematics faculty and 83 mathematics majors and no one is both a faculty member and a student.

Solution: By the sum rule it follows that there are $37 + 83 = 120$ possible ways to pick a representative.

The sum rule can be phrased in terms of sets. $|A \cup B| = |A| + |B|$ as long as A and B are disjoint sets.

- Or more generally, $|A_1 \cup A_2 \cup \dots \cup A_m| = |A_1| + |A_2| + \dots + |A_m|$
when $A_i \cap A_j = \emptyset$ for all i, j .

Combining the Sum and Product Rule

Example: Suppose statement labels in a programming language can be either a single letter or a letter followed by a digit. Find the number of possible labels.

Solution: Use the product rule.

$$26 + 26 \cdot 10 = 286$$

Basic Counting Principles: Subtraction Rule

Subtraction Rule: If a task can be done either in one of n_1 ways or in one of n_2 ways, then the total number of ways to do the task is $n_1 + n_2$ minus the number of ways to do the task that are common to the two different ways. Also known as, the *principle of inclusion-exclusion*:

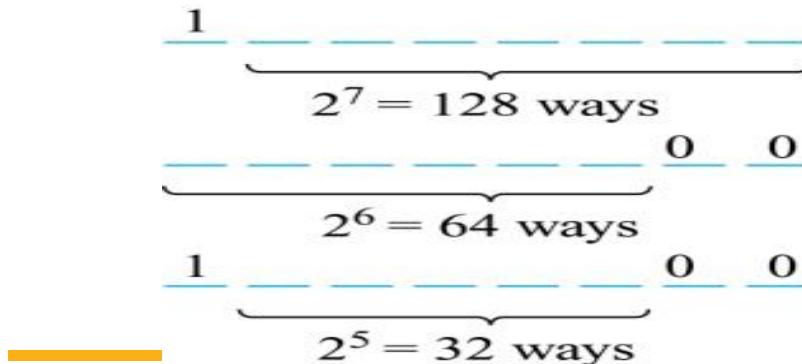
$$|A \cup B| = |A| + |B| - |A \cap B|$$

Example: How many bit strings of length eight either start with a 1 bit or end with the two bits 00?

Solution: Use the subtraction rule.

- Number of bit strings of length eight that start with a 1 bit: $2^7 = 128$
- Number of bit strings of length eight that end with bits 00: $2^6 = 64$
- Number of bit strings of length eight that start with a 1 bit and end with bits 00 : $2^5 = 32$

Hence, the number is $128 + 64 - 32 = 160$.



Basic Counting Principles: Division Rule

Division Rule: There are n/d ways to do a task if it can be done using a procedure that can be carried out in n ways, and for every way w , exactly d of the n ways correspond to way w .

- Restated in terms of sets: If the finite set A is the union of n pairwise disjoint subsets each with d elements, then $n = |A|/d$.
- In terms of functions: If f is a function from A to B , where both are finite sets, and for every value $y \in B$ there are exactly d values $x \in A$ such that $f(x) = y$, then $|B| = |A|/d$.

Basic Counting Principles: Division Rule

Example: How many ways are there to seat four people around a circular table, where two seatings are considered the same when each person has the same left and right neighbor?

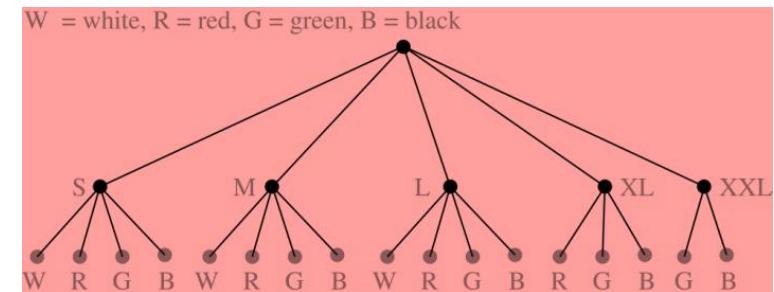
Solution: Number the seats around the table from 1 to 4 proceeding clockwise. There are four ways to select the person for seat 1, 3 for seat 2, 2, for seat 3, and one way for seat 4.

Thus there are $4! = 24$ ways to order the four people. But since two seatings are the same when each person has the same left and right neighbor, for every choice for seat 1, we get the same seating.

Therefore, by the division rule, there are $24/4 = 6$ different seating arrangements.

Tree Diagrams

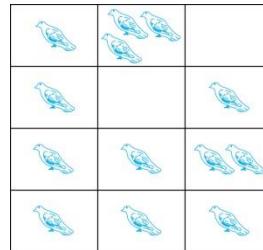
- **Tree Diagrams:** We can solve many counting problems through the use of *tree diagrams*, where a branch represents a possible choice and the leaves represent possible outcomes.
- **Example:** Suppose that “I Love Discrete Math” T-shirts come in five different sizes: S,M,L,XL, and XXL. Each size comes in four colors (white, red, green, and black), except XL, which comes only in red, green, and black, and XXL, which comes only in green and black. What is the minimum number of shirts that the campus book store needs to stock to have one of each size and color available?
- **Solution:** Draw the tree diagram.



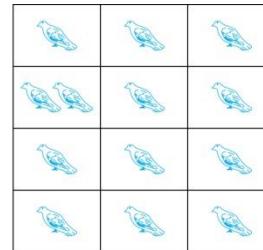
- The store must stock 17 T-shirts.

The Pigeonhole Principle

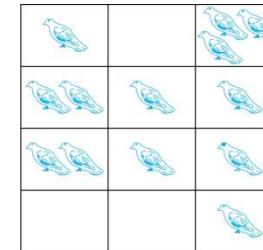
- If a flock of 20 pigeons roosts in a set of 19 pigeonholes, one of the pigeonholes must have more than 1 pigeon.



(a)



(b)



(c)

Pigeonhole Principle: If k is a positive integer and $k + 1$ objects are placed into k boxes, then at least one box contains two or more objects.

Proof: We use a proof by contraposition. Suppose none of the k boxes has more than one object. Then the total number of objects would be at most k . This contradicts the statement that we have $k + 1$ objects.

The Pigeonhole Principle

Corollary 1: A function f from a set with $k + 1$ elements to a set with k elements is not one-to-one.

Proof: Use the pigeonhole principle.

- Create a box for each element y in the codomain of f .
- Put in the box for y all of the elements x from the domain such that $f(x) = y$.
- Because there are $k + 1$ elements and only k boxes, at least one box has two or more elements.

Hence, f can't be one-to-one.

Pigeonhole Principle

Example: Among any group of 367 people, there must be at least two with the same birthday, because there are only 366 possible birthdays.

The Generalized Pigeonhole Principle

The Generalized Pigeonhole Principle: If N objects are placed into k boxes, then there is at least one box containing at least $\lceil N/k \rceil$ objects.

Proof: We use a proof by contraposition. Suppose that none of the boxes contains more than $\lceil N/k \rceil - 1$ objects. Then the total number of objects is at most

$$k \left(\left\lceil \frac{N}{k} \right\rceil - 1 \right) < k \left(\left(\frac{N}{k} + 1 \right) - 1 \right) = N,$$

where the inequality $\lceil N/k \rceil < \lceil N/k \rceil + 1$ has been used. This is a contradiction because there are a total of n objects.

Example: Among 100 people there are at least $\lceil 100/12 \rceil = 9$ who were born in the same month.

The Generalized Pigeonhole Principle

Example:

- 1) How many cards must be selected from a standard deck of 52 cards to guarantee that at least three cards of the same suit are chosen?

Answer

We assume four boxes; one for each suit. Using the generalized pigeonhole principle, at least one box contains at least $\lceil N/4 \rceil$ cards. At least three cards of one suit are selected if $\lceil N/4 \rceil \geq 3$. The smallest integer N such that $\lceil N/4 \rceil \geq 3$ is $N = 2 \cdot 4 + 1 = 9$.

The Generalized Pigeonhole Principle

2) How many must be selected to guarantee that at least three hearts are selected?

Answer

A deck contains 13 hearts and 39 cards which are not hearts. So, if we select 41 cards, we may have 39 cards which are not hearts along with 2 hearts. However, when we select 42 cards, we must have at least three hearts. (Note that the generalized pigeonhole principle is not used here.)

Permutations



Definition: A *permutation* of a set of distinct objects is an ordered arrangement of these objects. An ordered arrangement of r elements of a set is called an *r -permutation*.

Example: Let $S = \{1, 2, 3\}$.

- The ordered arrangement 3,1,2 is a permutation of S .
- The ordered arrangement 3,2 is a 2-permutation of S .
- The number of r -permutations of a set with n elements is denoted by $P(n,r)$.
 - The 2-permutations of $S = \{1, 2, 3\}$ are 1,2; 1,3; 2,1; 2,3; 3,1; and 3,2. Hence, $P(3,2) = 6$.

A Formula for the Number of Permutations

Theorem 1: If n is a positive integer and r is an integer with $1 \leq r \leq n$, then there are

$P(n, r) = n(n - 1)(n - 2) \cdots (n - r + 1)$ r -permutations of a set with n distinct elements.

Proof: Use the product rule. The first element can be chosen in n ways. The second in $n - 1$ ways, and so on until there are $(n - (r - 1))$ ways to choose the last element.

Corollary 1: If n and r are integers with $1 \leq r \leq n$, then

$$P(n, r) = \frac{n!}{(n-r)!}$$

Solving Counting Problems by Counting Permutations



Example: How many ways are there to select a first-prize winner, a second prize winner, and a third-prize winner from 100 different people who have entered a contest?

Solution: $P(100,3) = 100 \cdot 99 \cdot 98 = 970,200$

Example: How many permutations of the letters *ABCDEFGH* contain the string *ABC* ?

Solution: We solve this problem by counting the permutations of six objects, *ABC*, *D*, *E*, *F*, *G*, and *H*.

$$6! = 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 720$$

Combinations

Definition: An *r-combination* of elements of a set is an unordered selection of *r* elements from the set. Thus, an *r*-combination is simply a subset of the set with *r* elements.

- The number of *r*-combinations of a set with *n* distinct elements is denoted by $C(n, r)$. The notation $\binom{n}{r}$ is also used and is called a *binomial coefficient*.

Example: Let *S* be the set $\{a, b, c, d\}$. Then $\{a, c, d\}$ is a 3-combination from *S*. It is the same as $\{d, c, a\}$ since the order listed does not matter.

- $C(4,2) = 6$ because the 2-combinations of $\{a, b, c, d\}$ are the six subsets $\{a, b\}$, $\{a, c\}$, $\{a, d\}$, $\{b, c\}$, $\{b, d\}$, and $\{c, d\}$.

Theorem 2 : Combinations

Theorem 2: The number of r -combinations of a set with n elements, where $n \geq r \geq 0$, equals $C(n, r) = \frac{n!}{(n-r)!r!}$.

Proof: By the product rule $P(n, r) = C(n, r) \cdot P(r, r)$. Therefore,

$$C(n, r) = \frac{P(n, r)}{P(r, r)} = \frac{n!/(n-r)!}{r!/(r-r)!} = \frac{n!}{(n-r)!r!} .$$

Corollary 2: Let n and r be nonnegative integers with $r \leq n$. Then $C(n, r) = C(n, n - r)$.

Proof: From Theorem 2, it follows that

$$C(n, r) = \frac{n!}{(n-r)!r!}$$

and

$$C(n, n - r) = \frac{n!}{(n-r)![n-(n-r)]!} = \frac{n!}{(n-r)!r!} .$$

Hence, $C(n, r) = C(n, n - r)$.

Combinations

Example: How many poker hands of five cards can be dealt from a standard deck of 52 cards? Also, how many ways are there to select 47 cards from a deck of 52 cards?

Solution: Since the order in which the cards are dealt does not matter, the number of five card hands is:

$$C(52, 5) = \frac{52!}{5!47!} = \frac{52 \cdot 51 \cdot 50 \cdot 49 \cdot 48}{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 26 \cdot 17 \cdot 10 \cdot 49 \cdot 12 = 2,598,960$$

- The different ways to select 47 cards from 52 is

$$C(52, 47) = \frac{52!}{47!5!} = C(52, 5) = 2,598,960.$$



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 14

Agenda

- Combinatorial Proofs
 - Binomial Coefficients
 - Binomial theorem
 - Recurrence Relations
 - Application of Recurrence Relations
-

Combinatorial Proofs



- **Definition 1:** A *combinatorial proof* of an identity is a proof that uses one of the following methods.
 - A *double counting proof* uses counting arguments to prove that both sides of an identity count the same objects, but in different ways.
 - A *bijective proof* shows that there is a bijection between the sets of objects counted by the two sides of the identity.

Double Counting Proof

Combinatorial proofs that $C(n, r) = C(n, n - r)$ when r and n are nonnegative integers with $r \leq n$

- : By definition the number of subsets of S with r elements is $C(n, r)$.
- Each subset A of S can also be described by specifying which elements are not in A , i.e., those which are in \bar{A} .
- Since the complement of a subset of S with r elements has $n - r$ elements, there are also $C(n, n - r)$ subsets of S with r elements.



Bijective Proof

Combinatorial proofs that $C(n, r) = C(n, n - r)$ when r and n are nonnegative integers with $r \leq n$

- Suppose that S is a set with n elements.
- The function that maps a subset A of S to \bar{A} is a bijection between the subsets of S with r elements and the subsets with $n - r$ elements.
- Since there is a bijection between the two sets, they must have the same number of elements.



Combinations

Example: How many ways are there to select five players from a 10-member tennis team to make a trip to a match at another school.

Solution: By Theorem 2, the number of combinations is

$$C(10, 5) = \frac{10!}{5!5!} = 252.$$

Combinations

Example: A group of 30 people have been trained as astronauts to go on the first mission to Mars. How many ways are there to select a crew of six people to go on this mission?

Solution: By Theorem 2, the number of possible crews is

$$C(30, 6) = \frac{30!}{6!24!} = \frac{30 \cdot 29 \cdot 28 \cdot 27 \cdot 26 \cdot 25}{6 \cdot 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1} = 593,775 .$$

Powers of Binomial Expressions

Definition: A *binomial* expression is the sum of two terms, such as $x + y$.
(More generally, these terms can be products of constants and variables.)

- We can use counting principles to find the coefficients in the expansion of $(x + y)^n$ where n is a positive integer.
- To illustrate this idea, we first look at the process of expanding $(x + y)^3$.
- $(x + y)(x + y)(x + y)$ expands into a sum of terms that are the product of a term from each of the three sums.

Powers of Binomial Expressions

- Terms of the form x^3, x^2y, xy^2, y^3 arise. The question is what are the coefficients?
 - To obtain x^3 , an x must be chosen from each of the sums. There is only one way to do this. So, the coefficient of x^3 is 1.
 - To obtain x^2y , an x must be chosen from two of the sums and a y from the other. There are $\binom{3}{2}$ ways to do this and so the coefficient of x^2y is 3.
 - To obtain xy^2 , an x must be chosen from one of the sums and a y from the other two. There are $\binom{3}{1}$ ways to do this and so the coefficient of xy^2 is 3.
 - To obtain y^3 , a y must be chosen from each of the sums. There is only one way to do this. So, the coefficient of y^3 is 1.
- We have used a counting argument to show that $(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$.

Binomial Theorem

Binomial Theorem: Let x and y be variables, and n a nonnegative integer. Then:

$$(x+y)^n = \sum_{j=0}^n \binom{n}{j} x^{n-j} y^j = \binom{n}{0} x^n + \binom{n}{1} x^{n-1} y + \cdots + \binom{n}{n-1} x y^{n-1} + \binom{n}{n} y^n.$$

Proof: We use combinatorial reasoning .

The terms in the expansion of $(x + y)^n$ are of the form $x^{n-j}y^j$ for $j = 0, 1, 2, \dots, n$.

To form the term $x^{n-j}y^j$, it is necessary to choose $n-j$ x 's from the n sums.

Therefore, the coefficient of $x^{n-j}y^j$ is $\binom{n}{n-j}$ which equals $\binom{n}{j}$.



Using the Binomial Theorem

Example: What is the coefficient of $x^{12}y^{13}$ in the expansion of $(2x - 3y)^{25}$?

Solution: We view the expression as $(2x + (-3y))^{25}$. By the binomial theorem

$$(2x + (-3y))^{25} = \sum_{j=0}^{25} \binom{25}{j} (2x)^{25-j}(-3y)^j.$$

Consequently, the coefficient of $x^{12}y^{13}$ in the expansion is obtained when $j = 13$.

$$\binom{25}{13} 2^{12}(-3)^{13} = -\frac{25!}{13!12!} 2^{12}3^{13}.$$

A Useful Identity

Corollary 1: With $n \geq 0$,

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

Proof (using binomial theorem): With $x = 1$ and $y = 1$, from the binomial theorem we see that:

$$2^n = (1 + 1)^n = \sum_{k=0}^n \binom{n}{k} 1^k 1^{(n-k)} = \sum_{k=0}^n \binom{n}{k}.$$

Proof (combinatorial): Consider the subsets of a set with n elements. There are $\binom{n}{0}$ subsets with zero elements, $\binom{n}{1}$ with one element, $\binom{n}{2}$ with two elements, ..., and $\binom{n}{n}$ with n elements. Therefore, the total is

$$\sum_{k=0}^n \binom{n}{k}.$$

Since, we know that a set with n elements has 2^n subsets, we conclude:

$$\sum_{k=0}^n \binom{n}{k} = 2^n.$$

Pascal's Identity

Pascal's Identity: If n and k are integers with $n \geq k \geq 0$, then

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

Proof (combinatorial): Let T be a set where $|T| = n + 1$, $a \in T$, and $S = T - \{a\}$. There are $\binom{n+1}{k}$ subsets of T containing k elements.

Each of these subsets of T either:

- contains a with $k - 1$ other elements, or
- contains k elements of S and does not contain a .

Pascal's Identity

There are

- $\binom{n}{k-1}$ subsets of k elements that contain a , since there are $\binom{n}{k-1}$ subsets of $k-1$ elements of S ,
- $\binom{n}{k}$ subsets of k elements of T that do not contain a , because there are $\binom{n}{k}$ subsets of k elements of S .

Hence,

$$\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}.$$

Pascal's Triangle

The n th row in the triangle consists of the binomial coefficients $\binom{n}{k}$, $k = 0, 1, \dots, n$.

By Pascal's identity, adding two adjacent binomial coefficients results in the binomial coefficient in the next row between these two coefficients.

$$\begin{array}{c}
 \binom{0}{0} & & & & & & 1 \\
 \binom{1}{0} \binom{1}{1} & & & & & & 1 \quad 1 \\
 \binom{2}{0} \binom{2}{1} \binom{2}{2} & & & & & & 1 \quad 2 \quad 1 \\
 \binom{3}{0} \binom{3}{1} \binom{3}{2} \binom{3}{3} & & \text{By Pascal's identity:} & & \binom{6}{4} + \binom{6}{5} = \binom{7}{5} & & 1 \quad 3 \quad 3 \quad 1 \\
 \binom{4}{0} \binom{4}{1} \binom{4}{2} \binom{4}{3} \binom{4}{4} & & & & & & 1 \quad 4 \quad 6 \quad 4 \quad 1 \\
 \binom{5}{0} \binom{5}{1} \binom{5}{2} \binom{5}{3} \binom{5}{4} \binom{5}{5} & & & & & & 1 \quad 5 \quad 10 \quad 10 \quad 5 \quad 1 \\
 \binom{6}{0} \binom{6}{1} \binom{6}{2} \binom{6}{3} \binom{6}{4} \binom{6}{5} \binom{6}{6} & & & & & & 1 \quad 6 \quad 15 \quad 20 \quad 15 \quad 6 \quad 1 \\
 \binom{7}{0} \binom{7}{1} \binom{7}{2} \binom{7}{3} \binom{7}{4} \binom{7}{5} \binom{7}{6} \binom{7}{7} & & & & & & 1 \quad 7 \quad 21 \quad 35 \quad 35 \quad 21 \quad 7 \quad 1 \\
 \binom{8}{0} \binom{8}{1} \binom{8}{2} \binom{8}{3} \binom{8}{4} \binom{8}{5} \binom{8}{6} \binom{8}{7} \binom{8}{8} & & & & & & 1 \quad 8 \quad 28 \quad 56 \quad 70 \quad 56 \quad 28 \quad 8 \quad 1 \\
 \binom{n+1}{k} \stackrel{(a)}{=} \binom{n}{k-1} + \binom{n}{k} & & & & & & \dots \\
 & & & & & & \binom{n}{k} \stackrel{(b)}{=}
 \end{array}$$

Vandermonde's identity

Let m, n, r be nonnegative integers with r not exceeding m or n . then

$$\binom{m+n}{r} = \sum_{k=0}^r \binom{m}{r-k} \binom{n}{k}$$

Proof: Suppose that there are m items in one set and n items in a second set. Then the total number of ways to pick r elements from the union of these sets is $\binom{m+n}{r}$. Another way to pick r elements from the union is to pick k elements from the second set and then $r - k$ elements from the first set, where k is an integer with $0 \leq k \leq r$.

Because there are $\binom{n}{k}$ ways to choose k elements from the second set and $\binom{m}{r-k}$ ways to choose $r - k$ elements from the first set, the product rule tells us that this can be done in $\binom{m}{r-k} \binom{n}{k}$ ways. Hence, the total number of ways to pick r elements from the union also equals $\sum_{k=0}^r \binom{m}{r-k} \binom{n}{k}$

Vandermonde's identity : Corollary

If n is a nonnegative integer, then

$$\binom{2n}{n} = \sum_{k=0}^n \binom{n}{k}^2$$

Proof:

We use identity with $m=r=n$ to obtain

$$\binom{2n}{n} = \sum_{k=0}^n \binom{n}{n-k} \binom{n}{k} = \sum_{k=0}^n \binom{n}{k}^2$$

Theorem 4

Let n and r be nonnegative integers with $r \leq n$, then

$$\binom{n+1}{r+1} = \sum_{j=r}^n \binom{j}{r}$$

Proof:

We use a combinatorial proof. The left-hand side, $\binom{n+1}{r+1}$ counts the bit strings of length $n+1$ containing $r+1$ ones.

We show that the right-hand side counts the same objects by considering the cases corresponding to the possible locations of the final 1 in a string with $r+1$ ones. This final one must occur at position $r+1, r+2, \dots, n+1$. Furthermore, if the last one is the k th bit there must be r ones among the first $k-1$ positions. Consequently, there are $\binom{k-1}{r}$ such bit strings. Summing over k with $r+1 \leq k \leq n+1$, we find that there are $\sum_{k=r+1}^{n+1} \binom{k-1}{r} = \sum_{j=r}^n \binom{j}{r}$ bit strings of length n containing exactly $r+1$ ones.

Recurrence Relations

Definition

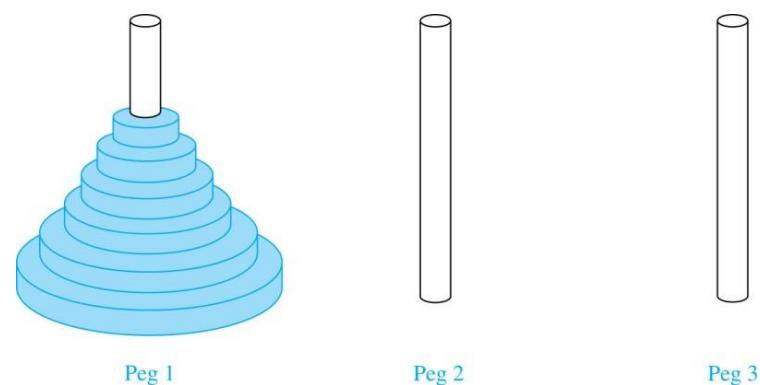
A **recurrence relation** for the sequence $\{a_n\}$ is an equation that expresses a_n in terms of one or more of the previous terms of the sequence, namely, a_0, a_1, \dots, a_{n-1} , for all integers n with $n \geq n_0$, where n_0 is a nonnegative integer.

- A sequence is called a **solution** of a recurrence relation if its terms satisfy the recurrence relation.
- The **initial conditions** for a sequence specify the terms that precede the first term where the recurrence relation takes effect.

The Tower of Hanoi

In the late nineteenth century, the French mathematician Édouard Lucas invented a puzzle consisting of three pegs on a board with disks of different sizes. Initially all of the disks are on the first peg in order of size, with the largest on the bottom.

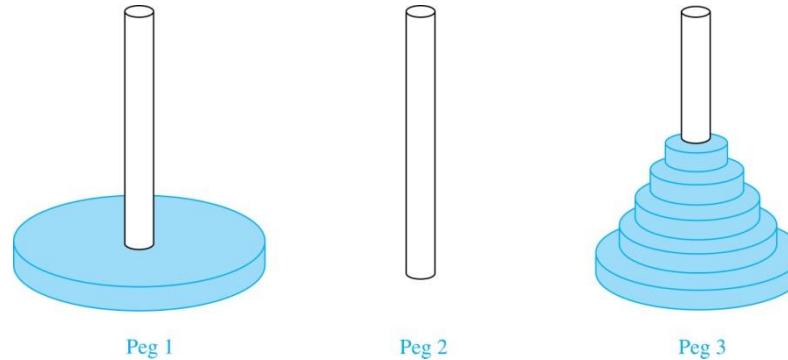
- **Rules:** You are allowed to move the disks one at a time from one peg to another as long as a larger disk is never placed on a smaller.
- **Goal:** Using allowable moves, end up with all the disks on the second peg in order of size with largest on the bottom.



The Initial Position in the Tower of Hanoi Puzzle

The Tower of Hanoi (*continued*)

Solution: Let $\{H_n\}$ denote the number of moves needed to solve the Tower of Hanoi Puzzle with n disks. Set up a recurrence relation for the sequence $\{H_n\}$. Begin with n disks on peg 1. We can transfer the top $n - 1$ disks, following the rules of the puzzle, to peg 3 using H_{n-1} moves.



First, we use 1 move to transfer the largest disk to the second peg. Then we transfer the $n - 1$ disks from peg 3 to peg 2 using H_{n-1} additional moves. This can not be done in fewer steps. Hence,

$$H_n = 2H_{n-1} + 1.$$

The initial condition is $H_1 = 1$ since a single disk can be transferred from peg 1 to peg 2 in one move.

The Tower of Hanoi (*continued*)

- We can use an iterative approach to solve this recurrence relation by repeatedly expressing H_n in terms of the previous terms of the sequence.

$$\begin{aligned}H_n &= 2H_{n-1} + 1 \\&= 2(2H_{n-2} + 1) + 1 = 2^2 H_{n-2} + 2 + 1 \\&= 2^2(2H_{n-3} + 1) + 2 + 1 = 2^3 H_{n-3} + 2^2 + 2 + 1 \\&\vdots \\&= 2^{n-1}H_1 + 2^{n-2} + 2^{n-3} + \dots + 2 + 1 \\&= 2^{n-1} + 2^{n-2} + 2^{n-3} + \dots + 2 + 1 \quad \text{because } H_1 = 1 \\&= 2^n - 1 \quad \text{using the formula for the sum of the terms of a geometric series}\end{aligned}$$

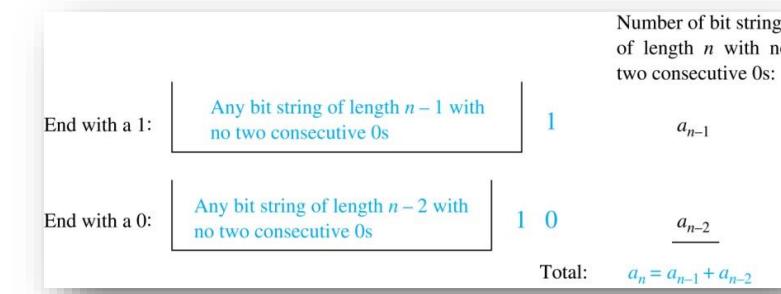
Counting Bit Strings

Example 3: Find a recurrence relation and give initial conditions for the number of bit strings of length n without two consecutive 0s. How many such bit strings are there of length five?

Solution: Let a_n denote the number of bit strings of length n without two consecutive 0s. To obtain a recurrence relation for $\{a_n\}$ note that the number of bit strings of length n that do not have two consecutive 0s is the number of bit strings ending with a 0 plus the number of such bit strings ending with a 1. Now assume that $n \geq 3$.

- The bit strings of length n ending with 1 without two consecutive 0s are the bit strings of length $n - 1$ with no two consecutive 0s with a 1 at the end. Hence, there are a_{n-1} such bit strings.
- The bit strings of length n ending with 0 without two consecutive 0s are the bit strings of length $n - 2$ with no two consecutive 0s with 10 at the end. Hence, there are a_{n-2} such bit strings.

We conclude that $a_n = a_{n-1} + a_{n-2}$ for $n \geq 3$.



Bit Strings (*continued*)

The initial conditions are:

- $a_1 = 2$, since both the bit strings 0 and 1 do not have consecutive 0s.
- $a_2 = 3$, since the bit strings 01, 10, and 11 do not have consecutive 0s, while 00 does.

To obtain a_5 , we use the recurrence relation three times to find that:

- $a_3 = a_2 + a_1 = 3 + 2 = 5$
- $a_4 = a_3 + a_2 = 5 + 3 = 8$
- $a_5 = a_4 + a_3 = 8 + 5 = 13$

Note that $\{a_n\}$ satisfies the same recurrence relation as the Fibonacci sequence. Since $a_1 = f_3$ and $a_2 = f_4$, we conclude that $a_n = f_{n+2}$.

Counting the Ways to Parenthesize a Product

Example

Find a recurrence relation for C_n , the number of ways to parenthesize the product of $n + 1$ numbers, $x_0 \cdot x_1 \cdot x_2 \cdot \dots \cdot x_n$, to specify the order of multiplication.

For example, $C_3 = 5$, since all the possible ways to parenthesize 4 numbers are

1. $((x_0 \cdot x_1) \cdot x_2) \cdot x_3$
2. $(x_0 \cdot (x_1 \cdot x_2)) \cdot x_3$
3. $(x_0 \cdot x_1) \cdot (x_2 \cdot x_3)$
4. $x_0 \cdot ((x_1 \cdot x_2) \cdot x_3)$
5. $x_0 \cdot (x_1 \cdot (x_2 \cdot x_3))$

Counting the Ways to Parenthesize a Product

Solution:

- Note that however parentheses are inserted in $x_0 \cdot x_1 \cdot x_2 \cdot \dots \cdot x_n$, one “.” operator remains outside all parentheses.
- This final operator appears between two of the $n + 1$ numbers, say x_k and x_{k+1} .
- Since there are C_k ways to insert parentheses in the product $x_0 \cdot x_1 \cdot x_2 \cdot \dots \cdot x_k$ and C_{n-k-1} ways to insert parentheses in the product $x_{k+1} \cdot x_{k+2} \cdot \dots \cdot x_n$, we have

The initial conditions are $C_0 = 1$ and $C_1 = 1$.

$$\begin{aligned}C_n &= C_0 C_{n-1} + C_1 C_{n-2} + \dots + C_{n-2} C_1 + C_{n-1} C_0 \\&= \sum_{k=0}^{n-1} C_k C_{n-k-1}\end{aligned}$$

The sequence $\{C_n\}$ is the sequence of **Catalan Numbers**. This recurrence relation can be solved using the method of generating functions; see Exercise 41 in Section 8.4.



BITS Pilani
Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 15

Agenda

- Solving linear recurrence relations
- Linear homogeneous recurrence relations with constant coefficients
- Characteristic equation and multiplicity
- Linear nonhomogeneous recurrence relations with constant coefficients

Linear Homogeneous Recurrence Relations

Definition: A *linear homogeneous recurrence relation of degree k with constant coefficients* is a recurrence relation of the form

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k},$$

where c_1, c_2, \dots, c_k are real numbers, and $c_k \neq 0$

1. it is ***linear*** because the right-hand side is a sum of the previous terms of the sequence each multiplied by a function of n .
2. it is ***homogeneous*** because no terms occur that are not multiples of the a_j s. Each coefficient is a constant.
3. the ***degree is k*** because a_n is expressed in terms of the previous k terms of the sequence.

Examples of Linear Homogeneous Recurrence Relations



- $P_n = (1.11)P_{n-1}$ linear homogeneous recurrence relation of degree one
- $f_n = f_{n-1} + f_{n-2}$ linear homogeneous recurrence relation of degree two
- $a_n = a_{n-1} + a_{n-2}^2$ not linear
- $H_n = 2H_{n-1} + 1$ not homogeneous
- $B_n = nB_{n-1}$ coefficients are not constants

Linear Homogeneous Recurrence Relations

Consider a linear homogeneous recurrence relation of degree k with constant coefficients

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k},$$

where c_1, c_2, \dots, c_k are real numbers, and $c_k \neq 0$

By strong induction, a sequence satisfying such a recurrence relation is uniquely determined by

1. the recurrence relation and
2. the k initial conditions $a_0 = C_1, a_1 = C_2, \dots, a_{k-1} = C_{k-1}$.

Solving Linear Homogeneous Recurrence Relations

innovate

achieve

lead

1. The basic approach is to look for solutions of the form $a_n = r^n$, where r is a constant.
2. Note that $a_n = r^n$ is a solution to the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k}$ if and only if $r^n = c_1 r^{n-1} + c_2 r^{n-2} + \dots + c_k r^{n-k}$.
3. Algebraic manipulation yields the ***characteristic equation***:

$$r^k - c_1 r^{k-1} - c_2 r^{k-2} - \dots - c_{k-1} r - c_k = 0$$

Solving Linear Homogeneous Recurrence Relations



- The sequence $\{a_n\}$ with $a_n = r^n$ is a solution if and only if r is a solution to the characteristic equation.
- The solutions to the characteristic equation are called the *characteristic roots* of the recurrence relation. The roots are used to give an explicit formula for all the solutions of the recurrence relation.

Theorem 1



Solving Linear Homogeneous Recurrence Relations of Degree Two

Let c_1 and c_2 be real numbers. Suppose that $r^2 - c_1r - c_2 = 0$ has two distinct roots r_1 and r_2 .

Then the sequence $\{a_n\}$ is a solution to the recurrence relation $a_n = c_1a_{n-1} + c_2a_{n-2}$ if and only if

$$a_n = \alpha r_1^n + \alpha_2 r_2^n$$

for $n = 0, 1, 2, \dots$, where α_1 and α_2 are constants.

Using Theorem 1

Example: What is the solution to the recurrence relation $a_n = a_{n-1} + 2a_{n-2}$ with $a_0 = 2$ and $a_1 = 7$?

Solution:

1. The characteristic equation is $r^2 - r - 2 = 0$. Its roots are $r = 2$ and $r = -1$. Therefore, $\{a_n\}$ is a solution to the recurrence relation if and only if $a_n = \alpha_1 2^n + \alpha_2 (-1)^n$, for some constants α_1 and α_2 .
2. To find the constants α_1 and α_2 , note that $a_0 = 2 = \alpha_1 + \alpha_2$ and $a_1 = 7 = \alpha_1 2 + \alpha_2 (-1)$.
3. Solving these equations, we find that $\alpha_1 = 3$ and $\alpha_2 = -1$.
4. Hence, the solution is the sequence $\{a_n\}$ with $a_n = 3 \cdot 2^n - (-1)^n$.

An Explicit Formula for the Fibonacci Numbers

We can use Theorem 1 to find an explicit formula for the Fibonacci numbers. The sequence of Fibonacci numbers satisfies the recurrence relation $f_n = f_{n-1} + f_{n-2}$ with the initial conditions: $f_0 = 0$ and $f_1 = 1$.

Solution: The roots of the characteristic equation $r^2 - r - 1 = 0$ are

$$r_1 = \frac{1+\sqrt{5}}{2} \quad r_2 = \frac{1-\sqrt{5}}{2} \quad \text{for some constants } \alpha_1 \text{ and } \alpha_2.$$

By Theorem 1

$$f_n = \alpha_1 \left(\frac{1+\sqrt{5}}{2} \right)^n + \alpha_2 \left(\frac{1-\sqrt{5}}{2} \right)^n$$

Using the initial conditions $f_0 = 0$ and $f_1 = 1$, we have

$$f_0 = \alpha_1 + \alpha_2 = 0 \quad f_1 = \alpha_1 \left(\frac{1+\sqrt{5}}{2} \right) + \alpha_2 \left(\frac{1-\sqrt{5}}{2} \right) = 1 \quad \alpha_1 = \frac{1}{\sqrt{5}} \quad \alpha_2 = -\frac{1}{\sqrt{5}}$$

$$f_n = \frac{1}{\sqrt{5}} \left(\frac{1+\sqrt{5}}{2} \right)^n - \frac{1}{\sqrt{5}} \left(\frac{1-\sqrt{5}}{2} \right)^n$$

Theorem 2



The Solution when there is a Repeated Root

Theorem 2: Let c_1 and c_2 be real numbers with $c_2 \neq 0$. Suppose that $r^2 - c_1r - c_2 = 0$ has one repeated root r_0 . Then the sequence $\{a_n\}$ is a solution to the recurrence relation $a_n = c_1a_{n-1} + c_2a_{n-2}$ if and only if

$$a_n = \alpha r_0^n + \alpha_2 n r_0^n$$

for $n = 0, 1, 2, \dots$, where α_1 and α_2 are constants.

Example: What is the solution to the recurrence relation $a_n = 6a_{n-1} - 9a_{n-2}$ with $a_0 = 1$ and $a_1 = 6$?

Solution:

- The characteristic equation is $r^2 - 6r + 9 = 0$. The only root is $r = 3$. Therefore, $\{a_n\}$ is a solution to the recurrence relation if and only if $a_n = \alpha_1 3^n + \alpha_2 n (3)^n$ where α_1 and α_2 are constants.
- To find the constants α_1 and α_2 , note that $a_0 = 1 = \alpha_1$ and $a_1 = 6 = \alpha_1 \cdot 3 + \alpha_2 \cdot 3$.
- Solving, we find that $\alpha_1 = 1$ and $\alpha_2 = 1$. Hence, $a_n = 3^n + n3^n$.

Theorem 3

Solving Linear Homogeneous Recurrence Relations of Arbitrary Degree



Theorem 3: Let c_1, c_2, \dots, c_k be real numbers. Suppose that the characteristic equation $r^k - c_1r^{k-1} - \dots - c_k = 0$ has k distinct roots r_1, r_2, \dots, r_k . Then a sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = c_1a_{n-1} + c_2a_{n-2} + \dots + c_k a_{n-k}$ if and only if for $n = 0, 1, 2, \dots$, where $\alpha_1, \alpha_2, \dots, \alpha_k$ are constants.

$$a_n = \alpha_1 r_1^n + \alpha_2 r_2^m + \dots + \alpha_k r_k^n$$

This theorem can be used to solve linear homogeneous recurrence relations with constant coefficients of any degree when the characteristic equation has distinct roots.

The General Case with Repeated Roots Allowed

Theorem 4: Let c_1, c_2, \dots, c_k be real numbers. Suppose that the characteristic equation $r^k - c_1 r^{k-1} - \dots - c_k = 0$ has t distinct roots r_1, r_2, \dots, r_t with multiplicities m_1, m_2, \dots, m_t , respectively so that $m_i \geq 1$ for $i = 0, 1, 2, \dots, t$ and $m_1 + m_2 + \dots + m_t = k$. Then a sequence $\{a_n\}$ is a solution of the recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k}$

if and only if

$$\begin{aligned} a_n = & (\alpha_{1,0} + \alpha_{1,1}n + \dots + \alpha_{1,m_1-1}n^{m_1-1})r_1^n \\ & + (\alpha_{2,0} + \alpha_{2,1}n + \dots + \alpha_{2,m_2-1}n^{m_2-1})r_2^n \\ & + \dots + (\alpha_{t,0} + \alpha_{t,1}n + \dots + \alpha_{t,m_t-1}n^{m_t-1})r_t^n \end{aligned}$$

for $n = 0, 1, 2, \dots$, where $\alpha_{i,j}$ are constants for $1 \leq i \leq t$ and $0 \leq j \leq m_{i-1}$.

Linear Non-homogeneous Recurrence Relations with Constant Coefficients

Definition: A *linear nonhomogeneous recurrence relation with constant coefficients* is a recurrence relation of the form:

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k} + F(n),$$

where c_1, c_2, \dots, c_k are real numbers, and $F(n)$ is a function not identically zero depending only on n .

The recurrence relation $a_n = c_1 a_{n-1} + c_2 a_{n-2} + \dots + c_k a_{n-k}$, is called the associated homogeneous recurrence relation.

Linear Non-homogeneous Recurrence Relations with Constant Coefficients (cont.)

innovate

achieve

lead

The following are linear nonhomogeneous recurrence relations with constant coefficients:

1. $a_n = a_{n-1} + 2^n,$
2. $a_n = a_{n-1} + a_{n-2} + n^2 + n + 1,$
3. $a_n = 3a_{n-1} + n3^n,$
4. $a_n = a_{n-1} + a_{n-2} + a_{n-3} + n!$

where the following are the associated linear homogeneous recurrence relations, respectively:

- a) $a_n = a_{n-1},$
- b) $a_n = a_{n-1} + a_{n-2},$
- c) $a_n = 3a_{n-1},$
- d) $a_n = a_{n-1} + a_{n-2} + a_{n-3}$

Solving Linear Non-homogeneous Recurrence Relations with Constant Coefficients

Theorem 5: If $\{a_n^{(p)}\}$ is a particular solution of the nonhomogeneous linear recurrence relation with constant coefficients

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k} + F(n),$$

then every solution is of the form $\{a_n^{(h)} + a_n^{(p)}\}$ where $\{a_n^{(h)}\}$ is a solution of the associated homogeneous recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k}.$$

Solving Linear Non-homogeneous Recurrence Relations with Constant Coefficients (*continued*)



Example: Find all solutions of the recurrence relation $a_n = 3a_{n-1} + 2n$. What is the solution with $a_1 = 3$?

Solution: The associated linear homogeneous equation is $a_n = 3a_{n-1}$.

Its solutions are $a_n^{(h)} = \alpha 3^n$, where α is a constant.

- Because $F(n)=2n$ is a polynomial in n of degree one, to find a particular solution we might try a linear function in n , say $p_n = cn + d$, where c and d are constants.
- Suppose that $p_n = cn + d$ is such a solution.
 - Then $a_n = 3a_{n-1} + 2n$ becomes $cn + d = 3(c(n-1) + d) + 2n$.

Solving Linear Non homogeneous Recurrence Relations with Constant Coefficients (*continued*)



- Simplifying yields $(2 + 2c)n + (2d - 3c) = 0$. It follows that $cn + d$ is a solution if and only if $2 + 2c = 0$ and $2d - 3c = 0$.
- Therefore, $cn + d$ is a solution if and only if $c = -1$ and $d = -3/2$. Consequently, $a_n^{(p)} = -n - 3/2$ is a particular solution.
- By Theorem 5, all solutions are of the form
$$a_n = a_n^{(p)} + a_n^{(h)} = -n - 3/2 + \alpha 3^n, \text{ where } \alpha \text{ is a constant.}$$
- To find the solution with $a_1 = 3$, let $n = 1$ in the above formula for the general solution. Then $3 = -1 - 3/2 + 3\alpha$, and $\alpha = 11/6$.
- Hence, the solution is

$$a_n = -n - 3/2 + (11/6)3^n.$$

Theorem 6

Suppose that $\{a_n\}$ satisfies the linear nonhomogeneous recurrence relation

$$a_n = c_1 a_{n-1} + c_2 a_{n-2} + \cdots + c_k a_{n-k} + F(n),$$

where c_1, c_2, \dots, c_k are real numbers, and

$$F(n) = (b_t n^t + b_{t-1} n^{t-1} + \cdots + b_1 n + b_0) s^n,$$

where b_0, b_1, \dots, b_t and s are real numbers.

- When s is not a root of the characteristic equation of the associated linear homogeneous recurrence relation, there is a particular solution of the form

$$(p_t n^t + p_{t-1} n^{t-1} + \cdots + p_1 n + p_0) s^n$$

- When s is a root of this characteristic equation and its multiplicity is m , there is a particular solution of the form

$$n^m (p_t n^t + p_{t-1} n^{t-1} + \cdots + p_1 n + p_0) s^n$$

.

Example

What form does a particular solution of the linear nonhomogeneous recurrence relation $a_n = 6a_{n-1} - 9a_{n-2} + F(n)$ have when $F(n) = 3^n, F(n) = n3^n, F(n) = n^22^n$, and $F(n) = (n^2 + 1)3^n$?

Solution: The associated linear homogeneous recurrence relation is $a_n = 6a_{n-1} - 9a_{n-2}$. Its characteristic equation, $r^2 - 6r + 9 = (r - 3)^2 = 0$, has a single root, 3, of multiplicity two.

To apply Theorem 6, with $F(n)$ of the form $P(n)s^n$, where $P(n)$ is a polynomial and s is a constant, we need to ask whether s is a root of this characteristic equation. Because $s = 3$ is a root with multiplicity $m = 2$ but $s = 2$ is not a root.

Theorem 6 tells us that a particular solution has the form

1. $p_0 n^2 3^n$ if $F(n) = 3^n$,
2. $n^2(p_1 n + p_0)3^n$ if $F(n) = n3^n$,
3. $(p_2 n^2 + p_1 n + p_0)2^n$ if $F(n) = n^22^n$, and
4. $n^2(p_2 n^2 + p_1 n + p_0)3^n$ if $F(n) = (n^2 + 1)3^n$.



BITS Pilani

Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 16

Agenda

- Generating functions
- Solving recurrence relations using generating functions
- Extended binomial theorem
- Counting problems and generating functions

Generating Functions

Definition: The *generating function for the sequence* $a_0, a_1, \dots, a_k, \dots$ of real numbers is the infinite series

$$G(x) = a_0 + a_1x + \cdots + a_kx^k + \cdots = \sum_{k=0}^{\infty} a_kx^k.$$

Examples:

- The sequence $\{a_k\}$ with $a_k = 3$ has the generating function $\sum_{k=0}^{\infty} 3x^k$.
- The sequence $\{a_k\}$ with $a_k = k + 1$ has the generating function has the generating function $\sum_{k=0}^{\infty} (k + 1)x^k$.
- The sequence $\{a_k\}$ with $a_k = 2^k$ has the generating function has the generating function $\sum_{k=0}^{\infty} 2^k x^k$.

Generating Functions for Finite Sequences



- Generating functions for finite sequences of real numbers can be defined by extending a finite sequence a_0, a_1, \dots, a_n into an infinite sequence by setting $a_{n+1} = 0, a_{n+2} = 0$, and so on.
- The generating function $G(x)$ of this sequence $\{a_n\}$ is a polynomial of degree n because no terms of the form $a_j x^j$ with $j > n$ occur, that is,

$$G(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Generating Functions for Finite Sequences (continued)

Example: What is the generating function for the sequence
1,1,1,1,1,1?

Solution: The generating function of 1,1,1,1,1,1 is

$$1 + x + x^2 + x^3 + x^4 + x^5.$$

we have

$$(x^6 - 1)/(x - 1) = 1 + x + x^2 + x^3 + x^4 + x^5$$

when $x \neq 1$.

Consequently $G(x) = (x^6 - 1)/(x - 1)$ is the generating function of the sequence.

Useful Facts : Power Series

The function $\frac{1}{1-x}$ is the generating function of the sequence 1,1,1,1,1.....

Because $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$ for $|x| < 1$

The function $\frac{1}{1-ax}$ is the generating function of the sequence 1,a, a^2 , a^3 ,....

Because $\frac{1}{1-ax} = 1 + ax + a^2x^2 + a^3x^3 + \dots$ for $|x| < \frac{1}{a}$ for $a \neq 0$

Theorem 1

Let $f(x) = \sum_{k=0}^{\infty} a_k x^k$ and $g(x) = \sum_{k=0}^{\infty} b_k x^k$. then

$$f(x) + g(x) = \sum_{k=0}^{\infty} (a_k + b_k) x^k \text{ and } f(x)g(x) = \sum_{k=0}^{\infty} \left(\sum_{j=0}^k a_k b_{k-j} \right) x^k$$

Useful Generating Functions

TABLE 1 Useful Generating Functions.

$G(x)$	a_k
$(1 + x)^n = \sum_{k=0}^n C(n, k)x^k$ $= 1 + C(n, 1)x + C(n, 2)x^2 + \cdots + x^n$	$C(n, k)$
$(1 + ax)^n = \sum_{k=0}^n C(n, k)a^kx^k$ $= 1 + C(n, 1)ax + C(n, 2)a^2x^2 + \cdots + a^nx^n$	$C(n, k)a^k$
$(1 + x^r)^n = \sum_{k=0}^n C(n, k)x^{rk}$ $= 1 + C(n, 1)x^r + C(n, 2)x^{2r} + \cdots + x^{rn}$	$C(n, k/r)$ if $r \mid k$; 0 otherwise
$\frac{1 - x^{n+1}}{1 - x} = \sum_{k=0}^n x^k = 1 + x + x^2 + \cdots + x^n$	1 if $k \leq n$; 0 otherwise
$\frac{1}{1 - x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots$	1

Useful Generating Functions

TABLE 1 Useful Generating Functions.

$G(x)$	a_k
$\frac{1}{1 - ax} = \sum_{k=0}^{\infty} a^k x^k = 1 + ax + a^2 x^2 + \dots$	a^k
$\frac{1}{1 - x^r} = \sum_{k=0}^{\infty} x^{rk} = 1 + x^r + x^{2r} + \dots$	1 if $r \mid k$; 0 otherwise
$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} (k+1)x^k = 1 + 2x + 3x^2 + \dots$	$k+1$
$\frac{1}{(1-x)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)x^k$ $= 1 + C(n, 1)x + C(n+1, 2)x^2 + \dots$	$C(n+k-1, k) = C(n+k-1, n-1)$
$\frac{1}{(1+x)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)(-1)^k x^k$ $= 1 - C(n, 1)x + C(n+1, 2)x^2 - \dots$	$(-1)^k C(n+k-1, k) = (-1)^k C(n+k-1, n-1)$
$\frac{1}{(1-ax)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)a^k x^k$ $= 1 + C(n, 1)ax + C(n+1, 2)a^2 x^2 + \dots$	$C(n+k-1, k)a^k = C(n+k-1, n-1)a^k$
$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$	$1/k!$
$\ln(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$	$(-1)^{k+1}/k$

Note: The series for the last two generating functions can be found in most calculus books when power series are discussed.

Solving Recurrence Relation

Solve $a_k = 3a_{k-1}$ for $k = 1, 2, 3 \dots$. With initial condition $a_0 = 2$



Let $G(x)$ be the generating function for the sequence $\{a_k\}$, $G(x) = \sum_{k=0}^{\infty} a_k x^k$

$$xG(x) = \sum_{k=0}^{\infty} a_k x^{k+1} = \sum_{k=1}^{\infty} a_{k-1} x^k$$

$$\begin{aligned} G(x) - 3xG(x) &= \sum_{k=0}^{\infty} a_k x^k - 3 \sum_{k=1}^{\infty} a_{k-1} x^k \\ &= a_0 + \sum_{k=1}^{\infty} (a_k - 3a_{k-1}) x^k = 2 \end{aligned}$$

Hence $(1 - 3x)G(x) = 2$.

$$\text{In summary } G(x) = \frac{2}{1-3x}$$

$$\text{Hence } G(x) = \sum_{k=0}^{\infty} 2 \cdot 3^k x^k$$

Solving Recurrence relation

Use generating functions to find an explicit formula for $a_n = 8a_{n-1} + 10^{n-1}$ and the initial condition $a_0 = 1$

$$a_n x^n = 8a_{n-1}x^n + 10^{n-1}x^n$$

Let $G(x) = \sum_{\{n=0\}}^{\infty} a_n x^n$ be the generating function of the sequence a_0, a_1, a_2, \dots . We sum both sides of the last equation starting with $n = 1$, to find that

$$\begin{aligned} G(x) &= 1 + \sum_{n=1}^{\infty} a_n x^n = 1 + 8 \sum_{n=1}^{\infty} (a_{n-1} x^n) + \sum_{n=1}^{\infty} (10^{n-1} x^n) \\ &= 1 + 8x \sum_{n=1}^{\infty} (a_{n-1} x^{n-1}) + x \sum_{n=1}^{\infty} (10^{n-1} x^{n-1}) = 1 + 8x \sum_{n=0}^{\infty} (a_n x^n) + x \sum_{n=0}^{\infty} (10^n x^n) \\ &= 1 + 8xG(x) + \frac{x}{1-10x} \end{aligned}$$

$$G(x) - 8xG(x) = \frac{1-9x}{1-10x}$$

$$G(x) = \frac{(1-9x)}{(1-8x)(1-10x)} = \frac{1}{2} \left(\frac{1}{1-8x} + \frac{1}{1-10x} \right) = \frac{1}{2} \left(\sum_{n=0}^{\infty} 8^n x^n + \sum_{n=0}^{\infty} 10^n x^n \right)$$

$$a_n = \frac{8^n + 10^n}{2}$$

Extended Binomial Coefficient



Let u be a real number and k be a nonnegative integer. Then the extended binomial coefficient $\binom{u}{k}$ is defined by

$$\binom{u}{k} = \begin{cases} \frac{u(u-1)\dots(u-k+1)}{k!} & \text{if } k > 0 \\ 1 & \text{if } k = 0 \end{cases}$$

Theorem 2 : Extended Binomial Theorem



Let x be a real number $|x| < 1$ and let u be a real number. Then

$$(1 + x)^u = \sum_{k=0}^{\infty} \binom{u}{k} x^k$$

Find the generating functions for $(1 + x)^{-n}$ and $(1 - x)^{-n}$, where n is a positive integer, using the extended binomial theorem.

Counting Problems and Generating Functions

Example: Find the number of solutions of $e_1 + e_2 + e_3 = 17$, where e_1, e_2 , and e_3 are nonnegative integers with $2 \leq e_1 \leq 5$, $3 \leq e_2 \leq 6$, and $4 \leq e_3 \leq 7$.

Solution:

- The number of solutions is the coefficient of x^{17} in the expansion of $(x^2 + x^3 + x^4 + x^5)(x^3 + x^4 + x^5 + x^6)(x^4 + x^5 + x^6 + x^7)$.
- This follows because a term equal to x^{17} is obtained in the product by picking a x^{e_1} term in the first sum x^{e_1} , a term in the second sum x^{e_2} , and a term in the third sum x^{e_3} , where $e_1 + e_2 + e_3 = 17$.
- There are three solutions since the coefficient of x^{17} in the product is 3.

Counting Problems and Generating Functions



Example: Use generating functions to find the number of k -combinations of a set with n elements, i.e., $C(n, k)$.

Solution: Each of the n elements in the set contributes the term $(1 + x)$ to the generating function

$$f(x) = \sum_{k=0}^n a^k x^k.$$

Hence $f(x) = (1 + x)^n$ where $f(x)$ is the generating function for $\{a^k\}$, where a^k represents the number of k -combinations of a set with n elements.

By the binomial theorem, we have

where
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

$$f(x) = \sum_{k=0}^n \binom{n}{k} x^k,$$

Hence,

$$C(n, k) = \frac{n!}{k!(n-k)!}.$$

Counting Problems and Generating Functions

In how many different ways can eight identical cookies be distributed among three distinct children if each child receives at least two cookies and no more than four cookies?

$$(x^2 + x^3 + x^4)^3$$

.



BITS Pilani

Pilani Campus

Mathematical Foundations for Data Science

MFDS Team



DSECL ZC416, MFDS

Lecture No. 16

Agenda

- Generating functions
- Extended binomial theorem
- Counting problems and generating functions
- Solving recurrence relations using generating functions
- Proving Identities via Generating Functions

Generating Functions

Definition: The *generating function for the sequence $a_0, a_1, \dots, a_k, \dots$* of real numbers is the infinite series

$$G(x) = \underline{a_0} + \underline{a_1}x + \cdots + \underline{a_k}x^k + \cdots = \sum_{k=0}^{\infty} \underline{a_k}x^k.$$

Examples:

- The sequence $\{a_k\}$ with $a_k = \underline{3}$ has the generating function $\sum_{k=0}^{\infty} \underline{3}x^k$. 3+3x+3x²...
- The sequence $\{a_k\}$ with $a_k = \underline{k+1}$ has the generating function $\sum_{k=0}^{\infty} (k+1)x^k$.
- The sequence $\{a_k\}$ with $a_k = \underline{2^k}$ has the generating function $\sum_{k=0}^{\infty} \underline{2^k}x^k$.

- Generating functions for finite sequences of real numbers can be defined by extending a finite sequence a_0, a_1, \dots, a_n into an infinite sequence by setting $\underline{a_{n+1} = 0, a_{n+2} = 0, \text{ and so on.}}$
- The generating function $G(x)$ of this sequence $\{a_n\}$ is a polynomial of degree n because no terms of the form $a_j x^j$ with $j > n$ occur, that is,

$$G(x) = a_0 + a_1 x + \cdots + a_n x^n + a_{n+1} x^{n+1} + a_{n+2} x^{n+2}$$

\uparrow

$$a_k = 0 \quad \underline{\underline{+ R7/n+1}}$$

Generating Functions for Finite Sequences (continued)

Example: What is the generating function for the sequence
1,1,1,1,1,1?

Solution: The generating function of 1,1,1,1,1,1 is

$$1 + x + x^2 + x^3 + x^4 + x^5.$$

we have

$$(x^6 - 1)/(x - 1) = 1 + x + x^2 + x^3 + x^4 + x^5$$

when $x \neq 1$.

Consequently $G(x) = (x^6 - 1)/(x - 1)$ is the generating function of the sequence.

Useful Facts : Power Series

The function $\frac{1}{1-x}$ is the generating function of the sequence 1,1,1,1,1.....

Because $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$ for $|x| < 1$

—

The function $\frac{1}{1-ax}$ is the generating function of the sequence 1, a, a^2, a^3, \dots

Because $\frac{1}{1-ax} = 1 + ax + a^2x^2 + a^3x^3 + \dots$ for $|x| < \frac{1}{a}$ for $a \neq 0$

—

$$\sum a^k x^k \quad a_0 = 1 \quad |a x| < 1$$

Theorem 1

Let $f(x) = \sum_{k=0}^{\infty} a_k x^k$ and $\underline{g(x)} = \sum_{k=0}^{\infty} b_k x^k$. then

$$\underline{f(x) + g(x)} = \sum_{k=0}^{\infty} (\underline{a_k + b_k}) x^k \text{ and } \underline{f(x)g(x)} = \sum_{k=0}^{\infty} \left(\sum_{j=0}^k a_j \cdot b_{k-j} \right) x^k$$

Extended Binomial Coefficient

Let u be a real number and k be a nonnegative integer. Then the extended binomial coefficient $\binom{u}{k}$ is defined by

$$\binom{u}{k} = \begin{cases} \frac{u(u-1)\dots(u-k+1)}{k!} & \text{if } k > 0 \\ 1 & \text{if } k = 0 \end{cases}$$

u C k

Theorem 2 : Extended Binomial Theorem

-1, -2, ...

Let x be a real number $|x| < 1$ and let u be a real number. Then

$$(1+x)^n = \sum_{k=0}^{\infty} \binom{n}{k} x^k \rightarrow BT$$

$\underbrace{\qquad\qquad}_{n+1 \text{ term}} \underbrace{\qquad\qquad}_{\text{by } x}$

$$(1+x)^u = \sum_{k=0}^{\infty} \binom{u}{k} x^k$$

$a_n \quad G(x) = \sum_{n=0}^{\infty} a_n x^n$

$|x| < 1 \quad \text{Converges for infinite series}$

(2+7)

Find the generating functions for $(1 + x)^{-n}$ and $(1 - x)^{-n}$, where n is a positive integer, using the extended binomial theorem.

$$(1+x)^{-n} = ? \quad (1+x)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} x^k$$

Counting Problems and Generating Functions



Example: Find the number of solutions of $e_1 + e_2 + e_3 = 17$, where e_1, e_2 , and e_3 are nonnegative integers with $2 \leq e_1 \leq 5, 3 \leq e_2 \leq 6$, and $4 \leq e_3 \leq 7$.

Solution:

- The number of solutions is the coefficient of $\underline{x^{17}}$ in the expansion of
 - $(x^2 + x^3 + x^4 + x^5)(x^3 + x^4 + x^5 + x^6)(x^4 + x^5 + x^6 + x^7)$.
- This follows because a term equal to x^{17} is obtained in the product by picking a x^{e_1} term in the first sum x^{e_1} , a term in the second sum x^{e_2} , and a term in the third sum x^{e_3} , where $e_1 + e_2 + e_3 = 17$.
- There are three solutions since the coefficient of x^{17} in the product is 3.

Counting Problems and Generating Functions

In how many different ways can eight identical cookies be distributed among three distinct children if each child receives at least two cookies and no more than four cookies?

$$(x^2 + x^3 + x^4)^3 - x^8$$
$$(x^2 + x^3 + x^4)(x^2 + x^3 + x^4)(x^2 + x^3 + x^4)$$

8

2, 3, 4

6

3, 3, 2 — (3)

2, 2, 4 — (3)

Useful Generating Functions

TABLE 1 Useful Generating Functions.

$G(x)$	a_k
$(1 + x)^n = \sum_{k=0}^n C(n, k)x^k$ $= 1 + C(n, 1)x + C(n, 2)x^2 + \cdots + x^n$	$C(n, k)$
$(1 + ax)^n = \sum_{k=0}^n C(n, k)a^kx^k$ $= 1 + C(n, 1)ax + C(n, 2)a^2x^2 + \cdots + a^nx^n$	$C(n, k)a^k$
$(1 + x^r)^n = \sum_{k=0}^n C(n, k)x^{rk}$ $= 1 + C(n, 1)x^r + C(n, 2)x^{2r} + \cdots + x^{rn}$	$C(n, k/r)$ if $r \mid k$; 0 otherwise
$\frac{1 - x^{n+1}}{1 - x} = \sum_{k=0}^n x^k = 1 + x + x^2 + \cdots + x^n$	1 if $k \leq n$; 0 otherwise
$\frac{1}{1 - x} = \sum_{k=0}^{\infty} x^k = 1 + x + x^2 + \cdots$	1

Useful Generating Functions

TABLE 1 Useful Generating Functions.

$G(x)$	a_k
$\frac{1}{1 - ax} = \sum_{k=0}^{\infty} a^k x^k = 1 + ax + a^2 x^2 + \dots$	a^k
$\frac{1}{1 - x^r} = \sum_{k=0}^{\infty} x^{rk} = 1 + x^r + x^{2r} + \dots$	1 if $r \mid k$; 0 otherwise
$\frac{1}{(1-x)^2} = \sum_{k=0}^{\infty} (k+1)x^k = 1 + 2x + 3x^2 + \dots$	$k+1$
$\frac{1}{(1-x)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)x^k$ $= 1 + C(n, 1)x + C(n+1, 2)x^2 + \dots$	$C(n+k-1, k) = C(n+k-1, n-1)$
$\frac{1}{(1+x)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)(-1)^k x^k$ $= 1 - C(n, 1)x + C(n+1, 2)x^2 - \dots$	$(-1)^k C(n+k-1, k) = (-1)^k C(n+k-1, n-1)$
$\frac{1}{(1-ax)^n} = \sum_{k=0}^{\infty} C(n+k-1, k)a^k x^k$ $= 1 + C(n, 1)ax + C(n+1, 2)a^2 x^2 + \dots$	$C(n+k-1, k)a^k = C(n+k-1, n-1)a^k$
$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \dots$	$1/k!$
$\ln(1+x) = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} x^k = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots$	$(-1)^{k+1}/k$

Note: The series for the last two generating functions can be found in most calculus books when power series are discussed.

Solving Recurrence relation

Solve $\underline{a_k = 3a_{k-1}}$ for $k = 1, 2, 3 \dots$ With initial condition $\underline{a_0 = 2}$.

Answer:

Let $G(x)$ be the generating function for the sequence $\{a_k\}$, that is

$$G(x) = \sum_{k=0}^{\infty} a_k x^k \text{ and}$$

$$xG(x) = \sum_{k=0}^{\infty} a_k x^{k+1} = \sum_{k=1}^{\infty} a_{k-1} x^k$$

$$G(x) - 3xG(x) = \sum_{k=0}^{\infty} a_k x^k - 3 \sum_{k=1}^{\infty} a_{k-1} x^k = a_0 + \sum_{k=1}^{\infty} (a_k - 3a_{k-1}) x^k = 2$$

$$\text{Hence } (1 - 3x)G(x) = 2.$$

$$\text{In summary } G(x) = \frac{2}{1-3x},$$

$$\text{Hence } G(x) = \sum_{k=0}^{\infty} 2 \cdot 3^k x^k$$

Solving Recurrence relation

$a_n = 8a_{n-1} + 10^{n-1}$ and the initial condition $a_0 = 1$. Use generating functions to find an explicit formula for a_n .

$$a_n x^n = 8a_{n-1} x^n + 10^{n-1} x^n$$

Let $G(x) = \sum_{n=0}^{\infty} a_n x^n$ be the generating function of the sequence a_0, a_1, a_2, \dots . We sum both sides of the last equation starting with $n = 1$, to find that

$$G(x) - 1 = \sum_{n=0}^{\infty} a_n x^n = \sum_{k=0}^{\infty} 8a_{k-1} x^k = 10^{k-1} x^k = 8xG(x) + \frac{x}{1-10x}$$

$$G(x) = \frac{(1-9x)}{(1-8x)(1-10x)} = \frac{1}{2} \left(\frac{1}{1-8x} + \frac{1}{1-10x} \right) = \frac{1}{2} \left(\sum_{n=0}^{\infty} 8^n x^n + \sum_{n=0}^{\infty} 10^n x^n \right)$$

$$a_n = \frac{8^n + 10^n}{2}$$