MAE 507 - Engineering Analysis

Joseph Marziale June 29, 2023

Contents

1 Mod1 Linear algebra

1.1 Lec 1a Matrices

Matrix equations

$$\mathbf{A}\mathbf{x} = \mathbf{b} \tag{1}$$

represent sets of linear algebraic equations or LAEs. $\mathbf{A}$ is the system matrix with shape mxn, where m is rows and n is columns. $\mathbf{x}$ is unknown vector with shape nx1, and $\mathbf{b}$ is

the known vector with shape nx1.

The matrix eigenproblem is

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{2}$$

$\mathbf{A}$ is the nxn system matrix, λ is a scalar eigenvalue, and $\mathbf{x}$ is the nx1 eigenvector. This is saying that there is some $\mathbf{x}\lambda$ such that a system matrix $\mathbf{A}$ transforms $\mathbf{x}$ into a vector parallel to itself $\lambda\mathbf{x}$. Rearranged,

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

Solving Eq. 1 can be done by isolating the unknown $\mathbf{x}$, so that

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}, \tag{4}$$

provided $\mathbf{A}^{-1}, \mathbf{b}$ are known or able to be calculated.

If a matrix $\mathbf{A}$ is mxn, and if m=n so that $\mathbf{A}$ is actually nxn, then the matrix is square. If $\mathbf{A}^T = \mathbf{A} \leftrightarrow A_{ij} = A_{ji}$ then $\mathbf{A}$ is called symmetric. If $A^T = -\mathbf{A} \leftrightarrow A_{ij} = -A_{ji}$ then $\mathbf{A}$ is called skew symmetric. In general $\mathbf{A}$ can always be broken down into symmetric and skew parts, in that

$$\mathbf{A} = \mathbf{A}_{sym} + \mathbf{A}_{skew} = \frac{1}{2}(\mathbf{A} + \mathbf{A}^T) + \frac{1}{2}(\mathbf{A} - \mathbf{A}^T). \tag{5}$$

This is shown using the simple 2d example

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix} A_{11} + A_{11} & A_{12} + A_{21} \\ A_{21} + A_{12} & A_{22} + A_{22} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} A_{11} - A_{11} & A_{12} - A_{21} \\ A_{21} - A_{12} & A_{22} - A_{22} \end{bmatrix}. \tag{6}$$

orthogonality of a system martrix $\mathbf{A}$ is defined by

$$\mathbf{A}^{-1} = \mathbf{A}^T, \tag{7}$$

and identity is defined by

$$\mathbf{A}^{-1} = \mathbf{A} = \mathbf{I}. \tag{8}$$

Matrices are communitative with respect to addition so that $\mathbf{B}+\mathbf{A} = \mathbf{A}+\mathbf{B}$, but are not with respect to multiplication so that $\mathbf{BA} \neq \mathbf{AB}$. The transpose of multiple matrices is

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T, \tag{9}$$

$$(\mathbf{A}_1\mathbf{A}_2 \ldots \mathbf{A}_{n-1}\mathbf{A}_n)^T = \mathbf{A}_n^T\mathbf{A}_{n-1}^T \ldots \mathbf{A}_2^T\mathbf{A}_1^T. \tag{10}$$

1.2 Lec 1b Norms and determinants

1.2.1 Norms

A norm is a measure of a vector. Euclidean or L_2 norm

$$||\mathbf{x}|| = (x_1^2 + x_2^2 + \ldots + x_n^2)^{1/2}. \tag{11}$$

It is a rating of the vector length. Properties of the Euclidean norm are

- $||\mathbf{x}|| \geq 0$. $||\mathbf{x}|| = 0$ iff $\mathbf{x} = \mathbf{0}$.
- $||k\mathbf{x}|| = k||\mathbf{x}||$ for all k.
- $||\mathbf{x} + \mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ for all $\mathbf{x}, \mathbf{y}$ of same dimension n (Triangle equality). The hypotenuse of the triangle represents the LHS and the opposite/adjacent sides represent the RHS.

Another representation of the L_2 norm is

$$||\mathbf{x}||_2 = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}. \tag{12}$$

In general, the L_p norm is

$$||\mathbf{x}||_p = \left(\sum_{i}^{n} x_i^p \right)^{1/p} \tag{13}$$

For matrices, the L_2 norm is

$$||\mathbf{A}||_2 = \left(\sum_{i}^{n} \sum_{j}^{n} A_{ij}^2 \right)^{1/2}. \tag{14}$$

Of an eigenproblem, the L_e norm of system matrix $\mathbf{A}$ is

$$||\mathbf{A}||_e = \max(\lambda_i), \tag{15}$$

or the greatest of the eigenvalues.

1.2.2 Determinants

Determinant of 2x2

$$\det \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 1 * 4 - 3 * 2 = -2. \tag{16}$$

Determinant of 3x3

$$\det \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = 1(5 * 9 - 8 * 6) - 2(4 * 9 - 7 * 6) + 3(4 * 8 - 7 * 5). \tag{17}$$

It is also permissible to do

$$\det \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = -4(2 * 9 - 8 * 3) + 5(1 * 9 - 7 * 3) - 6(1 * 8 - 7 * 2). \tag{18}$$

The coefficients must be $\sum_{j=1}^{n} a_{ij}(-1)^{i+j}$. In other words, some row i is picked as the set of coefficients, and the row is summed across the columsn j. The sign of the coefficient depends on the specific placement of the element: if $i + j$ is even, then the coefficient is

positive, but if $i+j$ is odd, then it is negative. Then the terms inside the parentheses are called the minor of a_{ij} denoted as M_{ij}. For a 3x3 matrix they are defined as the 2x2 matrix of elements not found in row i or column j. In total the determinant of $\mathbf{A}$ is

$$\det \mathbf{A} = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} M_{ij} = \sum_{j=1}^{n} a_{ij}\beta_{ij}, \tag{19}$$

where $\beta_{ij} = (-1)^{i+j} M_{ij}$.

Properties of determinants are

- $\det \mathbf{A} = \det \mathbf{A}^T$,
- An entire row or column of zeros in $\mathbf{A}$ implies $\det \mathbf{A} = 0$.
- Two proportional rows or columns implies singularity/linear dependence/$\det \mathbf{A} = 0$.
- Interchanging two rows switches the determinant's sign.
- Multiplying a row or column by a scalar multiplies the determinant by that scalar.
- Multiplying all n rows (thus multiplying the entire matrix) by a scalar c is the same as $\det(c\mathbf{A}) = c^n \det \mathbf{A}$.
- Adding a row by a multiple of another row does not change the determinant.

1.3 Lec 1c Linear algebraic equations (LAEs)

1.3.1 Nonsingular unique trivial

An example of a set of LAEs is

$$3x + 2y = 0, \quad -x + 5y = 0 \tag{20}$$

$$\rightarrow \begin{bmatrix} 3 & 2 \\ -1 & 5 \end{bmatrix} \begin{Bmatrix} x \\ y \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}. \tag{21}$$

Another representation is

$$\begin{bmatrix} 3 & 2 \\ -1 & 5 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}. \tag{22}$$

This way $\mathbf{Ax} = \mathbf{b} = \mathbf{0}$. A simple, trivial solution $x = \{0, 0\}^T$ exists. The determinant determines whether or not this is the only solution. $\det \mathbf{A} \neq 0$ implies the solution is unique/nonsingular. $\det \mathbf{A} = 0$ implies the solution is singular/linearly dependent.

1.3.2 Singular nonunique trivial/nontrivial

Another example is

$$\begin{bmatrix} 3 & 2 \\ -1 & -2/3 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \end{Bmatrix}. \tag{23}$$

Here $\det \mathbf{A} = 0$ meaning the matrix is singular/nonunique/linearly dependent. Therefore there are many solutions including the trivial solution. Particularly every point on the line is a solution.

1.3.3 Nonsingular unique nontrivial

Another example is

$$\begin{bmatrix} 3 & 2 \\ -1 & 5 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 4 \\ -6 \end{Bmatrix}. \tag{24}$$

Here $\det \mathbf{A} \neq 0$ and $\mathbf{x} = \mathbf{0}$ is not a solution. So the matrix is nonsingular/unique/linearly independent and the unique solution is nontrivial.

1.3.4 Singular nonunique nontrivial

Another example is

$$\begin{bmatrix} 3 & 2 \\ -1 & -2/3 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 4 \\ -4/3 \end{Bmatrix}. \tag{25}$$

Here $\det \mathbf{A} = 0$ but $\mathbf{x} = \mathbf{0}$ is not a solution. Therefore the matrix is linearly dependent/singular/nonunique and all the solutions are nontrivial.

1.3.5 Consistency vs. inconsistency

The matrix in Eq. 25 is singular and the solution elements $\mathbf{b}$ are similarly proportional to one another. Thuis means that they describe the same equation and so they are consistent. However, if $\mathbf{b}$ had elements that were not proportional then this would be describing equations with the same y, x, m but different b, meaning they are parallel but have different $y \leftrightarrow x_2$ intercepts. Because they are not the same equation, they are inconsistent.

1.3.6 Homogeneity vs nonhomogeneity

If the known vector $\mathbf{b}$=$\mathbf{0}$, then the equation system is homogeneous. If not, then it is nonhomogeneous.

1.4 Lec 1d Terminology and solution methods

To recap, $\mathbf{b} \neq \mathbf{0}$ implies nonhomogeneity of the system, and $\mathbf{x} \neq \mathbf{0}$ implies nontriviality of the solution.

1.4.1 Augmented matrix

An augmented matrix $[\mathbf{A}|\mathbf{b}] \longrightarrow^{\text{Gaussian elimination}} \longrightarrow [\mathbf{I}|\mathbf{x}]$. That is, Gaussian elimination methods can be used to transform the augmented matrix into an augmentation of the identity matrix and the solution vector. This is because $[\mathbf{A}|\mathbf{b}] \rightarrow [\mathbf{A}^{-1}\mathbf{A}|\mathbf{A}^{-1}\mathbf{b}] \rightarrow [\mathbf{I}|\mathbf{x}]$.

1.4.2 Rank

- If the largest square submatrix with a nonzero determinant has size mxm, then the rank of the matrix is m. If the entire matrix has a nonzero determinant then the rank is just the size.

- If the rank of a matrix **A**is the same as the rank of the augmented matrix $[\mathbf{A}|\mathbf{b}]$ then the solution to $\mathbf{Ax} = \mathbf{b}$ exists. Otherwise, there are no solutions. For example,

$$\text{rank}[\mathbf{A}|\mathbf{b}] = \text{r}\begin{bmatrix} 3 & 2 & 4 \\ -1 & -3/2 & 2 \end{bmatrix} = 2 \leftrightarrow 2 * 2 + (3/2) * 4 \neq 0, \tag{26}$$

$$\text{rank}\mathbf{A} = \text{r}\begin{bmatrix} 3 & 2 \\ -1 & -3/2 \end{bmatrix} = 1 \leftrightarrow 3 * (-3/2) + 1 * 2 = 0. \tag{27}$$

  Here the ranks are not equal and so the system

$$\mathbf{Ax} = \mathbf{b} \leftrightarrow \begin{bmatrix} 3 & 2 \\ -1 & -3/2 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} 4 \\ 2 \end{Bmatrix} \tag{28}$$

  has no solution (they are parallel equations).

- If **A** is nxn and r**A** =r$[\mathbf{A}|\mathbf{b}] = n$, then the solution is unique. In this case $\mathbf{b} = \mathbf{0} \rightarrow \mathbf{x} = \mathbf{0}$, $\mathbf{b} \neq \mathbf{0} \rightarrow \mathbf{x} \neq \mathbf{0}$. For example

$$\text{rank}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = 2, \tag{29}$$

$$\text{rank}\begin{bmatrix} 1 & 2 & 0 \\ 3 & 4 & 0 \end{bmatrix} = 2 \longrightarrow \mathbf{b} = \mathbf{0} \rightarrow \mathbf{x} = \mathbf{0}. \tag{30}$$

- If **A** is nxn and r**A** =r$[\mathbf{A}|\mathbf{b}] < n$, then many solutions exist. For example

$$\text{r}\begin{bmatrix} 1 & 3 \\ 2 & 6 \end{bmatrix} = 1, \tag{31}$$

$$\text{r}\begin{bmatrix} 1 & 3 & 1 \\ 2 & 6 & 2 \end{bmatrix} = 1 \longrightarrow \text{same equation} \longrightarrow \text{infinite solutions along line.} \tag{32}$$

1.4.3 Linear dependence vs independence

If one of the equations can be written as a linear combination of the other equations in a system, then that system is linearly dependent. Otherwise, it is linearly independent. Linear independence implies full rank.

1.4.4 Cramer's rule

If $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{A}_i$ is the same as **A** except the column i is replaced by **b**, then

$$x_i = \frac{\det \mathbf{A}_i}{\det \mathbf{A}}. \tag{33}$$

For example

$$\mathbf{Ax} = \mathbf{b} \longrightarrow \begin{bmatrix} 3 & 2 \\ -1 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{Bmatrix} 4 \\ -6 \end{Bmatrix} \tag{34}$$

implies

$$\mathbf{A}_1 = \begin{bmatrix} 4 & 2 \\ -6 & 5 \end{bmatrix}, \quad \mathbf{A}_2 = \begin{bmatrix} 3 & 4 \\ -1 & -6 \end{bmatrix} \tag{35}$$

and

$$x_1 = \frac{\det \mathbf{A}_1}{\det \mathbf{A}} = \frac{4*5+6*2}{3*5+1*2} = \frac{32}{17}, \tag{36}$$

$$x_2 = \frac{\det \mathbf{A}_2}{\det \mathbf{A}} = \frac{3*-6+1*4}{3*5+1*2} = \frac{-14}{17}. \tag{37}$$

1.5 Lec 1e Elimination and decomposition methods

The three elementary row operations or EROs are the last three of the list in Sec. 1.2.2 and are

- Multiply row by scalar: $\det \mathbf{A} \to c \det \mathbf{A}$
- Swap rows: $\det \mathbf{A} \to -\det \mathbf{A}$
- Add a linear multiple of one row to another row: $\det \mathbf{A} \to \det \mathbf{A}$.

1.5.1 Gaussian elimination

Gaussian elimination seeks to reduce some system matrix $\mathbf{A}$ into a triangular matrix so that, for example,

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} b_1 \\ b_2 \\ b_3 \end{Bmatrix}. \tag{38}$$

Then

$$A_{33}x_3 = b_3 \to x_3 = \frac{b_3}{A_{33}}, \tag{39}$$

$$A_{22}x_2 + A_{23}x_3 = b_2 \to x_2 = \frac{b_2 - A_{23}x_3}{A_{22}}, \tag{40}$$

$$A_{11}x_1 + A_{12}x_2 + A_{13}x_3 = b_1 \to x_1 = \frac{b_1 - A_{12}x_2 - A_{13}x_3}{A_{11}}. \tag{41}$$

More generally, for a matrix of size n,

$$x_n = \frac{b_n}{A_{nn}}, \tag{42}$$

$$x_{n-1} = \frac{b_{n-1} - A_{n-1,n}x_n}{A_{n-1,n-1}}, \tag{43}$$

and so on until x_1 is solved for.

1.5.2 Gauss Jordan elimination

Gauss Jordan continues using EROs until the identity matrix is the system matrix so that $\mathbf{b} = \mathbf{x}$.

1.5.3 LU decomposition

If $\mathbf{Ax} = \mathbf{b}$ then let $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is a lower triangular matrix and $\mathbf{U}$ is an upper triangular matrix. Then

$$\mathbf{LUx} = \mathbf{b}. \tag{44}$$

If it is supposed that $\mathbf{Ux} = \mathbf{z}$ then the two equations in the order of

$$\mathbf{Lz} = \mathbf{b}, \quad \mathbf{Ux} = \mathbf{z} \tag{45}$$

can be solved for. That is because $\mathbf{L}$,$\mathbf{U}$,$\mathbf{b}$ are known, therefore $\mathbf{z}$ can be solved for, therefore $\mathbf{x}$ can be solved for.

1.5.4 Positive definiteness

This is for symmetric positive definite matrices. A positive matrix $\mathbf{A}$ satisfies the criterion

$$\mathbf{x}^T\mathbf{Ax} > 0 \tag{46}$$

for all $\mathbf{x}$. In 2d, this looks like

$$\begin{Bmatrix} x_1 & x_2 \end{Bmatrix} \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix} = \begin{Bmatrix} x_1 & x_2 \end{Bmatrix} \begin{Bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + x_{22}x_2 \end{Bmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + (a_{12}+a_{21})x_1x_2 > 0. \tag{47}$$

Note that the skew symmetric part of $\mathbf{A}$ goes to zero because $a_{12} = -a_{21}$ and $a_{ii} = -a_{ii} = 0$.

1.5.5 Positive definiteness/Cholesky

If $\mathbf{Ax} = \mathbf{b}$, Cholesky decomposition assumes $\mathbf{A} = \mathbf{LU} = \mathbf{U}^T\mathbf{U}$ so that

$$\mathbf{U}^T \underbrace{\mathbf{Ux}}_{\mathbf{y}} = \mathbf{b}. \tag{48}$$

If $\mathbf{A}$ is known, $\mathbf{U}$ can be found. In 2d,

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} u_{11} & 0 \\ u_{12} & u_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix} = \begin{bmatrix} u_{11}^2 & u_{11}u_{12} \\ u_{12}u_{11} & u_{12}^2 + u_{22}^2 \end{bmatrix} \tag{49}$$

$$\rightarrow a_{11} = u_{11}^2 \rightarrow u_{11} = \sqrt{a_{11}}, \tag{50}$$

$$a_{12} = u_{11}u_{12} \rightarrow u_{12} = \frac{a_{12}}{\sqrt{a_{11}}}, \tag{51}$$

$$a_{22} = u_{12}^2 + u_{22}^2 \rightarrow u_{22} = \sqrt{a_{22} - \frac{a_{12}^2}{a_{11}}}. \tag{52}$$

The general formula is

$$u_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} u_{ki}^2}, \tag{53}$$

$$u_{ij} = \begin{cases} \frac{1}{u_{ii}} \left(a_{ij} - \sum_{k=1}^{i-1} u_{ki} u_{kj} \right), & i < j, \\ 0, & i > j \text{ (upper triangular).} \end{cases} \tag{54}$$

Once $\mathbf{U}$ is known and if $\mathbf{b}$ is known, then the two equations

$$\mathbf{U}^T \mathbf{y} = \mathbf{b}, \quad \mathbf{U}\mathbf{x} = \mathbf{y} \tag{55}$$

can be solved for in the listed order to find $\mathbf{y}$, then $\mathbf{x}$.

Matrices that are not symmetric and positive definite cannot be decomposed with the Cholesky method.

The operation counts of some popular linear solvers are

- Cramer's: $n!$ $(20! \approx 2 \times 10^{18})$
- Gauss: $n^3/3$ $(20^3/3 = 2600)$
- Gauss Jordan: $n^3/2$ $(20^3/2 = 4000)$
- Cholesky: $n^3/6$ $(20^3/6 = 13000)$.

1.6 Lec 1f Determinants and iterative methods

1.6.1 Using Gaussian elimination

The determinant of a triangular matrix is the product of the diagonal entries. For this reason it is efficient to use Gauss elimintation to turn $\mathbf{A}$ into a triangular matrix $\bar{\mathbf{A}}$. Then $\det \mathbf{A} = (-1)^{n_p} \det \bar{\mathbf{A}}$, where n_p is the number of row swapping operations done. (This assumes none of the rows were multiplied by a constant. This would not be necessary unless Gauss Jordan was being done.) This implies

$$\det \mathbf{A} = (-1)^{n_p} \prod_i \bar{A}_{ii}. \tag{56}$$

1.6.2 Using LU decomposition

$$\det \mathbf{A} = \det \mathbf{L}\mathbf{U} = \det \mathbf{L} \det \mathbf{U}. \tag{57}$$

Again for triangular matrices, $\det \mathbf{U} = \prod_i u_{ii}$. If $\mathbf{U}$ is a unit triangular matrix, then the diagonal entries are 1 and so $\det \mathbf{U} = 1$. Then $\det \mathbf{A} = \prod_i l_{ii}$.

1.6.3 Jacobi iteration

For any $\mathbf{A}$ in $\mathbf{A}\mathbf{x} = \mathbf{b}$ there is the admissible decomposition

$$\mathbf{A} = \mathbf{A}_d + \mathbf{A}_o, \tag{58}$$

where $\mathbf{A}_d$ denotes a matrix containing only the diagonal elements of $\mathbf{A}$ and $\mathbf{A}_o$ denotes a matrix containing all of the other elements with zeros on the diagonal. Then

$$(\mathbf{A}_d + \mathbf{A}_0)\mathbf{x} = \mathbf{b} \tag{59}$$

implies

$$\mathbf{A}_d\mathbf{x} = -\mathbf{A}_0\mathbf{x} + \mathbf{b} \tag{60}$$

which implies

$$\mathbf{x} = \mathbf{A}_d^{-1}(-\mathbf{A}_o\mathbf{x} + \mathbf{b}) \tag{61}$$

in which $\mathbf{A}_d^{-1}$ is simply a matrix containing elements a_{ii}^{-1} on the diagonals. Given an initial guess $\mathbf{x}_{(0)}$, Jacobi iteration goes like

$$\mathbf{x}_{(1)} = \mathbf{A}_d^{-1}(-\mathbf{A}_o\mathbf{x}_{(0)} + \mathbf{b}), \tag{62}$$

$$\mathbf{x}_{(2)} = \mathbf{A}_d^{-1}(-\mathbf{A}_o\mathbf{x}_{(1)} + \mathbf{b}), \tag{63}$$

etc. Iteration continues until

$$||\mathbf{x}_{(k+1)} - \mathbf{x}_{(k)}|| \leq \epsilon_x \tag{64}$$

where ϵ_x denotes a convergence criterion or a tolerance. It says, if the difference between iterations is sufficiently small, stop iterating because an approximate solution has been reached.

This system will converge if $\mathbf{A}$ is diagonally dominant, i.e. if $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for all i.

1.6.4 Gauss Seidel iteration

Jacobi updates all $x_i \in \mathbf{x}$ simultaneously. The Gauss Seidel method on the other hand updates x_i one at a time. Given some $\mathbf{x}_{(0)}$, iteration is represented by

$$x_{i,(k+1)} = a_{ii}^{-1}\left(b_i - \sum_{j<i} a_{ij}x_{j,(k+1)} - \sum_{j>i} a_{ij}x_{j,(k)}\right). \tag{65}$$

Iteration step by step in 2d goes like

$$\mathbf{x}_{(1)} = \begin{cases} x_{1,(1)} = a_{11}^{-1}\left(b_1 - \cancel{\sum_{j<1} a_{ij}x_{j,(1)}} - a_{12}x_{2,(0)}\right), \\ x_{2,(1)} = a_{22}^{-1}\left(b_2 - a_{21}x_{1,(1)} - \cancel{\sum_{j>i} a_{ij}x_{j,(0)}}\right), \end{cases} \tag{66}$$

$$\mathbf{x}_{(2)} = \begin{cases} x_{1,(2)} = a_{11}^{-1}\left(b_1 - \cancel{\sum_{j<1} a_{ij}x_{j,(2)}} - a_{12}x_{2,(1)}\right), \\ x_{2,(2)} = a_{22}^{-1}\left(b_2 - a_{21}x_{1,(2)} - \cancel{\sum_{j>i} a_{ij}x_{j,(1)}}\right), \end{cases} \tag{67}$$

etc. Diagonal dominance is still required for convergence. Gauss Siedel usually converges faster than Jacobi iteration.

1.6.5 Southwell relaxation method

Eq. 65 can be rewritten as

$$x_{i,(k+1)} = a_{ii}^{-1}\left(b_i - \sum_{j<i} a_{ij}x_{j,(k+1)} - \underbrace{\sum_{j>i}}_{\mathbf{I}} a_{ij}x_{j,(k)}\right)$$

$$\longrightarrow x_{i,(k+1)} = \underbrace{x_{i,(k)}}_{\mathbf{I}} + a_{ii}^{-1}\left(b_i - \sum_{j<i} a_{ij}x_{j,(k+1)} - \underbrace{\sum_{j=1}}_{\mathbf{I}} a_{ij}x_{j,(k)}\right) \tag{68}$$

$$\longrightarrow x_{i,(k+1)} = x_{i,(k)} + \Delta x_{i,(k+1)}. \tag{69}$$

The term $\Delta x_{i,(k+1)}$ is called a correction term and is the term that drives iteration. If this term is weighted by a coefficient ω, which is called the relaxation parameter, then the Southwell relaxation method

$$\longrightarrow x_{i,(k+1)} = x_{i,(k)} + \omega\Delta x_{i,(k+1)} \tag{70}$$

emerges, where

$$\begin{cases} 0 < \omega < 1, & \text{successive under relaxation,} \\ 1, & \text{Gauss Seidel,} \\ 1 < \omega < 2, & \text{successive over relaxation,} \\ \omega > 2, & \text{divergence.} \end{cases} \tag{71}$$

1.7 Lec 1g Advanced solution methods

1.7.1 Conjugate gradient method

If $\mathbf{A}$ is symmetric and positive definite, let there be some

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} - \mathbf{b}\mathbf{x} \tag{72}$$

so that

$$0 = f'(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} = \nabla\mathbf{f}. \tag{73}$$

Iterate the solution using the Jacobi method in Sec. 1.6.3. During this iteration, let $\mathbf{p}_k = -\nabla\mathbf{f}(\mathbf{x}_k) = \mathbf{b} - \mathbf{A}\mathbf{x}_k$ be the residual at the kth step and

$$\alpha_k = \frac{\mathbf{p}_k^T\mathbf{p}_k}{\mathbf{p}_k\mathbf{A}\mathbf{p}_k}. \tag{74}$$

Then it can be written that $\mathbf{x}_{(k+1)} = \mathbf{x}_{(k)} + \alpha_k\mathbf{p}_k$.

1.7.2 Biconjugate gradient method

"Uses biorthogonality and biconjugacy conditions to establish $\mathbf{p}_k, \alpha_k$ to drive residual to zero."

1.7.3 Preconditioned biconjugate gradient method

If $\mathbf{Ax} = \mathbf{b}$ and Gaussian elimination is done to reduce $\mathbf{A}$ to some diagonally dominant matrix then

$$\hat{\mathbf{A}}\mathbf{Ax} \approx \mathbf{x} = \hat{\mathbf{A}}^{-1}\mathbf{b}. \tag{75}$$

For large systems, the amount of operations done can be cut down dramatically with sparse matrices, or a matrices where most of the elements are zeros except for elements on the diagonal or tridiagonal. The solver can then avoid operating on zeros and prevent the (tri-)diagonal band from growing.

1.8 Lec 1h Matrix eigenproblem

1.8.1 Mass spring systems

If a set of masses in a row m_i moving in distances u_i have a set of springs s_i with coefficients k_i attached to their backs, and springs s_{i+1} with coefficients k_{i+1} attached to their fronts. The total number of contributions to the force on m_i, called F_i, is four. Those are

- u_{i-1} moving forward will shorten s_i and push m_i forward. $\Longrightarrow$
- u_i moving forward will elongate s_i and push m_i backward. $\Longleftarrow$
- u_i moving forward will also shorten s_{i+1} and prevent m_i from moving forward. $\Longleftarrow$
- u_{i+1} moving forward will elongate s_{i+1} and push m_i forward. $\Longrightarrow$

This is because a long s_i works against u_i, but a long s_{i+1} works with u_i. So a shortening of s_i suppresses its negative effect on m_i, but a shortening of s_{i+1} suppresses its positive effect on m_i.

Altogether,

$$F_i = k_i u_{i-1} - k_i u_i - k_{i+1} u_i + k_{i+1} u_{i+1}. \tag{76}$$

Also, Newton's second law states

$$F_i = m_i \ddot{u}_i. \tag{77}$$

Therefore,

$$m_i \ddot{u}_i = k_i u_{i-1} - k_i u_i - k_{i+1} u_i + k_{i+1} u_{i+1} \tag{78}$$

which implies

$$m_i \ddot{u}_i + u_i(k_i + k_{i+1}) - k_i u_{i-1} - k_{i+1} u_{i+1} = 0. \tag{79}$$

If $n = 3$,

$$\begin{bmatrix} m_1 & 0 & 0 \\ 0 & m_2 & 0 \\ 0 & 0 & m_3 \end{bmatrix} \begin{Bmatrix} \ddot{u}_1 \\ \ddot{u}_2 \\ \ddot{u}_3 \end{Bmatrix} + \begin{bmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_1 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{bmatrix} \begin{Bmatrix} u_1 \\ u_2 \\ u_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix}. \tag{80}$$

In the equation for u_1, there is no such term as $u_0 = u_{i-1}$, and u_3 is irrelevant. In the equation for u_3, there is no such term as $u_4 = u_{i+1}$, and u_1 is irrelevant. Represented otherwise,

$$\mathbf{M}\ddot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{0}, \tag{81}$$

where $\mathbf{K}$ is a tridiagonal stiffness matrix and $\mathbf{M}$ is a mass matrix. Now if we let the solution

$$\mathbf{u} = \boldsymbol{\phi} e^{\alpha t} \longrightarrow \ddot{\mathbf{u}} = \alpha^2 \boldsymbol{\phi} e^{\alpha t}, \tag{82}$$

then

$$\alpha^2 \mathbf{M}\boldsymbol{\phi} e^{\alpha t} + \mathbf{K}\boldsymbol{\phi} e^{\alpha t} = \mathbf{0} \longrightarrow \alpha^2 \mathbf{M}\boldsymbol{\phi} + \mathbf{K}\boldsymbol{\phi} = \mathbf{0}. \tag{83}$$

Letting $\alpha = i\omega \longrightarrow \alpha^2 = i^2\omega^2 = -\omega^2$,

$$-\omega^2 \mathbf{M}\boldsymbol{\phi} + \mathbf{K}\boldsymbol{\phi} = \mathbf{0} \longrightarrow (\mathbf{K} - \omega^2 \mathbf{M})\boldsymbol{\phi} = \mathbf{0}. \tag{84}$$

If $\omega^2 = \lambda$, then

$$(\mathbf{K} - \lambda \mathbf{M})\boldsymbol{\phi} = \mathbf{0}, \tag{85}$$

and this is the general matrix eigenproblem with eigenvalues λ and eigenvectors $\boldsymbol{\phi}$. The solution to this equation is either the trivial $\boldsymbol{\phi} = \mathbf{0}$ or the nontrivial

$$0 = \det(\mathbf{K} - \lambda \mathbf{M}). \tag{86}$$

If for example $n = 2, k_1 = k_2 = k, m_1 = m_2 = m$, then

$$0 = \det \begin{bmatrix} k_1 + k_2 - \lambda m_1 & -k_2 \\ -k_1 & k_2 - \lambda m_2 \end{bmatrix} = (k_1 + k_2 - \lambda m_1)(k_2 - \lambda m_2) - k_1 k_2$$

$$= (2k - \lambda m)(k - \lambda m) - k^2$$

$$= k^2 - 3k\lambda m + \lambda^2 m^2 \longrightarrow \lambda = \frac{3km \pm \sqrt{9k^2m^2 - 4m^2k^2}}{2m^2} = \left(\frac{3 \pm \sqrt{5}}{2}\right)\frac{k}{m} = \lambda = \omega^2.$$

If $\omega_0 = \sqrt{k/m} \rightarrow \omega_0^2 = k/m$, then

$$\omega^2 = \begin{cases} \omega_1^2 = (3 + \sqrt{5})\omega_0^2/2 \\ \omega_2^2 = (3 - \sqrt{5})\omega_0^2/2 \end{cases} \longrightarrow \omega = \begin{cases} \omega_1 = 0.618\omega_0 \\ \omega_2 = 1.618\omega_0. \end{cases} \tag{87}$$

Then the eigenvectors $\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ are calculated by plugging the known λ_1, λ_2 into the general matrix eigenproblem Eq. 85.

1.8.2 Stress tensor

The rotation matrix

$$\mathbf{T} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{88}$$

is orthogonal because

$$\mathbf{T}^T\mathbf{T} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} \cos^2\theta + \sin^2\theta & 0 \\ 0 & \cos^2\theta + \sin^2\theta \end{bmatrix} = \mathbf{I} \tag{89}$$

which implies

$$\mathbf{T}^T\mathbf{T}\mathbf{T}^{-1} = \mathbf{I}\mathbf{T}^{-1} \rightarrow \mathbf{T}^T = \mathbf{T}^{-1}. \tag{90}$$

An orthogonal transformation of the stress tensor $\boldsymbol{\sigma}$ is

$$\bar{\boldsymbol{\sigma}} = \mathbf{T}^T\boldsymbol{\sigma}\mathbf{T}. \tag{91}$$

This implies

$$\mathbf{T}\bar{\boldsymbol{\sigma}} = \boldsymbol{\sigma}\mathbf{T} \longrightarrow \boldsymbol{\Phi}\boldsymbol{\Lambda} = \mathbf{A}\boldsymbol{\Phi}. \tag{92}$$

There is some pair of $\mathbf{T}$,$\boldsymbol{\Lambda}$ where $\boldsymbol{\Lambda} = \lambda\mathbf{I}$. Then this becomes the general eigenproblem. If $\boldsymbol{\lambda} \leftrightarrow \bar{\boldsymbol{\sigma}}$ is a diagonal matrix then the entries $\bar{\sigma}$ are the principal stresses/eigenvalues, and the eigenvectors $\mathbf{T}$ are the principal directions. An example of this is

$$\boldsymbol{\sigma} = \begin{bmatrix} 50 & 30 \\ 30 & -20 \end{bmatrix}. \tag{93}$$

Eigenvalues are found through

$$0 = \det \begin{bmatrix} 50-\lambda & 30 \\ 30 & -20-\lambda \end{bmatrix} = (50-\lambda)(-20-\lambda) - 900 = \lambda^2 - 30\lambda - 1900 \tag{94}$$

$$\rightarrow \lambda = \frac{30 \pm \sqrt{2800}}{2} = 61.1, -31.1 = \lambda_{(1)}, \lambda_{(2)}. \tag{95}$$

Eigenvectors are found through

$$\begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = \begin{bmatrix} 50-61.1 & 30 \\ 30 & -20-61.1 \end{bmatrix} \begin{Bmatrix} \phi_{1,(1)} \\ \phi_{2,(1)} \end{Bmatrix}, \tag{96}$$

$$\begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = \begin{bmatrix} 50+31.1 & 30 \\ 30 & -20+31.1 \end{bmatrix} \begin{Bmatrix} \phi_{1,(2)} \\ \phi_{2,(2)} \end{Bmatrix}. \tag{97}$$

Then $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_{(1)}, \boldsymbol{\phi}_{(2)}\}$ and $\boldsymbol{\Lambda} = \begin{bmatrix} \lambda_{(1)} & 0 \\ 0 & \lambda_{(2)} \end{bmatrix}$ so that $\mathbf{A}\boldsymbol{\Phi} = \boldsymbol{\Phi}\boldsymbol{\Lambda}$ or

$$\begin{cases} \mathbf{A}\boldsymbol{\Phi}_{(1)} = \lambda_{(1)}\boldsymbol{\Phi}_{(1)}, \\ \mathbf{A}\boldsymbol{\Phi}_{(2)} = \lambda_{(2)}\boldsymbol{\Phi}_{(2)}. \end{cases} \tag{98}$$

1.8.3 Intertia tensor

The symmetric inertia tensor is

$$\mathbf{G} = \begin{bmatrix} \int_V \rho(y^2+z^2)dV & -\int_V \rho xy dV & -\int_V \rho xz dV \\ -\int_V \rho yx dV & \int_V \rho(x^2+z^2)dV & -\int_V \rho yz dV \\ -\int_V \rho zx dV & -\int_V \rho zy dV & \int_V \rho(x^2+y^2)dV \end{bmatrix} \tag{99}$$

and some rotational transformation is

$$\bar{\mathbf{G}} = \mathbf{T}^T\mathbf{G}\mathbf{T} \rightarrow \mathbf{T}\bar{\mathbf{G}} = \mathbf{G}\mathbf{T} \rightarrow (\mathbf{G} - \bar{\mathbf{G}}\mathbf{I})\mathbf{T} = \mathbf{0}. \tag{100}$$

Here $\bar{\mathbf{G}}$ are the principal inertias and the principal directions are $\mathbf{T}$.

1.8.4 Quadratic forms

The mass matrix $\mathbf{M}$ is positive definite provided $m_i > 0$ and $\mathbf{x}$ is nontrivial. The stiffness matrix $\mathbf{K}$ is positive definite if rigid body motion is prevented. It is positive semi definite if one or more rigid body motion is allowed. The identity matrix $\mathbf{I}$ is real, symmetric, and positive definite. The last statement is true because

$$\mathbf{x}^T\mathbf{I}\mathbf{x} = \mathbf{x}^T\mathbf{x} = \mathbf{x}\cdot\mathbf{x} = x_i^2 > 0 \forall \mathbf{x} \neq \mathbf{0}. \tag{101}$$

1.9 Lec 1i Standard eigenproblem

The standard eigenproblem is

$$(\mathbf{A} - \lambda\mathbf{I})\boldsymbol{\phi} = \mathbf{0}, \tag{102}$$

and the characteristic polynomial in which to find roots λ is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0. \tag{103}$$

If $\mathbf{A}$ is real, then λ are usually complex. However if $\mathbf{A}$ is symmetric then λ are all real. This is because if eigenvalues are complex, then they are conjugates by virtue of the quadratic formula having a $\pm$ discriminant. So if $\mathbf{A} = \mathbf{A}^T$,

$$\mathbf{A}\boldsymbol{\phi} = \lambda\boldsymbol{\phi} \rightarrow \boldsymbol{\phi}^T\mathbf{A}\boldsymbol{\phi} = \boldsymbol{\phi}^T\lambda\boldsymbol{\phi} \rightarrow \lambda = \frac{\boldsymbol{\phi}^T\mathbf{A}\boldsymbol{\phi}}{||\boldsymbol{\phi}||^2} \rightarrow \lambda^* = \frac{\boldsymbol{\phi}^T\mathbf{A}^T\boldsymbol{\phi}}{||\boldsymbol{\phi}||^2} = \frac{\boldsymbol{\phi}^T\mathbf{A}\boldsymbol{\phi}}{||\boldsymbol{\phi}||^2} = \lambda. \tag{104}$$

Since $\lambda = \lambda^*$, the eigenvalues must be real.

1.9.1 orthogonality

Because the eigenvalues are real for symmetric matrices, the eigenvectors are orthogonal. This is because if $\lambda_1, \boldsymbol{\phi}_1$ and $\lambda_2, \boldsymbol{\phi}_2$ are distinct eigenpairs, then

$$\lambda_1\boldsymbol{\phi}_1\cdot\boldsymbol{\phi}_2 = \mathbf{A}\boldsymbol{\phi}_1\cdot\boldsymbol{\phi}_2 \leftrightarrow A_{ij}\phi_{(1)j}\phi_{(2)i} = \phi_{(1)j}A_{ji}\phi_{(2)i} \leftrightarrow \boldsymbol{\phi}_1\cdot\mathbf{A}^T\boldsymbol{\phi}_2 = \boldsymbol{\phi}_1\cdot\mathbf{A}\boldsymbol{\phi}_2 = \boldsymbol{\phi}_1\cdot\lambda_2\boldsymbol{\phi}_2. \tag{105}$$

Therefore,

$$(\lambda_1 - \lambda_2)(\boldsymbol{\phi}_1\cdot\boldsymbol{\phi}_2) = 0, \tag{106}$$

but we assumed $\lambda_1 \neq \lambda_2$, and therefore $\boldsymbol{\phi}_1\cdot\boldsymbol{\phi}_2 = 0$ which implies orthogonality. In this proof the statement

$$\mathbf{A}\mathbf{u}\cdot\mathbf{v} = \mathbf{u}\cdot\mathbf{A}^T\mathbf{v} \longleftrightarrow A_{ij}u_jv_i = u_jA_{ji}v_i \tag{107}$$

was also made, and this how a transposed matrix is defined, so that $A^T_{ij} = A_{ji}$, $[A^T_{ij}] = A_{ji}(\mathbf{e}_i \otimes \mathbf{e}_j)$.

If $\mathbf{A}$ is also positive definite as well as symmetric, then the eigenvalues must be positive because

$$0 < \mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{x}^T\lambda\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x} = \lambda||\mathbf{x}||^2. \tag{108}$$

If the whole term is positive and since x_ix_i must be positive, λ must also be positive, and this is true for any number of eigenpairs because positive definiteness implies this relationship is true for any $\mathbf{x}$.

1.9.2 Spectral decomposition properties

An orthonormal set is defined by the dot products of two of any of the elements being zero and the magnitude of every individual element in the set being one. For example, $([0,1],[1,0])$ is orthonormal because $0*1=1*0=0$ and $\sqrt{0^2+1^2}=\sqrt{1^2+0^2}=1$. Let there be some orthonormal set $\boldsymbol{\Phi}=[\boldsymbol{\phi}_1\boldsymbol{\phi}_2\dots\boldsymbol{\phi}_n]$. Then

$$\boldsymbol{\Phi}^T\boldsymbol{\Phi}=\begin{bmatrix}\boldsymbol{\phi}_1^T\\ \boldsymbol{\phi}_2^T\\ \dots\\ \boldsymbol{\phi}_n^T\end{bmatrix}[\boldsymbol{\phi}_1\boldsymbol{\phi}_2\dots\boldsymbol{\phi}_n]=\begin{bmatrix}\boldsymbol{\phi}_1^T\boldsymbol{\phi}_1 & \boldsymbol{\phi}_1^T\boldsymbol{\phi}_2 & \dots & \boldsymbol{\phi}_1^T\boldsymbol{\phi}_n\\ \boldsymbol{\phi}_2^T\boldsymbol{\phi}_1 & \boldsymbol{\phi}_2^T\boldsymbol{\phi}_2 & \dots & \boldsymbol{\phi}_2^T\boldsymbol{\phi}_n\\ \dots & \dots & \dots & \dots\\ \boldsymbol{\phi}_n^T\boldsymbol{\phi}_1 & \boldsymbol{\phi}_n^T\boldsymbol{\phi}_2 & \dots & \boldsymbol{\phi}_n^T\boldsymbol{\phi}_n\end{bmatrix}=\mathbf{I}. \tag{109}$$

Therefore,

$$\boldsymbol{\Phi}^T=\boldsymbol{\Phi}^{-1}, \tag{110}$$

and this is the definition of an orthogonal tensor/orthogonal matrix. If one eigenpair is represented as

$$\mathbf{A}\boldsymbol{\phi}_i=\lambda_i\boldsymbol{\phi}_i=\boldsymbol{\phi}_i\lambda_i, \tag{111}$$

and this is the standard eigenproblem, then the set of all eigenpairs is represented as

$$\mathbf{A}\boldsymbol{\Phi}=\boldsymbol{\Phi}\boldsymbol{\Lambda}=\big[\mathbf{A}\big]_{3\text{x}3}\big[\boldsymbol{\phi}_1\quad\boldsymbol{\phi}_2\quad\boldsymbol{\phi}_3\big]_{3\text{x}3}=\big[\boldsymbol{\phi}_1\quad\boldsymbol{\phi}_2\quad\boldsymbol{\phi}_3\big]_{3\text{x}3}\begin{bmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{bmatrix}_{3\text{x}3}=\begin{cases}\mathbf{A}\boldsymbol{\phi}_1=\lambda_1\boldsymbol{\phi}_1\\ \mathbf{A}\boldsymbol{\phi}_2=\lambda_2\boldsymbol{\phi}_2\\ \mathbf{A}\boldsymbol{\phi}_3=\lambda_3\boldsymbol{\phi}_3.\end{cases} \tag{112}$$

and this is the matrix eigenproblem. Then because of the orthogonality property Eq. 110,

$$\mathbf{A}=\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T, \tag{113}$$

and this is called the spectral decomposition of $\mathbf{A}$.

If $\mathbf{A}\mathbf{x}=\mathbf{b}$ and the eigenvectors are used as a basis for the set of solutions so that $\mathbf{x}=\boldsymbol{\Phi}\mathbf{c}$ where $\mathbf{c}$ is a coefficient vector, then

$$\mathbf{A}\boldsymbol{\Phi}\mathbf{c}=\mathbf{b}\rightarrow\boldsymbol{\Phi}^T\mathbf{A}\boldsymbol{\Phi}\mathbf{c}=\boldsymbol{\Phi}^T\mathbf{b}\rightarrow\boldsymbol{\Lambda}\mathbf{c}=\boldsymbol{\Phi}^T\mathbf{b}\rightarrow\mathbf{c}=\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}^T\mathbf{b}, \tag{114}$$

where $\boldsymbol{\Lambda}^{-1}$ is simply a matrix with the reciprocals of the eigenvalues on the diagonals.

Now, consider that if the eigenvectors are used as an orthonormal basis/coordinate system then

$$A_{ij}=\boldsymbol{\phi}_i\cdot\mathbf{A}\boldsymbol{\phi}_j. \tag{115}$$

For instance, in the Cartesian coordinate system $\mathbf{e}_i$,

$$\mathbf{e}_1^T\begin{bmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{bmatrix}\mathbf{e}_2=\begin{Bmatrix}1 & 0 & 0\end{Bmatrix}\begin{bmatrix}A_{11} & A_{12} & A_{13}\\ A_{21} & A_{22} & A_{23}\\ A_{31} & A_{32} & A_{33}\end{bmatrix}\begin{Bmatrix}0\\ 1\\ 0\end{Bmatrix}=\begin{Bmatrix}1 & 0 & 0\end{Bmatrix}\begin{Bmatrix}A_{12}\\ A_{22}\\ A_{32}\end{Bmatrix}=A_{12}. \tag{116}$$

Substituting the eigenproblem into Eq. 115,

$$A_{ij}=\boldsymbol{\phi}_i\cdot\lambda_j\boldsymbol{\phi}_j. \tag{117}$$

Then

$$\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \begin{bmatrix} \lambda_1(\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_1) & \lambda_2(\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_2) & \lambda_3(\boldsymbol{\phi}_1 \cdot \boldsymbol{\phi}_3) \\ \lambda_1(\boldsymbol{\phi}_2 \cdot \boldsymbol{\phi}_1) & \lambda_2(\boldsymbol{\phi}_2 \cdot \boldsymbol{\phi}_2) & \lambda_3(\boldsymbol{\phi}_2 \cdot \boldsymbol{\phi}_3) \\ \lambda_1(\boldsymbol{\phi}_3 \cdot \boldsymbol{\phi}_1) & \lambda_2(\boldsymbol{\phi}_3 \cdot \boldsymbol{\phi}_2) & \lambda_3(\boldsymbol{\phi}_3 \cdot \boldsymbol{\phi}_3) \end{bmatrix} = \sum_i \lambda_i(\boldsymbol{\phi}_i \otimes \boldsymbol{\phi}_i) \quad (118)$$

$$= \sum_i \lambda_i \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T = \mathbf{A}. \quad (119)$$

Taking the inverse,

$$\mathbf{A}^{-1} = \sum_i \boldsymbol{\phi}_i^{-T} \boldsymbol{\phi}_i^{-1} \frac{1}{\lambda_i} = \sum_i \boldsymbol{\phi}_i^{T^T} \boldsymbol{\phi}_i^T \frac{1}{\lambda_i} = \sum_i \frac{1}{\lambda_i} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^T = \mathbf{A}^{-1}. \quad (120)$$

Therefore the eigenvectors of $\mathbf{A}$ and its inverse are identical, and the eigenvalues are the reciprocals. Another consequence of the spectral decomposition is

$$\mathbf{A}^2 = \mathbf{A}\mathbf{A} = (\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T)(\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T) = \boldsymbol{\Phi}\boldsymbol{\Lambda}^2\boldsymbol{\Phi}^T \quad (121)$$

and to that effect,

$$\mathbf{A}^n = \boldsymbol{\Phi}\boldsymbol{\Lambda}^n\boldsymbol{\Phi}^T. \quad (122)$$

1.9.3 Functions of square matrices

If there is some function $p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x^1 + a_0$ then we can say

$$p(\mathbf{A}) = a_n\mathbf{A}^n + a_{n-1}\mathbf{A}^{n-1} + \ldots + a_1\mathbf{A} + a_0\mathbf{I} \quad (123)$$

$$= a_n\boldsymbol{\Phi}\boldsymbol{\Lambda}^n\boldsymbol{\Phi}^T + a_{n-1}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{n-1}\boldsymbol{\Phi}^T + \ldots + a_1\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T + a_0\boldsymbol{\Phi}\mathbf{I}\boldsymbol{\Phi}^T = \boldsymbol{\Phi}p(\boldsymbol{\Lambda})\boldsymbol{\Phi}^T. \quad (124)$$

The Cayley Hamilton theorem states that if the characteristic equation to solve for the eigenvalues of system matrix$\mathbf{A}$ is

$$p(x) = \lambda_n + a_{n-1}\lambda^{n-1} + a_{n-2}\lambda^{n-2} + \ldots + a_1\lambda + a_0\mathbf{I} = 0, \quad (125)$$

then the matrix $\mathbf{A}$ also satisfies

$$p(\mathbf{A}) = \mathbf{A}^n + a_{n-1}\mathbf{A}^{n-1} + a_{n-2}\mathbf{A}^{n-2} + \ldots + a_i\mathbf{A} + a_0\mathbf{I} = 0. \quad (126)$$

So $\mathbf{A}$ satisfies its own characteristic equation. This is because the spectral decomposition $\mathbf{A}^m = \boldsymbol{\Phi}\boldsymbol{\Lambda}^m\boldsymbol{\Phi}^T$ can be done and then the whole equation can be premultiplied/postmultiplied by $\boldsymbol{\Phi}^T/\boldsymbol{\Phi}$ to just leave $\boldsymbol{\Lambda}^m$. Then the equation

$$p(\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T) = \boldsymbol{\Phi}\boldsymbol{\Lambda}^n\boldsymbol{\Phi}^T + a_{n-1}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{n-1}\boldsymbol{\Phi}^T + a_{n-2}\boldsymbol{\Phi}\boldsymbol{\Lambda}^{n-2}\boldsymbol{\Phi}^T + \ldots + a_i\boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{\Phi}^T + a_0\mathbf{I} = 0 \quad (127)$$

$$\rightarrow p(\boldsymbol{\Lambda}) = \boldsymbol{\Lambda}^n + a_{n-1}\boldsymbol{\Lambda}^{n-1} + a_{n-2}\boldsymbol{\Lambda}^{n-2} + \ldots + a_i\boldsymbol{\Lambda} + a_0\mathbf{I} = 0 \quad (128)$$

emerges. Then Eq. 128 is a set of three equations corresponding to each eigenvalue. Another consequence of Eq. 126 is that any $\mathbf{A}^n$ can be written in terms of $\sum_{m \leq n} \mathbf{A}^m$. Particularly

$$-(a_{n-1}\mathbf{A}^{n-1} + a_{n-2}\mathbf{A}^{n-2} + \ldots + a_1\mathbf{A} + a_0\mathbf{I}) = \mathbf{A}^n. \quad (129)$$

Multiplying everthing by $\mathbf{A}$,

$$-(a_{n-1}\mathbf{A}^n + a_{n-2}\mathbf{A}^{n-1} + \ldots + a_1\mathbf{A}^2 + a_0\mathbf{A}) = \mathbf{A}^{n+1}. \quad (130)$$

1.10 Lec 1j General eigenproblem

The standard eigenproblem is

$$(\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}, \tag{131}$$

and the general eigenproblem is

$$(\mathbf{A} - \lambda \mathbf{B})\mathbf{x} = \mathbf{0} \longrightarrow \mathbf{A}\mathbf{x} = \lambda \mathbf{B}\mathbf{x}. \tag{132}$$

A,**B** being real and **B** being positive definite constitutes real eigenvalues. There also exist a set of eigenvectors orthonormal with respect to **B**. **A** being positive definite constitutes positive eigenvalues. The general matrix eigenproblem is

$$\mathbf{A}\mathbf{\Phi} = \mathbf{B}\mathbf{\Phi}\mathbf{\Lambda} \tag{133}$$

implying

$$\mathbf{\Phi}^T\mathbf{A}\mathbf{\Phi} = \mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi}\mathbf{\Lambda}. \tag{134}$$

As the eigenvectors are orthonormal with respect to **B**,

$$\mathbf{\Phi}^T\mathbf{B}\mathbf{\Phi} = \mathbf{I} \longrightarrow \mathbf{\Phi}^T\mathbf{A}\mathbf{\Phi} = \mathbf{\Lambda}. \tag{135}$$

1.10.1 Convert general to standard eigenproblem

The conversion of the general eigenproblem to standard form can be done in one of two ways. Of course in general

$$\mathbf{B}^{-1}\mathbf{A}\mathbf{x} = \lambda \mathbf{x}, \tag{136}$$

provided $\mathbf{B}^{-1}$ exists/**B** is invertible/$\det \mathbf{B} \neq 0$. However $\mathbf{B}^{-1}\mathbf{A}$ is rarely symmetric and so this is not a good approach. Instead we can let $\mathbf{B} = \mathbf{U}^T\mathbf{U}$ so that

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{U}^T\mathbf{U}\mathbf{x} \longrightarrow \mathbf{U}^{-T}\mathbf{A}\mathbf{x} = \lambda \mathbf{U}\mathbf{x}. \tag{137}$$

Introducing $\mathbf{Y} = \mathbf{U}\mathbf{x} \rightarrow \mathbf{x} = \mathbf{U}^{-1}\mathbf{Y}$,

$$\underbrace{\mathbf{U}^{-T}\mathbf{A}\mathbf{U}^{-1}}_{\mathbf{D}}\mathbf{Y} = \lambda \mathbf{Y} \longleftrightarrow \underbrace{(\mathbf{U}\mathbf{A}^{-1}\mathbf{U}^T)^{-1}}_{\mathbf{D}}\mathbf{Y} = \lambda \mathbf{Y} \tag{138}$$

$$\rightarrow \mathbf{D}\mathbf{Y} = \lambda \mathbf{Y} \tag{139}$$

in which **D** will be symmetric, provided **A** is symmetric. This is because if a matrix is symmetric then so is its inverse. An example of this is in the calculation of the natural frequencies ω of the mass spring system

$$\mathbf{K}\mathbf{\Phi} = \lambda \mathbf{M}\mathbf{\Phi} \leftrightarrow \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\mathbf{\Phi} = \lambda \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}\mathbf{\Phi} \tag{140}$$

Let $\mathbf{M} = \mathbf{U}^T\mathbf{U}$. Then using Cholesky decomposition in Sec. 1.5.5,

$$\mathbf{U} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \rightarrow \mathbf{U}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} \rightarrow \mathbf{D} = \mathbf{U}^{-T}\mathbf{K}\mathbf{U}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1/2 \end{bmatrix} \tag{141}$$

$$= \begin{bmatrix} 2 & -1/2 \\ -1/2 & 1/2 \end{bmatrix} = \mathbf{D}. \tag{142}$$

Then

$$0 = (\mathbf{D} - \lambda\mathbf{I})\mathbf{Y} \longrightarrow 0 = \det(\mathbf{D} - \lambda\mathbf{I}) = (2-\lambda)(1/2-\lambda) - 1/4 = 3/4 - 5\lambda/2 + \lambda^2 \tag{143}$$

$$\rightarrow \lambda = \frac{5/2 \pm \sqrt{13/4}}{2} = \frac{5 \pm \sqrt{13}}{2}, \tag{144}$$

and the eigenvalues are computed thereafter.

1.10.2 Principal invariants/characteristic equation

Another example is one of the stress tensor $\boldsymbol{\sigma}$ and the principal stresses/eigenvalues in relationship to it $\bar{\boldsymbol{\sigma}}$. The characteristic equation for a 3x3 is

$$-\bar{\sigma}^3 + I_1\bar{\sigma}^2 - I_2\bar{\sigma} + I_3 = 0. \tag{145}$$

The I are called the principal invariants/stress invariants, calculated by

$$I_1 = \mathrm{tr}\boldsymbol{\sigma}, \quad I_2 = \det\begin{bmatrix} \sigma_{22} & \sigma_{23} \\ \sigma_{32} & \sigma_{33} \end{bmatrix} + \det\begin{bmatrix} \sigma_{11} & \sigma_{13} \\ \sigma_{31} & \sigma_{33} \end{bmatrix} + \det\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix}, \quad I_3 = \det\boldsymbol{\sigma}. \tag{146}$$

An assymetric $\mathbf{D}$ will yield complex eigenvalues. Complex matrices will yield real eigenvalues provided $\mathbf{F}$ is Hermitian, meaning that

$$F_{ji} = \bar{F}_{ij}, \tag{147}$$

or that the transpose of $\mathbf{F}$ is equal its complex conjugate.

1.11 Lec 1k Eigensolution methods

1.11.1 Power method

If $\mathbf{Ax} = \lambda\mathbf{x}$ and the eigenvalues have a hierarchy of magnitude $|\lambda_1| < |\lambda_2| < \ldots < |\lambda_n|$ with corresponding eigenvectors $|\boldsymbol{\phi}|_1 < |\boldsymbol{\phi}|_2 < \ldots < |\boldsymbol{\phi}|_n$. The eigenvectors corresponding to distinct eigenvalues are linearly independent because

$$\mathbf{0} = a_1\boldsymbol{\phi}_1 + a_2\boldsymbol{\phi}_2 = \begin{cases} \lambda_1 a_1\boldsymbol{\phi}_1 + \lambda_1 a_2\boldsymbol{\phi}_2 \\ \mathbf{T}a_1\boldsymbol{\phi}_1 + \mathbf{T}a_2\boldsymbol{\phi}_2 = \lambda_1 a_1\boldsymbol{\phi}_1 + \lambda_2 a_2\boldsymbol{\phi}_2 \end{cases} \rightarrow (\lambda_1 - \lambda_2)a_2\boldsymbol{\phi}_2 = 0 \tag{148}$$

$$\rightarrow a_2 = 0 \rightarrow a_1 = 0 \rightarrow \boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \neq \mathbf{0}. \tag{149}$$

In other words, the only solution to $\mathbf{0} = a_1\boldsymbol{\phi}_1, \boldsymbol{\phi}_2$ is the coefficients $a_1, a_2 = 0$, meaning there is no other solution, such as the linear combination of the eigenvectors. This conveys independence.

Since the eigenvectors are linearly independent then any vector

$$\mathbf{x}_1 = c_1\boldsymbol{\phi}_1 + c_2\boldsymbol{\phi}_2 + \ldots + c_n\boldsymbol{\phi}_n = \sum_i c_i\boldsymbol{\phi}_i. \tag{150}$$

Then

$$\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1 = \mathbf{A}\sum_i c_i\boldsymbol{\phi}_i = \sum_i c_i\lambda_i\boldsymbol{\phi}_i \tag{151}$$

serves as an approximation to the eigenproblem $\mathbf{A}\boldsymbol{\phi}_i = \lambda\boldsymbol{\phi}$. Iterating further,

$$\mathbf{x}_3 = \mathbf{A}\mathbf{x}_2 = \mathbf{A}(\sum_i c_i\lambda_i\boldsymbol{\phi}_i) = \sum_i c_i\lambda_i^2\boldsymbol{\phi}_i, \tag{152}$$

$$\mathbf{x}_4 = \sum_i c_i\lambda_i^3\boldsymbol{\phi}_i, \quad \ldots, \tag{153}$$

$$\mathbf{x}_{r+1} = \sum_i c_i\lambda_i^r\boldsymbol{\phi}_i. \tag{154}$$

If r is very large then $\lambda_n >> \sum_{m\neq n}\lambda_m$ and so

$$\mathbf{x}_{r+1} \approx c_n\lambda_r^n\boldsymbol{\phi}_n \tag{155}$$

and so the largest eigenpair n can be found because the process converges to $\boldsymbol{\phi}_n$ associated with λ_n. For example, given

$$\boldsymbol{\sigma} = \begin{bmatrix} 33 & 16 & 18 \\ 16 & -5 & 0 \\ 18 & 0 & 42 \end{bmatrix} \tag{156}$$

we take initial guess

$$\mathbf{n}_1 = \begin{Bmatrix} 0 & 0 & 1 \end{Bmatrix}^T \tag{157}$$

and iterate like

$$\mathbf{n}_2 = \boldsymbol{\sigma}\mathbf{n}_1 = \begin{Bmatrix} 18 & 0 & 42 \end{Bmatrix}^T = 42\begin{Bmatrix} 0.4286 & 0 & 1 \end{Bmatrix}^T, \tag{158}$$

$$\mathbf{n}_3 = \boldsymbol{\sigma}\mathbf{n}_2 = \begin{Bmatrix} 32.1429 & 6.8571 & 49.7143 \end{Bmatrix}^T = 49.7143\begin{Bmatrix} 0.6466 & 0.1379 & 1 \end{Bmatrix}^T, \tag{159}$$

$$\ldots, \mathbf{n}_{12} = \boldsymbol{\sigma}\mathbf{n}_{11} = \begin{Bmatrix} 50.3525 & 12.8466 & 57.7043 \end{Bmatrix} = 57.7043\begin{Bmatrix} 0.8726 & 0.2226 & 1 \end{Bmatrix}^T \approx \bar{\sigma}_n\mathbf{n}_n \tag{160}$$

$$\rightarrow \bar{\sigma}_n \approx 57.7043, \quad \mathbf{n}_n \approx \begin{Bmatrix} 0.8726 & 0.2226 & 1 \end{Bmatrix}^T = \begin{Bmatrix} 0.6485 & 0.1655 & 0.7431 \end{Bmatrix}^T. \tag{161}$$

1.11.2 Inverse power method

The above power method solves for the largest eigenpair. The inverse power method is basically using the power method on the eigenproblem

$$\mathbf{A}^{-1}\boldsymbol{\phi}^{-1} = \frac{1}{\lambda}\boldsymbol{\phi} \tag{162}$$

to reveal the largest eigenpair of that problem which in turn is the smallest eigenpair of the original problem.

1.12 Lec 1l Vector spaces, subspaces I

1.12.1 Vector space rules

Real and complex vectors of size n live in $\mathcal{R}^n$, $\mathcal{C}^n$. Real matrices/tensors of size $m \times n$ live in $\mathcal{R}^{m \times n}$. Vector spaces are collections of vectors with the same dimensions that follow the following rules: if $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$ then

- $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x} \in \mathcal{V}$,
- $\mathbf{x} + (\mathbf{y} + \mathbf{z}) = (\mathbf{x} + \mathbf{y}) + \mathbf{z} \in \mathcal{V}$
- $\exists \mathbf{0} | \mathbf{0} + \mathbf{x} = \mathbf{x} + 0 = \mathbf{x}$,
- $\exists -\mathbf{x} | \mathbf{x} + (-\mathbf{x}) = (-\mathbf{x}) + \mathbf{x} = 0$,
- $a(\mathbf{x} + \mathbf{y}) = a\mathbf{x} + a\mathbf{y} \in \mathcal{V}$,
- $(a + b)\mathbf{x} = a\mathbf{x} + b\mathbf{x} \in \mathcal{V}$,
- $a(b\mathbf{x}) = b(a\mathbf{x}) \in \mathcal{V}$,
- $1\mathbf{x} = \mathbf{x}$.

A vector space that follows these rules is called closed.

1.12.2 Subspace rules

If $\mathcal{W} \in \mathcal{V}$ then $\mathcal{W}$ is a subset of $\mathcal{V}$. If $\mathcal{W}$ is closed under addition and multiplication then it is a subspace of $\mathcal{V}$. An example of a subset not being a subspace is

$$\underbrace{\begin{Bmatrix} a & b \end{Bmatrix}^T}_{\mathcal{W}} \in \underbrace{\mathcal{R}^2, a \geq 0, b \geq 0.}_{\mathcal{V}} \tag{163}$$

in the sense that

$$-k \begin{Bmatrix} a & b \end{Bmatrix} = \begin{Bmatrix} -ka & -kb \end{Bmatrix} \notin \mathcal{V}. \tag{164}$$

That is, the subset is not closed under multiplication.

1.12.3 Span

The span of a space $\mathcal{S}$ is the set of all linear combinations of the vectors in that space. For example

$$\mathcal{S} = \left\{ \begin{Bmatrix} 1 & 0 \end{Bmatrix}^T \quad \begin{Bmatrix} 0 & 1 \end{Bmatrix} \right\} \longrightarrow \text{span}\mathcal{S} = \mathcal{R}^2 \tag{165}$$

in the sense that any vector in $\mathcal{R}^2$ can be written as a linear combination of the two vectors in $\mathcal{S}$. Another example is

$$\mathcal{S} = \left\{ \begin{Bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{Bmatrix} \quad \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{Bmatrix} \right\} \longrightarrow \text{span}\mathcal{S} = \mathcal{R}^4 \text{ of the form } \begin{Bmatrix} a \\ 0 \\ 0 \\ b \end{Bmatrix}. \tag{166}$$

It is a subspace in that it is closed under addition and multiplication.

1.13 Lec 1m Vector spaces, subspaces II

Of course, the idea of spaces, subspaces, and spans extend to matrices. For example

$$\mathcal{S} = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\} \longrightarrow \text{span}\mathcal{S} = \mathcal{R}^{2\times 2} \text{ of the form } \begin{bmatrix} a & b \\ c & d \end{bmatrix}. \tag{167}$$

The statement $\mathcal{S} \in \text{span}\mathcal{S}$ is true in the sense that any element in $\mathcal{S}$ can be written as a linear combination of the span of $\mathcal{S}$. For example, $\begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} = a \begin{Bmatrix} 1 & 0 \\ 0 & 1 \end{Bmatrix} + 0 \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} + \dots$.

1.13.1 Linear independence/dependence

If the only combination of $\mathbf{v}_1, \mathbf{v}_2, \dots$ that results in the zero vector is $0\mathbf{v}_1 + 0\mathbf{v}_2 + \dots$, then $\mathbf{v}_i$ are linearly independent with respect to one another. If there is some nontrivial linear combination that results in $\mathbf{0}$ then the vectors are linearly dependent. For example,

$$2\begin{Bmatrix} 1 \\ 2 \\ 2 \end{Bmatrix} + \begin{Bmatrix} -1 \\ -4 \\ -5 \end{Bmatrix} - \begin{Bmatrix} 1 \\ 0 \\ 1 \end{Bmatrix} = \mathbf{0} \rightarrow \left\{ \begin{Bmatrix} 1 \\ 2 \\ 2 \end{Bmatrix} \quad \begin{Bmatrix} -1 \\ -4 \\ -5 \end{Bmatrix} \quad \begin{Bmatrix} 1 \\ 0 \\ 1 \end{Bmatrix} \right\} \text{ are linearly dependent.} \tag{168}$$

1.13.2 Bases

$\mathcal{B}$ is a basis of $\mathcal{V}$ if $\text{span}\mathcal{B} = \mathcal{V}$ and elements of $\mathcal{B}$ are linearly independent. For instance

$$\left\{ \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}, \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix}, \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix} \right\} \text{ is a basis for } \mathcal{R}^3. \tag{169}$$

Also

$$\left\{ \begin{Bmatrix} 1 \\ 2 \\ 1 \end{Bmatrix}, \begin{Bmatrix} 2 \\ 3 \\ 1 \end{Bmatrix}, \begin{Bmatrix} -1 \\ 2 \\ -3 \end{Bmatrix} \right\} \text{ is a basis for } \mathcal{R}^3. \tag{170}$$

That is to say bases are not unique. Multiple linearly independent sets of vectors span the same space.

1.13.3 Dimension

The dimension of a vector space $\mathcal{V}$ is the minimum number of vectors needed in a basis $\mathcal{B}$of that $\mathcal{V}$. For example, $\dim(\mathcal{R}^3) = 3$ because $\left\{ \begin{Bmatrix} 1 \\ 0 \\ 0 \end{Bmatrix}, \begin{Bmatrix} 0 \\ 1 \\ 0 \end{Bmatrix}, \begin{Bmatrix} 0 \\ 0 \\ 1 \end{Bmatrix} \right\}$ is one of the shortest bases of $\mathcal{R}^3$. No fewer number of vectors can be used to span all of $\mathcal{R}^3$.

If all polynomials of order n and below is $P_n = \left\{1, x, x^2, \dots, x^n\right\}$, then $\dim(P_n) = n+1$. The dimension of a Taylor series is then ∞ because it does not end. Lastly, the dimension of an $m \times n$ matrix is mn.

1.14 Lec 1n Four subspaces of a matrix

1.14.1 Different types of matrices

To review, the different types of matrices are

- Symmetric: $\mathbf{A}=\mathbf{A}^T$,
- Skew: $\mathbf{A}=-\mathbf{A}^T$,
- Orthogonal: $\mathbf{Q}^T = \mathbf{Q}^{-1}$,
- Hermitian: $\bar{\mathbf{Q}}^T = \mathbf{Q}$
- Skew Hermitian: $-\bar{\mathbf{Q}}^T = \mathbf{Q}$
- Unitary: $\bar{\mathbf{Q}}^{-1} = \mathbf{Q}$
- Normal: $\bar{\mathbf{A}}^T\mathbf{A} = \mathbf{A}\bar{\mathbf{A}}^T$

where the overbar indicates a complex conjugate. That means for real matrices, Hermitian is the same as symmetric, and unitary is the same as orthogonal.

1.14.2 Four subspaces

The four subspaces of an $m \times n$ matrix $\mathbf{A}$ are the column space, the null space, the row space, and the left nullspace.

1.14.3 Column space

Notice that $\mathbf{Ax} = \mathbf{b}$ implies

$$\begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ \underbrace{a_{m1}}_{\mathbf{a}_1} & \underbrace{a_{m2}}_{\mathbf{a}_2} & \dots & \underbrace{a_{mn}}_{\mathbf{a}_n} \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{Bmatrix} = \begin{Bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \end{Bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{Bmatrix} = \mathbf{b} \tag{171}$$

$$\rightarrow \mathbf{b} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + \ldots + x_n\mathbf{a}_n. \tag{172}$$

The column space of $\mathbf{A}$, denoted as $\mathcal{C}(\mathbf{A})$, is the set of all versions of $\mathbf{b}$ based on all combinations of x_i. Note that $\mathcal{C}(\mathbf{A}) \in \mathcal{R}^m$. If a known solution vector $\mathbf{b}$ in $\mathbf{Ax} = \mathbf{b}$ satisfies Eq. 172, that is if the given $\mathbf{b} \in \mathcal{C}(\mathbf{A})$, then the matrix equation has at least one solution.

If $\mathbf{A}$ is invertible, then any $\mathbf{b}$ admits a solution because $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ is always a solution. However, if $\mathbf{A}$ is not invertible then $\mathbf{b}$ must be in the column space (must satisfy Eq. 172)in order for there to exist a solution to $\mathbf{Ax} = \mathbf{b}$. For example, given

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{Bmatrix} x \\ y \\ z \end{Bmatrix} = \begin{Bmatrix} 1 \\ 1 \\ 0 \end{Bmatrix}, \tag{173}$$

the fact $\det \mathbf{A} = 0$ reveals $\mathbf{A}$ is not invertible. Therefore $\mathbf{b}$ must be in the column space of $\mathbf{A}$ in order for there to exist a solution. It is, because it satisfies Eq. 172, in that

$$\begin{Bmatrix} 1 \\ 1 \\ 0 \end{Bmatrix} = \begin{Bmatrix} 1 \\ 0 \\ 1 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 1 \\ -1 \end{Bmatrix} + a \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} = x_1\mathbf{a}_1 + x_2\mathbf{a}_2 + x_3\mathbf{a}_3. \tag{174}$$

However, given

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{Bmatrix} x \\ y \\ z \end{Bmatrix} \begin{Bmatrix} 1 \\ 1 \\ 1 \end{Bmatrix}, \tag{175}$$

$\mathbf{b}$ is not in the column space of $\mathbf{A}$ and so there is no solution.

1.14.4 Null/Kernel space

The null space of $\mathbf{A}$, denoted as $N(\mathbf{A})$, is the set of $\mathbf{x}$ such that $\mathbf{A}\mathbf{x} = \mathbf{0}$. The null space is a vector space in that if both $\mathbf{x}_1, \mathbf{x}_2$ fulfill $\mathbf{A}\mathbf{x}_i = \mathbf{0}$, then $\mathbf{A}(\mathbf{x}_1 + \mathbf{x}_2) = \mathbf{A}\mathbf{x}_1 + \mathbf{A}\mathbf{x}_2 = \mathbf{0} + \mathbf{0} = \mathbf{0}$, $c\mathbf{A}\mathbf{x}_1 = c\mathbf{0} = \mathbf{0}$. That is, it is closed under addition and multiplication.

IF $\mathbf{A}$ is invertible $\leftrightarrow \det \mathbf{A} \neq 0$, then the null space will contain only the zero vector. For example

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{Bmatrix} x_1 \\ x_2 \\ x_3 \end{Bmatrix} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \rightarrow \mathbf{x} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} = N(\mathbf{A}). \tag{176}$$

However, if the nullspace contains anything else then $\det \mathbf{A} = 0 \leftrightarrow \mathbf{A}$ is not invertible.

For example let us obtain the null space $N(\mathbf{A})$ of

$$\mathbf{A} = \begin{bmatrix} 1 & -1 & 1 \\ 2 & -1 & 0 \\ 0 & -1 & 2 \end{bmatrix} \longrightarrow^{rref} \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -2 \\ 0 & 0 & 0 \end{bmatrix} \mathbf{x} = \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \tag{177}$$

$$\rightarrow x_1 = x_3,\ x_2 = 2x_3,\ x_3 = x_3 \rightarrow c \begin{Bmatrix} 1 \\ 2 \\ 1 \end{Bmatrix} \in N(\mathbf{A}) \tag{178}$$

$$\rightarrow N(\mathbf{A}) = \left\{ \begin{Bmatrix} 0 \\ 0 \\ 0 \end{Bmatrix} \quad c \begin{Bmatrix} 1 \\ 2 \\ 1 \end{Bmatrix} \right\}. \tag{179}$$

Null space extends not just to square matrices. For example if

$$\mathbf{A} = \begin{bmatrix} 1 & 3 & 2 & 3 \\ 2 & 6 & 8 & 10 \\ 3 & 9 & 10 & 13 \end{bmatrix} \longrightarrow^{rref} \begin{bmatrix} 1 & 3 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{x} = \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \tag{180}$$

$$\rightarrow x_1 = -3x_2 - x_4,\ x_2 = x_2,\ x_3 = -x_4,\ x_4 = x_4. \tag{181}$$

The free variables are x_2, x_4. To solve this we set one variable to 1 and the rest to 0. That is

$$x_2 = 1, x_4 = 0 \rightarrow x_1 = -3x_2, x_2 = x_2, x_3 = 0, x_4 = 0 \rightarrow c \begin{Bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{Bmatrix} \in N(\mathbf{A}), \tag{182}$$

$$x_2 = 0, x_4 = 1 \rightarrow x_1 = -x_4, x_2 = 0, x_3 = -x_4, x_4 = x_4 \rightarrow c \begin{Bmatrix} -1 \\ 0 \\ -1 \\ 1 \end{Bmatrix} \in N(\mathbf{A}). \tag{183}$$

The full null space is

$$N(\mathbf{A}) = \left\{ c \begin{Bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{Bmatrix} \quad c \begin{Bmatrix} -1 \\ 0 \\ -1 \\ 1 \end{Bmatrix} \quad c \begin{Bmatrix} -3 \\ 1 \\ 0 \\ 0 \end{Bmatrix} \right\}, \tag{184}$$

noting that $\mathbf{0}$ is always a member of the null space.

1.14.5 Row space

The row space is the column space of $\mathbf{A}^T$. It is denoted as $\mathcal{C}(\mathbf{A}^T)$ and is the set of all $\mathbf{b}$ such that

$$\mathbf{b} = x_1\mathbf{a}_1^T + x_2\mathbf{a}_2^T + \ldots + x_m\mathbf{a}_m^T. \tag{185}$$

Simply transposing Eq. 171 and doing the same procedure of the column space, one finds the row space.

1.14.6 Left nullspace

The left null space is the set of all $\mathbf{x}$ such that $\mathbf{A}^T\mathbf{x} = \mathbf{0}$. Transposing $\mathbf{A}$ and following the same procedure as the null space Eq. 177/Eq. 180, one obtains the left null space.

1.15 Lec 1o Single value decomposition

Eigenproblems require square $\mathbf{A}$, but single value decomposition of SVD is for all rectangular system matrices. Instead of finding eigenvalues let us find so-called single values σ such that

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u} \leftrightarrow [\mathbf{A}]_{m\times n}\{\mathbf{v}\}_{n\times 1} = \sigma\{\mathbf{u}\}_{m\times 1}. \tag{186}$$

$\mathbf{u}$ is in the column space of $\mathbf{A}$, meaning the equation $\mathbf{u} = \mathbf{a}_1 v_1 + \mathbf{a}_2 v_2 + \ldots$ is satisfied for $\mathbf{v}$, and $\mathbf{v}$ is in the row space of $\mathbf{A}$, meaning the equation $\mathbf{v} = \mathbf{a}_1^T x_1 + \mathbf{a}_2^T x_2 + \ldots$ is satisfied for some other $\mathbf{x}$. Converting Eq. 186 to a matrix problem,

$$\mathbf{A}\mathbf{V} = \mathbf{U}\mathbf{\Sigma} \rightarrow \begin{cases} \mathbf{A}\mathbf{v}_1 = \mathbf{u}_1\sigma_1 \\ \mathbf{A}\mathbf{v}_2 = \mathbf{u}_2\sigma_2 \\ \ldots \end{cases} \tag{187}$$

. This represents a circle with radius $\mathbf{v}_i \in \mathbf{V}$ being transformed by $\mathbf{A}$ into an ellipse with principal directions $\sigma_{ii} \in \mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots \\ 0 & \sigma_2 & \dots \\ \dots & \dots & \dots \end{bmatrix}$.

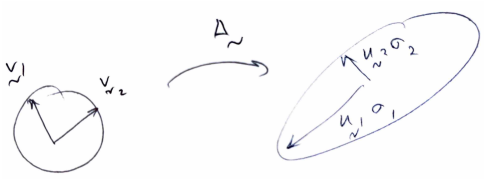

Figure 1: Transformation from hypercircle $\mathbf{v}$ to hyperellipse $\mathbf{u}$ with principal directions σ through $\mathbf{A}$.

$\mathbf{U}$ and $\mathbf{V}$ are like eigenvectors, in that they are unitary/orthonormal with respect to $\mathbf{B}$= $\mathbf{I}$, meaning

$$\mathbf{U}^T\mathbf{I}\mathbf{U} = \mathbf{I} \rightarrow \mathbf{U}^T\mathbf{U} = \mathbf{I} \rightarrow \mathbf{U}^T = \mathbf{U}^{-1}. \tag{188}$$

Therefore,

$$\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \begin{Bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots \end{Bmatrix} \begin{bmatrix} \sigma_1 & 0 & \dots \\ 0 & \sigma_2 & \dots \\ \dots & \dots & \dots \end{bmatrix} \begin{Bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \dots \end{Bmatrix}. \tag{189}$$

If rank $r\mathbf{A} < m$ and $r\mathbf{A} < n$, then the null space contains more than just the zero vector. The null space contains vectors $\mathbf{v}$ for which

$$\mathbf{A}\mathbf{v} = \sigma\mathbf{u} = 0\mathbf{u} = \mathbf{0}. \tag{190}$$

The quantity

$$\mathbf{A}\mathbf{A}^T = (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T)^T\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}^2\mathbf{V}^T \tag{191}$$

is positive semi definite because

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = (\mathbf{A}\mathbf{x})^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{y} = \mathbf{y}\cdot\mathbf{y} \geq 0. \tag{192}$$

As a consequence, its eigenvalues are nonnegative (Eq. 108). As $\mathbf{\Sigma}$ is the matrix of eigenvalues corresponding to $\mathbf{A}\mathbf{A}^T$, therefore $\sigma_j = \sqrt{\lambda_j}$ will be nonnegative and real.

To solve an SVD problem given $\mathbf{A}$ ($\mathbf{A} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$),

- Compute $\mathbf{A}^T\mathbf{A}$,
- Find eigenvectors λ such that $0 = \det(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})$.
- Compute single values $\sigma = \sqrt{\lambda}$
- Populate $\mathbf{\Sigma} = \begin{bmatrix} \sigma_1 & 0 & \dots \\ 0 & \sigma_2 & \dots \\ \dots & \dots & \dots \end{bmatrix}$
- Calculate eigenvectors $\mathbf{v}$ such that $(\mathbf{A}^T\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$.

- Populate $\mathbf{V} = \begin{bmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots \end{bmatrix}$.
- Solve for $\mathbf{U} = \mathbf{A}\boldsymbol{\Sigma}^T\mathbf{V}$.
- Then you have $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T$.

Always order the eigenvalues from largest to smallest. Always normalize the eigenvectors - they must form an orthonormal basis.

2 Mod2 ODEs

2.1 Lec 2a Physical prototypes and classification

2.1.1 Mass spring system

The mass spring system is introduced in the most basic way in Sec. 1.8.1. This model excludes the possibility of a damper and a forcing term. Altogether the displacement equation for a singular cart is

$$m\frac{d^2u}{dt^2} + c\frac{du}{dt} + ku = f(t), \quad u(0) = 0, \frac{du}{dt}(0) = 0, \tag{193}$$

meaning the cart's initial position is zero and its initial velocity is zero. A system of carts is characterized by

$$\mathbf{M}\ddot{\mathbf{u}} + \cancel{\mathbf{C}}\dot{\mathbf{u}} + \mathbf{K}\mathbf{u} = \mathbf{f}(t). \tag{194}$$

2.1.2 Rigid body dynamics

If $\mathbf{p}$ is momentum then force $\mathbf{F} = \dot{\mathbf{p}}$. Likewise if angular momentum is $\mathbf{H} = \mathbf{r} \times \mathbf{p}$ then torque $\mathbf{M} = \dot{\mathbf{H}} = \mathbf{r} \times \dot{\mathbf{p}}$. Angular momentum has units $m \times kg \times m/s = kgm^2/s$. Therefore torque has units kgm^2/s^2. If arclength is s and radius is r then angle $s = \theta r$. Taking the derivative with respect to time, $\mathbf{v} = \boldsymbol{\omega} r$. Angular velocity $\boldsymbol{\omega} = \mathbf{v}/r$ has units $m/sm = 1/s$. Then angular momentum $\mathbf{H} \sim kgm^2/s$ is some product of $\boldsymbol{\omega} \sim 1/s$ and a quantity with units $kgm^2 \sim mr^2 = I$ which we call moment of inertia. Therefore $\mathbf{H} = I\boldsymbol{\omega}$, and either $\mathbf{M} = I\dot{\boldsymbol{\omega}}$ or $\mathbf{M} = I\boldsymbol{\omega}^2$ to satisfy units. To be exact, torque is some combination of those two terms. Particularly,

$$M_x = I_{xx}\dot{\omega}_x + (I_{zz} - I_{yy})\omega_y\omega_z, \tag{195}$$

$$M_y = I_{yy}\dot{\omega}_y + (I_{xx} - I_z)\omega_z\omega_x, \tag{196}$$

$$M_z = I_{zz}\dot{\omega}_z + (I_{yy} - I_{xx})\omega_x\omega_y, \tag{197}$$

with boundary conditions $\omega_x(0) = 0, \omega_y(0) = 0, \omega_z(0) = 0$.

2.1.3 Boundary value problems

Newton's third law requires that for a beam subjected to an external distributed force $p(x)$ on the top surface,

$$0 = \sum F_z = -V + (V + dV) + p(x)dx \rightarrow p(x) = \frac{dV}{dx}. \tag{198}$$

Torque/bending moment in relationship to force is

$$M = V\hat{z} \times x\hat{x} = Vx\hat{n} \rightarrow |V| = |\frac{dM}{dx}| \longrightarrow p(x) = \frac{d^2M}{d^2x}. \tag{199}$$

Moment for an elastic beam can also be expressed as

$$M = EI\kappa = EI\frac{d^2w}{dx^2} \longrightarrow p(x) = \frac{d^2}{d^2x}(EI\frac{d^2w}{dx^2}). \tag{200}$$

If the bar is fixed at both ends then $w(0) = w(L) = 0, M(L) = 0$.

2.1.4 Classification and terminology

- Independent variables are ones with respect to which derivatives are taken. x in du/dx
- Dependent variables are unknowns of the problem. u in du/dx
- Order is highest order derivative in the equation.
- Homogeneity of an equation is when all dependent variables being set to zero satisfies the equation.
- Lienarity of an equation is when the dependent variables are linear. d^2u/dx^2 is permissible but u^2 is not.
- Initial value problem is one in which the independent variable is time and the initial conditions of the dependent variables are given at the initial time step.
- Boundary value problem is one in which is the independent variable is spatial and the boundary conditions of the dependent variables are given at certain points in space.

2.2 Lec 2b Linear ODEs and power series

2.2.1 Homogeneous linear ODE with constant coefficients

Homogeneity implies there is no constant term. Then a linear homogeneous ODE with constant coefficients is some

$$\frac{d^ny}{dx^n} + c_1\frac{d^{n-1}y}{dx^{n-1}} + \ldots + c_{n-1}\frac{dy}{dx} + c_ny = 0. \tag{201}$$

Letting $y = Ce^{\lambda x}$, we get

$$0 = \frac{d^n}{dx^n}(Ce^{\lambda x}) + c_1\frac{d^{n-1}}{dx^{n-1}}(Ce^{\lambda x}) + \ldots + c_{n-1}\frac{d}{dx}(Ce^{\lambda x}) + c_n(Ce^{\lambda x}) \tag{202}$$

$$\rightarrow 0 = C\lambda^n e^{\lambda x} + c_1 C\lambda^{n-1}e^{\lambda x} + \ldots + c_{n-1}C\lambda e^{\lambda x} + c_n Ce^{\lambda x} \tag{203}$$

$$\rightarrow 0 = \lambda^n + c_1\lambda^{n-1} + \ldots + c_{n-1}\lambda + c_n, \tag{204}$$

dividing across by $Ce^{\lambda x}$.

2.2.2 Cauchy Euler equation

The Cauchy Euler equation is some

$$x^n\frac{d^n y}{dx^n} + c_1 x^{n-1}\frac{d^{n-1}y}{dx^{n-1}} + \ldots + c_{n-1}x\frac{dy}{dx} + c_n y = 0. \tag{205}$$

Again if $y = Ce^{\lambda x}$ then

$$0 = x^n\frac{d^n}{dx^n}(Ce^{\lambda x}) + c_1 x^{n-1}\frac{d^{n-1}}{dx^{n-1}}(Ce^{\lambda x}) + \ldots + c_{n-1}x\frac{d}{dx}(Ce^{\lambda x}) + c_n(Ce^{\lambda x}) \tag{206}$$

$$\rightarrow 0 = \lambda^n x^n + c_1\lambda^{n-1}x^{n-1} + \ldots + c_{n-1}\lambda x + c_n\lambda x. \tag{207}$$

2.2.3 Power series

A power series centered around x_0 is an infinite sum of the form

$$\sum_{n=0}^{\infty} a_n(x - x_0)^n = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \ldots . \tag{208}$$

If there is a limit to the infinite series then it is said to converge. If the absolute value of the series converges then it is said to converge absolutely. There exists a so-called radius of convergence R such that the power series converges absolutely for $|x - x_0| < R$ and diverges for $|x - x_0| > R$. So if the chosen point x_0 is more than R away from x then the series diverges. $R = 0$ indicates that the series only converges at $x = x_0$. $R = \infty$ indicates that the series converges $\forall x$.

2.2.4 Ratio test

The the ratio between adjacent terms in a power series converges to

$$\lim_{n\to\infty}\left|\frac{a_{n+1}(x - x_0)^{n+1}}{a_n(x - x_0)}\right| = (x - x_0)\lim_{n\to\infty}\frac{a_{n+1}}{a_n} = \alpha, \tag{209}$$

and this informs the convergence behavior of the series, where $\begin{cases} \alpha > 1, & \text{divergence,} \\ \alpha < 1, & \text{convergence,} \\ \alpha = 1, & \text{inconclusive.} \end{cases}$

The radius of convergence

$$R = \left(\lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right|\right)^{-1} \tag{210}$$

An example is to find the R of

$$\sum_{n=1}^{\infty} \frac{1}{2^n n}(x+1)^n. \tag{211}$$

We identify $x_0 = -1$, $a_n = 1/2^n n$ and compute

$$R = \left(\lim_{n\to\infty} \frac{2^n n}{2^{n+1}(n+1)} \right)^{-1} = \left(1/2 \right)^{-1} = 2 \longrightarrow -3 < x < 1 \leftrightarrow \text{ convergence.} \tag{212}$$

If two power series $f(x) = \sum_{n=0}^{\infty} a_n(x-x_0)^n$, $g(x) = \sum_{n=0}^{\infty} b_n(x-x_0)^n$ converge within the interval $I = x_0 - R < x < x_0 + R$ then

- $f \pm g = \sum_{n=0}^{\infty}(a_n \pm b_n)(x-x_0)^n$ convergees in I,
- $df/dx = \sum_{n=0}^{\infty} n a_n (x-x_0)^{n-1}$ converges in I,
- $\int_a^b f dx = \frac{1}{n+1} a_n[(b-x_0)^{n+1} - (a-x_0)^{n+1}]$ converges in I,
- $fg = \sum_{n=\infty}(a_0 b_n + a_1 b_{n-1} + \ldots + a_n b_0)(x-x_0)^n$ converges in I.

Note that if $f = g$ then all $a_n = b_n$. Consequently, if $f = 0$ then all $a_n = 0$.

2.2.5 Taylor series

A taylor series expanded around x_0 is

$$TSf|_{x_0} = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n, \tag{213}$$

i.e. a power series in which $a_n = f^{(n)}(x_0)/n!$. We say that $TSf|_{x=0} = f(x)$ if the conditions

- $f(x_0)$ is infinitely differentiable,
- $\exists R > 0$

are met, then

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n, \tag{214}$$

and $f(x)$ is said to be analytic at x_0 in that it can be solved for in exact terms as a function of an infinite series. The opposite of analytic is singular. If $x_0 = 0$, then the Taylor series is a simple Maclaurin series.

Trig functions, e^x, and polynomials are analytic everywhere. However, consider $f(x) = (x+1)^{3/2} \to f''(x) = (3/4)(x+1)^{-1/2}$. This is not analytic at $x = -1$ because that value does not exist. Therefore, $x = -1$ is a singular point of $f(x)$ in the sense that $f(x)$ is singular at $x = -1$. Also, consider x^α, which is analytic everywhere except at $x = 0$, for which the value does not exist.

2.3 Lec 2c Linear ODEs analytic coefficients

A linear homogeneous second order ODE of the form

$$c_0(x)y'' + c_1(x)y' + c_2(x)y = 0 \tag{215}$$

has the normal form

$$y'' + \frac{c_1}{c_0}y' + \frac{c_2}{c_0}y = py' + qy = 0. \tag{216}$$

The classification of x_0 depends on

$$\begin{cases} c_0(x_0) \neq 0, & x_0 \text{ is ordinary,} \\ c_0(x_0) = 0, & x_0 \text{ is singular } \ (p, q \to \infty). \end{cases} \tag{217}$$

If x_0 is a singular point then it can either be

$$\begin{cases} \text{regular singular,} & (x - x_0)p \ \text{ and } \ (x - x_0)^2 q \text{ are analytic at } x_0, \\ \text{irregular singular,} & \text{otherwise.} \end{cases} \tag{218}$$

Once the singular points are determined, it is determined that R is at least as great as the distance between x_0 and the singular point that is nearest to R. That is, the smallest interval of convergence is $I = (x_0 - R, x_0 + R)$. The way in which to find a solution to Eq. 215 is the following procedure.

- Assume solution $y = \sum_{n-0}^{\infty} a_n(x - x_0)^n$,
- take necessary derivatives and substitute them back in,
- expand the infinite sum and group terms according to powers of $(x - x_0)$,
- solve for a_n in terms of a_0, a_1 by virtue of every set of terms grouped by a particular x^m needing to go to zero: $\circ = 0$ in $(\circ)x^0 + (\circ)x^1 + \ldots = 0$.
- Let the solution be $y = a_0 y_1 + a_1 y_2$, where y_1, y_2 are the sets of terms within the expanded y that contain coefficients a_0, a_1 respectively.

2.3.1 Solution near an ordinary point

For example, suppose we wish to find the solution to

$$(1 - x^2)y'' - 2xy' + \lambda y = 0. \tag{219}$$

about the origin $x_0 = 0$. We note $c_0 = 1 - x^2,\ c_1 = -2x,\ c_2 = \lambda$, all of which are analytic everywhere. x_0 is ordinary because $c_0(0) = 1 \neq 0$. We note also that $p = 2x/(1 - x^2),\ q = \lambda/(1 - x^2)$. Therefore we can assume the solution

$$y = \sum_{n=0}^{\infty} a_n(x - x_0)^n = \sum_{n=0}^{\infty} a_n x^n \tag{220}$$

$$\to y' = \sum_{n=1}^{\infty} n a_n x^{n-1}, \quad y'' = \sum_{n=2}^{\infty} n(n-1) a_n x^{n-2}. \tag{221}$$

Notice that the indices change because of the fact that the derivative of the term $a_0 x^0$ is just zero. Therefore, it is excluded. Substituting this in,

$$0 = \underbrace{(1-x^2)\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2}} - \overbrace{2x\sum_{n=1}^{\infty} n a_n x^{n-1}} + \lambda \sum_{n=0}^{\infty} a_n x^n \tag{222}$$

$$= \underbrace{\sum_{n=2}^{\infty} n(n-1)a_n x^{n-2} - \sum_{n=2}^{\infty} n(n-1)a_n x^n} - \overbrace{\sum_{n=1}^{\infty} 2n a_n x^n} + \lambda \sum_{n=0}^{\infty} a_n x^n. \tag{223}$$

Reindexing each term to get x^n everywhere,

$$0 = \sum_{(n+2)=2}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=2}^{\infty} n(n-1)a_n x^n - \sum_{n=1}^{\infty} 2n a_n x^n + \lambda \sum_{n=0}^{\infty} a_n x^n \tag{224}$$

$$= \sum_{n=0}^{\infty} (n+2)(n+1)a_{n+2}x^n - \sum_{n=2}^{\infty} n(n-1)a_n x^n - \sum_{n=1}^{\infty} 2n a_n x^n + \lambda \sum_{n=0}^{\infty} a_n x^n. \tag{225}$$

Expanding the sum, noting well where each index starts,

$$0 = x^0[(2)(1)a_2 + \lambda a_0] + x^1[(3)(2)a_3 - 2a_1 + \lambda a_1] + \ldots + x^s[(s+2)(s+1)a_{s+2} - s(s-1)a_s - 2s a_s + \lambda a_s] \tag{226}$$

for $s \geq 2$. This implies

$$2a_2 + \lambda a_0 = 0 \longrightarrow a_2 = -\frac{\lambda a_0}{2}, \tag{227}$$

$$6a_3 + (\lambda - 2)a_1 = 0 \longrightarrow a_3 = \frac{2-\lambda}{6} a_1, \tag{228}$$

$$(s+2)(s+1)a_{s+2} + [-s(s-1) - 2s + \lambda]a_s = 0 \longrightarrow a_{s+2} = \frac{s(s-1) + 2s - \lambda}{(s+2)(s+1)} a_s = \frac{s(s+1) - \lambda}{(s+2)(s+1)} a_s. \tag{229}$$

Note that every even a_s will contain a_0 because it is contained in a_2, and every odd a_s will contain a_1 beacuse it is contained in a_3. Then,

$$y = \sum_{n=0}^{\infty} a_n x^n \tag{230}$$

$$= a_0 + a_1 x - \frac{\lambda a_0}{2} x^2 + \frac{(2-\lambda)a_1}{6} x^3 + \ldots \tag{231}$$

$$= a_0 (1 - \frac{\lambda}{2} x + \ldots) + a_1 (x + \frac{2-\lambda}{6} x^3 + \ldots) = a_0 y_1 + a_1 y_2 = y(x). \tag{232}$$

Note that the two parts of the solution y_1, y_2 are linearly independent in that y_1/y_2 is not a constant.

If $\lambda = n(n+1)$, where n is NOT the same as the indexe quantity, then the coefficients become

$$a_2 = -\frac{n(n+1)}{2}a_0, \tag{233}$$

$$a_3 = \frac{2-n(n+1)}{6}a_1, \tag{234}$$

$$a_{s+2} = \frac{s(s+1)-n(n+1)}{(s+2)(s+1)}a_s = \frac{s^2+s-n^2-n}{s^2+2s+2}a_s = \frac{s^2+sn+s-ns-n^2-n}{(s+2)(s+1)}a_s$$
$$= \frac{(s-n)(s+n+1)}{(s+2)(s+1)}a_s = a_{s+2}. \tag{235}$$

Lastly the singular points of the ODE occur where $0 = c_0(x) = 1-x^2 \to x = \pm 1$. $R = 1$ is then the distance from the origin to this singular point. Therefore the interval $I = (-1, 1)$.

2.3.2 Legendre polynomials

Legendre polynomials

$$P_n(x) = \sum_{m=0}^{M}(-1)^m \frac{(2n-2m)!}{2^n m!(n-m)!(n-2m)!}x^{n-2m}, \quad M = \begin{cases} n/2, & n \text{ even}, \\ (n-1)/2, & n \text{ odd}. \end{cases} \tag{236}$$

They are orthogonal over $-1 < x < 1$ so that

$$\int_{-1}^{1} P_m(x)P_n(x)dx = 0, \quad m \neq n, \tag{237}$$

and these are solutions to Sturm Liouville BVPs. Some of the first of these are

$$\begin{cases} P_0(x) = 1, \\ P_1(x) = x, \\ P_2(x) = \frac{1}{2}(3x^2-1), \\ P_3(x) = \frac{1}{2}(5x^3-3x), \\ P_4(x) = \frac{1}{8}(35x^4-30x^2+3), \\ P_5(x) = \frac{1}{8}(63x^5-70x^3+15x). \end{cases} \tag{238}$$

Even/odd Legendre polynomials have the quality

$$P_{2m}(-x) = P_{2m}(x), \quad P_{2m-1}(-x) = -P_{2m-1}(x) \tag{239}$$

respectively. That is, they are literally even and odd functions.

2.4 Lec 2d Linear ODEs regular singular points

Let $x_0 = 0$. That is, shift the coordinate system as necessary so that $x_0 = 0$ if it is not already. Recall from 217,218 that x being a regular singular point requires $c_0(x) = 0$, and that xp, x^2q are analytic. Therefore both xp and x^2q have power series expansions that converge within some interval I.

2.4.1 Frobenius method

Just like in Sec. 2.3, a linear homogeneous second order ODE of the form

$$c_0(x)y'' + c_1(x)y' + c_2(x)y = 0 \tag{240}$$

has the normal form

$$y'' + \frac{c_1}{c_0}y' + \frac{c_2}{c_0}y = y'' + py' + qy = 0 \tag{241}$$

which implies

$$x^2y'' + x(xp)y' + (x^2q) = 0. \tag{242}$$

The method of Frobenius, used to solve equations in the form of Eq. 242, is to assume a solution

$$y = \sum_{n=0}^{\infty} a_n x^{n+r} \tag{243}$$

with derivatives

$$y' = \sum_{n=0}^{\infty}(n+r)a_n x^{n+r-1}, \qquad y'' = \sum_{n=0}^{\infty}(n+r)(n+r-1)a_n x^{n+r-2}. \tag{244}$$

Unlike in Sec. 2.3, the indices here do not change because there is no term that goes to zero necessarily, unlike a_0x^0.

The next step is to substitute the derivatives back into the original equation, and combine like powers of x. Notice that the coefficient x^2 on each of the terms eliminates the possibility of terms such as x^{r-1}; every exponent will be $\geq r$.

Once like powers of x are combined, such as in Eq. 226, the coefficient associated with x^r being set to zero can be used to determine the value of r. The final solution to the equation will always be some

$$y(x) = ay_1(x) + by_2(x), \tag{245}$$

but the nature of y_1, y_2 change based on the solutions of r. Particularly,

$$\begin{cases} y_1 = |x|^{r_1}(a_0 + a_1x + \dots), & \text{Frobenius Case I,} \\ y_2 = |x|^{r_2}(b_0 + b_1x + \dots), & r_1, r_2 \text{ are distinct and do not differ by an integer,} \\ \\ y_1 = |x|^{r_1}(a_0 + a_1x + \dots), & \text{Frobenius Case II,} \\ y_2 = y_1\ln|x| + |x|^{r_1}(b_1x + b_2x^2 + \dots), & r_1 = r_2, \\ \\ y_1 = |x|^{r_1}(a_0 + a_1x + \dots), & \text{Frobenius Case III,} \\ y_2 = \kappa y_1\ln|x| + |x|^{r_2}(b_0 + b_1x + \dots), & r_1, r_2 \text{ are distinct and differ by an integer.} \end{cases} \tag{246}$$

2.4.2 Solution near a regular singular point

For example suppose we want to find a general solution of the Bessel equation of order ν

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0 \tag{247}$$

about the origin $x_0 = 0$. The normal form is

$$y'' + \frac{1}{x}y' + (1 - \frac{\nu^2}{x^2})y = 0. \tag{248}$$

the term $c_0(x = 0) = 0^2 = 0$ implies that $x = 0$ is a singular point, and the terms $xp = 1,\ x^2q = x^2 - \nu^2$ being analytic (as they are polynomials) imply that $x = 0$ is a regular singular point. Therefore the Frobenius method is used, in which the solution

$$y(x) = \sum_{n=0}^{\infty} a_n x^{n+r} \tag{249}$$

is assumed, which has derivatives

$$y' = \sum_{n=0}^{\infty}(n + r)a_n x^{n+r-1}, \qquad y'' = \sum_{n=0}^{\infty}(n + r)(n + r - 1)a_n x^{n+r-2}. \tag{250}$$

Substituting them into Eq. 247,

$$0 = \sum_{n=0}^{\infty}\Big[(n + r)(n + r - 1)a_n x^{n+r} + (n + r)a_n x^{n+r} + \underbrace{a_n x^{n+r+2}} - \nu^2 a_n x^{n+r}\Big] \tag{251}$$

$$= \sum_{n=0}^{\infty}\Big[(n + r)(n + r - 1)a_n x^{n+r} + (n + r)a_n x^{n+r} - \nu^2 a_n x^{n+r}\Big] + \underbrace{\sum_{n=2}^{\infty} a_{n-2} x^{n+r}} \tag{252}$$

Expanding,

$$\begin{aligned} 0 = &\Big(r(r - 1)a_0 x^r + r a_0 x^r - \nu^2 a_0 x^r\Big) \\ &+\Big((r + 1)r a_1 x^{r+1} + (r + 1)a_1 x^{r+1} - \nu^2 a_1 x^{r+1}\Big) \\ &+\Big((r + 2)(r + 1)a_2 x^{r+2} + (r + 2)a_2 x^{r+2} - \nu^2 a_2 x^{r+2} + a_0 x^{r+2}\Big) + \dots \\ &+\Big((r + s)(r + s - 1)a_s x^{r+s} + (r + s)a_s x^{r+s} - \nu^2 a_s x^{r+s} + a_{s-2} x^{r+s}\Big) + \dots . \end{aligned} \tag{253}$$

Solving for r in the first term,

$$0 = (r(r - 1) + r - \nu^2)x^r a_0 \to r = \pm\nu. \tag{254}$$

Now that r is known it can be substituted into other terms. Substituting it in to the coefficient for x^{r+1},

$$0 = ((\pm\nu + 1)(\pm\nu) + \pm\nu + 1 - \nu^2)x^{r+1}a_1 \rightarrow \begin{cases} r_1 = \nu : (2\nu + 1)a_1 = 0 \rightarrow a_1 = 0, \\ r_2 = -\nu : (-2\nu + 1)a_1 = 0 \rightarrow a_1 = 0. \end{cases} \tag{255}$$

Both equations indicate $a_1 = 0$. Further, notice that in the the coefficient of x^{r+s}, a_s depends on a_{s-2}. Particularly

$$(2\nu s + s^2)a_s = -a_{s-2} \rightarrow a_3 = a_5 = \ldots = 0 \tag{256}$$

because $a_1 = 0$.

Because the odd terms cancel, the solution becomes some

$$y_1 = \sum_{n=0}^{\infty} a_n x^{n+\nu} \tag{257}$$

where all the a_n depend on a_0. If we define it as

$$a_0 = \frac{a}{2^\nu \Gamma(\nu + 1)} \tag{258}$$

where $\Gamma(n + 1) = n!$ ($\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$), then the solution takes the form of

$$y_1(x) = a\mathcal{J}_\nu(x) \tag{259}$$

where the Bessel function of the first kind is, for $r_1 = \nu$

$$\mathcal{J}_\nu(x) = x^\nu \sum_{n=0}^{\infty} \frac{(-1)^n x^{2n}}{2^{2n+\nu} n! \Gamma(n + \nu + 1)}. \tag{260}$$

If $r_2 = -\nu$, the second linearly independent solution

$$y_2(x) = b\mathcal{J}_{-\nu}(x) \tag{261}$$

emerges. Then the general solution is

$$y(x) = a\mathcal{J}_\nu + b\mathcal{J}_{-\nu}. \tag{262}$$

Note: the value of ν determines whether or not r_1, r_2 differ by an integer and thus whether or not the problem is of Frobenius Case I or Frobenius Case III. For example, $\nu = 1/2$ indicates $\Delta r_{12} = 1 \in \mathcal{Z}$, but $\nu = 1/4$ indicates $\Delta r_{12} = 1/2 \notin \mathcal{Z}$.

If $\nu = n$ with $n \in \mathcal{Z}$, then we must get Frobenius Case III and

$$y(x) = a\mathcal{J}_n + b\mathcal{Y}_n, \tag{263}$$

where $\mathcal{Y}_n$ is the Bessel function of the second kind. Approximations for the Bessel functions are

$$\mathcal{J}_n \approx \frac{1}{\Gamma(n + 1)} \left(\frac{x}{n}\right)^n, \quad \mathcal{Y}_n(x) \approx \begin{cases} \frac{2}{\pi}\left[\ln \frac{x}{x} + \gamma\right], & n = 0, \\ -\frac{\Gamma(n)}{\pi}\left(\frac{2}{x}\right)^n, & n = 1, 2, \ldots \end{cases} \tag{264}$$

where $n = 1, 2, 3, \ldots$.

2.5 Lec 2e Bessel functions

As in Eq. 247, the Bessel equation is

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0. \tag{265}$$

On a graph with axes $(x, y) = (x, \mathcal{J}(x))$, Bessel function of the first kind of order zero, $\mathcal{J}_\prime$, starts at 1 and oscillates around 0 as it slowly diminishes. Whereas, $\mathcal{J}_1, \mathcal{J}_\in$ start at zero and do the same. Bessel function of the second kind $\mathcal{Y}_n(x)$ is singular as $x \to 0$, meaning it tends towards $-\infty$ as $x \to 0$ because there is no solution at that point. As x increases, $\mathcal{Y}_n$ approaches 0 and then oscillates around it until it diminishes.

Some bessel function properties are

- Bessel functions of successive orders are given by $\mathcal{J}_{\nu-1}+\mathcal{J}_{\nu+1} = \frac{2\nu}{x}\mathcal{J}_\nu$, $\mathcal{Y}_{\nu-1}+\mathcal{Y}_{\nu+1} = \frac{2\nu}{x}\mathcal{Y}_\nu$. Particularly, $\mathcal{Y}_2 = \frac{2}{x}\mathcal{Y}_1 - \mathcal{Y}_0$ and the same for $\mathcal{J}$.
- Derivatives are $d\mathcal{J}_0/dx = -\mathcal{J}_1$, $d\mathcal{J}_1/dx = \mathcal{J}_0 - \mathcal{J}_1/x$, and the same for $\mathcal{Y}$.
- Bessel function of the second kind $\mathcal{Y}_\nu(x)$ tends towards $\ln x$ if $\nu = 0$ and towards $x^{-\nu}$ if $\nu \neq 0$.

The Hankel functions are

$$H_\nu^{(1)}(x) = \mathcal{J}_\nu(x) + i\mathcal{Y}_\nu(x), \qquad H_\nu^{(2)}(x) = \mathcal{J}_\nu(x) - i\mathcal{Y}_\nu(x), \tag{266}$$

so that a general solution to the Bessel equation of order ν is

$$y(x) = aH_\nu^{(1)} + bH_\nu^{(2)} \tag{267}$$

which is analogous to the harmonic functions in that $0 = y'' + y$ has general solution $y = ae^{ix} + be^{-ix}$ where $e^{\pm x} = \cos x \pm i \sin x$.

The Bessel Equation of order ν is, as seen in Eq. 247,

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0. \tag{268}$$

The modified Bessel equation is

$$x^2y'' + xy' - (x^2 + \nu^2)y = 0 \tag{269}$$

with general solution

$$y(x) = \hat{a}\mathcal{J}_\nu(ix) + \hat{b}\mathcal{Y}_\nu(ix) = aI_\nu + bK_\nu, \tag{270}$$

where I_ν, K_ν are modified Bessel functions of the first and second kind. Properties of the modified Bessel functions are

- $I_{\nu-1} - I_{\nu+1} = \frac{2\nu}{x}I_\nu, -K_{\nu+1} + K_{\nu+1} = \frac{2\nu}{x}K_\nu$,
- Derivatives are $dI_0/dx = I_1$, $dK_0/dx = -K_1$, $dI_1/dx = I_0 - I_1/x, dK_1/dx = -K_0 - K_1/x$.
- Modified Bessel function of the second kind $K_\nu(x)$ tends towards $\ln x$ if $\nu = 0$ and towards $x^{-\nu}$ if $\nu \neq 0$.

The modified Bessel functions are not oscillatory like the original Bessel functions. The analogy is the governing equation $y'' - y = 0$ for which the general solution is $y = a\cosh x + b\sinh x$ and this does not oscillate.

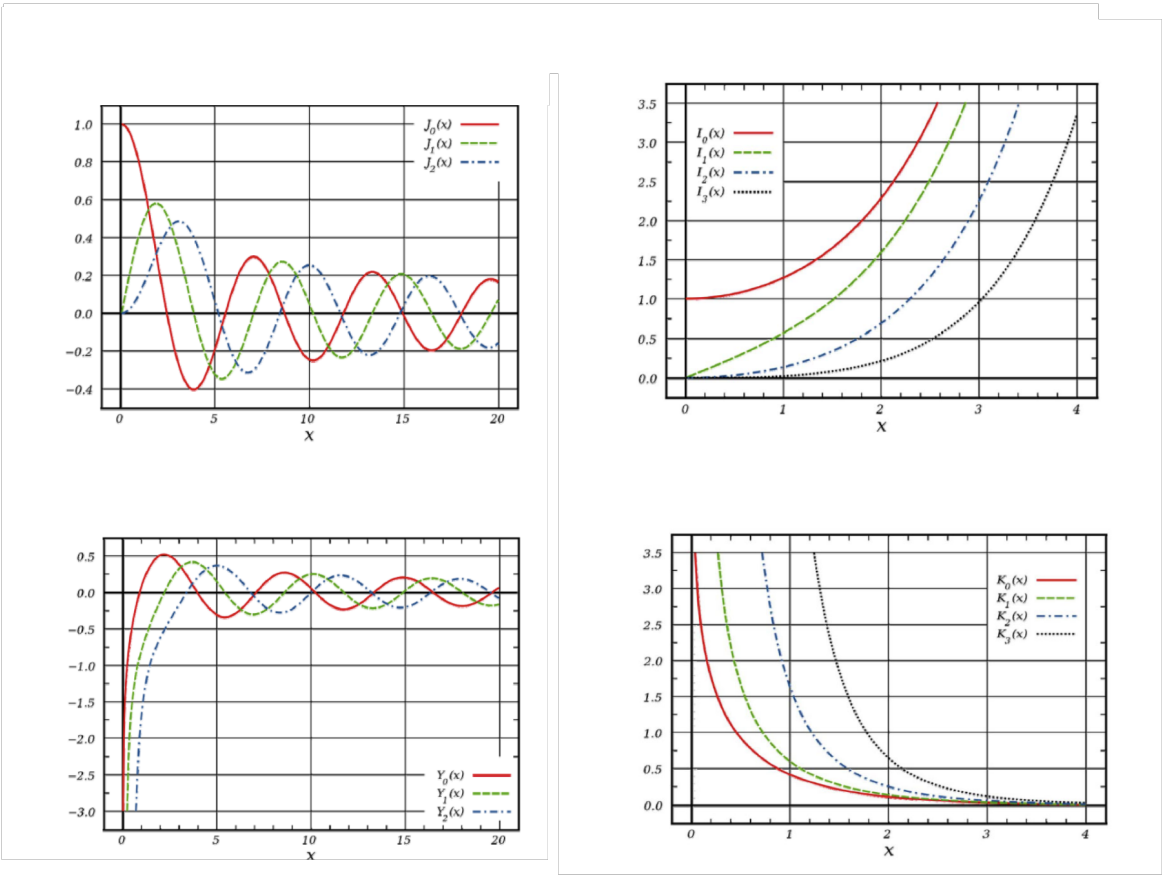

Figure 2: Bessel/modified bessel functions of the first, second kind of orders $0, 1, 2/0, 1, 2, 3$.

2.6 Lec 2f Sturm Liouville eigenproblem

If $p > 0, p', q, w > 0$ are real and continuous on $[x_1, x_2]$, then λ in the ODE

$$\frac{d}{dx}[p(x)y'] + [q(x) + \lambda w(x)]y = 0, \quad x_1 < x < x_2 \tag{271}$$

$$\leftrightarrow p'y' + py'' + qy + \lambda wy = 0, \quad x_1 < x < x_2 \tag{272}$$

with boundary conditions

$$a_1 y(x_1) + b_1 y'(x_1) = 0, \; a_2 y(x_2) + b_2 y'(x_2) = 0 \tag{273}$$

represents an eigenvalue with corresponding eigenfunction y_n. So the eigenpair is λ_n, y_n. There are infinite real eigenvalues. Two eigenfunctions y_m, y_n are orthogonal with respect to weighting function w so that $\int_{x_1}^{x_2} y_m(x) y_n(x) w(x) dx = 0, \; m \neq n$.

The Sturm Liouville bonudary value eigenproblem is analogous to the matrix eigenproblem $\mathbf{Ax} = \lambda \mathbf{Bx}$.

2.6.1 Bessel equation

An example of a Sturm Liouville eigenproblems is the Bessel equation

$$x^2y'' + xy' + (\bar{\lambda}^2x^2 - \nu^2)y = 0 \tag{274}$$

in that

$$y' + xy'' + (\frac{-\nu^2}{x} + \bar{\lambda}^2x)y = 0. \tag{275}$$

For this to match up with Eq. 271 ($p'y' + py'' + qy + \lambda wy = 0, \quad x_1 < x < x_2$), it must be that $p = w = x,\ q = -\nu^2/x,\ \lambda = \bar{\lambda}^2$.

2.6.2 Legendre equation

The Legendre equation is

$$(1 - x^2)y'' - 2xy' + n(n+1)y = 0 \tag{276}$$

which implies

$$[(1 - x^2)y]' + (0 + n(n+1))y = 0 \rightarrow p = 1 - x^2, w = 1, \lambda = n(n+1). \tag{277}$$

2.7 Lec 2g IVP numerical solutions

Consider the first order ODE

$$\frac{dy}{dx} = f(x, y). \tag{278}$$

First of all there is the numerical integration method where one considers

$$dy = f(x, y)dx \rightarrow \int_{y_i}^{y_{i+1}} dy = \int_{x_i}^{x_{i+1}} f(x, y)dx \rightarrow y_{i+1} - y_i = \int_{x_i}^{x_{i+1}} f(x, y)dx. \tag{279}$$

Then there are finite difference methods. Consider again

$$\frac{dy}{dx} = f(x, y). \tag{280}$$

Then Euler's method

$$dy/dx \approx \Delta y/\Delta x \approx y_{i+i} - y_i/x_{i+1} - x_i = f(x_i, y_i). \tag{281}$$

Let

$$x_{i+1} - x_i = h. \tag{282}$$

Then

$$y_{i+1} = y_i + hf(x_i, y_i). \tag{283}$$

h is called the step size. This is called Euler's method. It is known as a forward method because you are iterating forward. It is called explicit because the RHS only involves x_i and y_i where LHS involves y_{i+1}. So you are finding the unknown $i + 1$ term using the

known i terms.
An alternative approach is the backward method. Let once again

$$dy/dx = f(x, y) \tag{284}$$

but

$$\nabla y/\nabla x \approx f(x, y). \tag{285}$$

Then

$$\frac{y_i - y_{i-1}}{x_i - x_{i-1}} = \frac{y_i - y_{i-1}}{h} \approx f(x_i, y_i). \tag{286}$$

So

$$y_i = y_{i-1} + hf(x_i, y_i). \tag{287}$$

This is a backward method because you are iterating backwards. It is called implicit because both the LHS and the RHS involve some i term. So there are unknowns on both sides, theoretically. Some assumption is required. Let us again consider Euler's forward method. If

$$dy/dx = f(x, y) \tag{288}$$

so that

$$y_{i+1} = y_i + hf(x_i, y_i), \tag{289}$$

we might ask, what is the error involved in this approach? We do not yet know exactly but we can approximate the order of magnitude of the error in the Taylor series expansion of $y(x)$ at x_i. That is,

$$y_{i+1} = y_i + \frac{h^1}{1!}hy_i' + \frac{h^2}{2!}y_i'' + \frac{h^3}{3!}y_i''' + \dots . \tag{290}$$

Recall that $\frac{dy}{dx} = y' = f(x_i, y_i)$. Then in approximating the infinite series into something containing only known values we might say

$$y_{i+1} = y_i + hf(x_i, y_i) + \frac{h^2}{2!}y_i'' + \ldots = y_{i+1} = y_i + hf(x_i, y_i) + \mathcal{O}(h^2) \tag{291}$$

where $\mathcal{O}$ is called the local truncation error and its order (h^2) is just the order of magnitude of the approximation error. That is error at the local level, or at this particular step in the iteration. However if this approximation is used over many steps then then error will accumulate. In this way it is then known that if local truncation error is $\mathcal{O}(h^2)$ then total truncation error is $\mathcal{O}(h)$. That applies to this particular example. Another example is: if local error is $\mathcal{O}(h^3)$ then total error is $\mathcal{O}(h^2)$. This generally applies.
This method has low accuracy. There is also potential for instability numerically. That is, the solve may not converge. How do we improve this accuracy? One way is to obtain steps at the beginning and end of the interval and then average. That is, let us improve

$$y_{i+1} = y_i + hf(x_i, y_i) \tag{292}$$

by saying instead that

$$y_{i+1} = y_i + \frac{h}{2}\bigg(f(x_i, y_i) + f(x_{i+1}, y_{i+1})\bigg). \tag{293}$$

But, note that terms on the RHS are unknown. So let us use Euler method to estimate the y_{i+1} term on that side. Let us create a value which serves as an approximation

$$y_{i+1}^e = y_i + hf(x_i, y_i). \tag{294}$$

Plugging Eq. 294 into the RHS of Eq. 293,

$$y_{i+1} = y_i + \frac{h}{2}\bigg(f(x_i, y_i) + f(x_{i+1}, y_{i+1}^e)\bigg) = y_i + \frac{h}{2}\bigg(f(x_i, y_i) + f(x_{i+1}, y_i + hf(x_i, y_i))\bigg). \tag{295}$$

Now let us create a general model using as a characteristic example Eq. 295. First of all let us recall that

$$x_i + h = x_{i+1} \tag{296}$$

by virtue of Eq. 282. Now let

$$y_{i+1} = y_i + h\bigg(af(x_i, y_i) + bf(x_i + \alpha h, y_i + \beta hf(x_i, y_i))\bigg) \tag{297}$$

where α, β, a, b are constants. Eq. 297 is a model of 295 if

$$\alpha = 1,\ \beta = 1,\ a = 1/2,\ b = 1/2. \tag{298}$$

2.8 Lec 2h Higher order methods

Reminiscent of Lec 2.7, consider the first order ODE

$$f(x, y) = \frac{dy}{dx} \tag{299}$$

and the forward Euler method

$$y_{i+1} = y_i + hf(x_i, y_i). \tag{300}$$

This essentially uses the slopes at x_i and at y_i (slopes because $\frac{dy}{dx}$ informs slope) to estimate y_{i+1}. Recall otherwise that for this example, truncation error $\mathcal{O} = \mathcal{O}(h^2)$ at the local (per step) level and that $\mathcal{O} = \mathcal{O}(h)$ at the total level.
Let us try to improve beyond the Euler method. Recall Eq. 20 with parameters α, β, a, b. It is generally true for this model that

$$\text{if } a + b = 1,\ \alpha b = 1/2,\ \beta b = 1/2, \tag{301}$$

then the resultant iteration will have $\mathcal{O}(h^3)$ locally and $\mathcal{O}(h^2)$ totally. We remember that Eq. 295 obeys Eq. 297 if

$$\alpha = 1, \beta = 1, a = 1/2, b = 1/2. \tag{302}$$

Let us move on to the Runge Kutta (run guh kud duh; RK) method. Let us rewrite the previous formula as what is called a second order Runge Kutta method

$$y_{i+1} = y_i + h(c_1 k_1 + c_2 k_2) \tag{303}$$

since $i = \{1, 2\}$ on $c_i k_i$. Eq. 303 equals Eq. 295 only if

$$c_1 = 1/2, \quad c_2 = 1/2, \quad k_1 = f(x_i, y_i), \quad k_2 = f(\underbrace{x_i + h}_{x_{i+1}}, y_i + hf(x_i, y_i)). \tag{304}$$

A step further, we see that k_1 is contained in part of k_2. So we can instead say

$$c_1 = 1/2, \quad c_2 = 1/2, \quad k_1 = f(x_i, y_i), \quad k_2 = f(\underbrace{x_i + h}_{x_{i+1}}, y_i + hk_2). \tag{305}$$

Note that k_i can only be a function of k_j for $i > j$. We can use this same Runge Kutta procedure to develop higher order methods. A fourth order Runge Kutta method

$$y_{i+1} = y_1 + h(c_1 k_1 + c_2 k_2 + c_3 k_3 + c_4 k_4) \tag{306}$$

equals the example

$$y_{i+1} = y_i + \frac{h}{2}\Big(f(x_i, y_i) + f(x_{i+1}, y^e_{i+1})\Big) = y_i + \frac{h}{2}\Big(f(x_i, y_i) + f(x_{i+1}, y_i + hf(x_i, y_i))\Big) \tag{307}$$

(which is Eq. 295 but repeated here for convenience) if

$$c_1 = 1/6, \quad c_2 = 1/3, \quad c_3 = 1/3, \quad c_4 = 1/6; \tag{308}$$

moreover that

$$k_1 = f(x_i, y_i), \quad k_2 = f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hk_1),$$
$$k_3 = f(x_i + \frac{1}{2}h, y_i + \frac{1}{2}hk_2), k_4 = f(x_i + h, y_i + hk_3). \tag{309}$$

Now a fourth order Runge Kutta method admits a total $\mathcal{O}(h^4)$ (so local $\mathcal{O}(h^5)$). So the order of the method is the order of magnitude of the total truncation error. (A higher magnitude order for truncation error is actually good. Higher order terms are later on in the Taylor series expansion.)
So we know arbitrarily of our error but how do we better estimate it? Then, how do we control it? First we can try solving the problem multiple times with different h but with the same formula. This is "OK" but not preferred; some times you would need a very small h (Dargush). Instead you can use an adaptive algorithm that estimates the error at each step. Let us consider the "adaptive" Runge Kutta method. Here you combine a fourth order RK with a fifth order RK using the same evaluation points but with different coefficients. The psuedocode of this is

```
I.   Take a step of size h.
II.  Evaluate y_{i+1} and estimate the error e_{i+1}.
III. Is e_{i+1} < e_{tolerance}?
     A. If yes, then accept y_{i+1} and perhaps increase h;
     B. if no, reduce h and repeat step.
```

The MATLAB implementation of this is

ode45 .

Let us now consider the Runge Kutta Fehlberg method. Another type of fourth order RK is

$$\hat{y}_{i+1} = y_i + h(c_1k_1 + 0 + c_3k_3 + c_4k_4 + c_5k_5); \tag{310}$$

similarly, a fifth order RK can be

$$y_{i+1} = y_i + h(c_1k_1 + 0 + c_3k_3 + c_4k_4 + c_5k_5 + c_6k_6), \tag{311}$$

noticing for both cases that the $i = 2$ term is diminished. For these RK,

$$k_1 = f(x_i, y_i), \tag{312}$$

$$k_2 = f(x_i + a_2h, y_i + b_1hk_1), \tag{313}$$

$$k_3 = f(x_i + a_3h, y_i + b_2hk_1 + b_3hk_2), \tag{314}$$

$$k_4 = f(x_i + a_4h, y_i + b_4hk_1 + b_5hk_2 + b_6hk_3), \tag{315}$$

$$k_5 = f(x_i + a_5h, y_i + b_7hk_1 + b_8hk_2 + b_9hk_3 + b_{10}hk_4), \tag{316}$$

$$k_6 = f(x_i + a_6h, y_i + b_{11}hk_1 + b_{12}hk_2 + b_{13}hk_3 + b_{14}hk_4 + b_{15}hk_5). \tag{317}$$

For the coefficients a_i, b_i, see Rao (2002). Then the error can be computed as

$$e = y_{i+1} - \hat{y}_{i+1}. \tag{318}$$

We can compare this error to some tolerance that we have established and modify h accordingly (increase with success, decrease with failure).
Let us consider a multi step method. Here we use information from several previous steps to inform y_{i+1}. We interpolate over the previous steps $\Big(x_i, x_{i-1}, x_{i-2}, \dots\Big)+\Big(y_i, y_{i-1}, y_{i-2}, \dots\Big)$. Then we extrapolate to estimate y_{i+1} at x_{i+1}. This process is guided by a Taylor series expansion.
The Adams Bashford formulas are examples of this. The fourth order method (total $\mathcal{O}(h^4)$)

$$y_{i+1} = y_i + \frac{h}{24}\Big(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}\Big) \tag{319}$$

where $f_i = f(x_i, y_i)$ is explicit. It only uses known values on RHS. On the other hand the Adams Moulton formulas $\mathcal{O}(h^4)$

$$y_{i+1} = y_i + \frac{h}{24}\Big(9f_{i+1} + 19f_{i-1} - 5f_{i-1} + f_{i-2}\Big) \tag{320}$$

is implicit in that f_{i+1} is unknown. So that term is estimated or found simultaneously. Lastly a predictor-corrector method uses an explicit formula to predict y_{i+1}, then uses an implicit formula to correct in order to improve the solution. The $y^{(1)}_{i+1}$ term in

$$y^{(1)}_{i+1} = y_i + \frac{h}{24}\Big(55f_i - 59f_{i-1} + 37f_{i-2} - 9f_{i-3}\Big) \tag{321}$$

is used in the $f_{i+1}^{(j)}$ term in

$$y_{i+1}^{(j+1)} = y_i + \frac{h}{24}\left(9f_{i+1}^{(j)} + 19f_i - 5f_{i-1} + f_{i-2}\right). \tag{322}$$

This is called the Adams predictor-corrector. You can iterate over j until convergence is achieved.

2.9 Lec 2i Simultaneous ODEs

Now consider a set of first order ODEs

$$\frac{dy_1}{dx} = f_1(x, y_1, y_2, ..., y_n) \tag{323}$$

$$\frac{dy_2}{dx} = f_2(x, y_1, y_2, ..., y_n)\ldots \tag{324}$$

$$\frac{dy_i}{dx} = f_i(x, y_1, y_2, ..., y_n)\ldots \tag{325}$$

$$\frac{dy_n}{dx} = f_n(x, y_1, y_2, ..., y_n) \tag{326}$$

with the corresponding set of initial conditions

$$y_1(x_0) = y_{1,0}, \quad y_2(x_0) = y_{2,0}, \quad ..., y_n(x_0) = y_{n,0}. \tag{327}$$

We can rewrite this in vector notation as

$$\frac{d\mathbf{y}}{dx} = \mathbf{f}(x, \mathbf{y}), \quad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{328}$$

$$\mathbf{y} = \begin{Bmatrix} y_1(x) \\ y_2(x) \\ \ldots \\ y_n(x) \end{Bmatrix}. \tag{329}$$

We can use all previous methodologies to solve a system. The extension is straightforward. For example the Euler forward method for the system is

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}(x_i, \mathbf{y}_i) \tag{330}$$

and the Adams predictor-corrector is

$$\mathbf{y}_{i+1}^{(1)} = \mathbf{y}_i + \frac{h}{24}\left(55\mathbf{f}_i - 59\mathbf{f}_{i-1} + 37\mathbf{f}_{i-2} - 9\mathbf{f}_{i-3}\right) \tag{331}$$

where the $\mathbf{y}_{i+1}^{(1)}$ term is plugged into $\mathbf{f}_{i+1}^{(j)} = \mathbf{f}^{(j)}(x_{i+1}, \mathbf{y}_{i+1})$ in

$$\mathbf{y}_{i+1}^{(j+1)} = \mathbf{y}_i + \frac{h}{24}\left(9\mathbf{f}_{i+1}^{(j)} + 19\mathbf{f}_i - 5\mathbf{f}_{i-1} + \mathbf{f}_{i-2}\right). \tag{332}$$

As far as the Runge Kutta Fehlberg method for the ODE set, we have

$$\hat{\mathbf{y}}_{i+1} = \mathbf{y}_i + h(c_1\mathbf{k}_1 + 0 + c_3\mathbf{k}_3 + c_4\mathbf{k}_4 + c_5\mathbf{k}_5); \tag{333}$$

and

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h(c_1\mathbf{k}_1 + 0 + c_3\mathbf{k}_3 + c_4\mathbf{k}_4 + c_5\mathbf{k}_5 + c_6\mathbf{k}_6), \tag{334}$$

where

$$\mathbf{k}_1 = \mathbf{f}(x_i, \mathbf{y}_i), \tag{335}$$

$$\mathbf{k}_2 = \mathbf{f}(x_i + a_2h, \mathbf{y}_i + b_1h\mathbf{k}_1), \tag{336}$$

$$\mathbf{k}_3 = \mathbf{f}(x_i + a_3h, \mathbf{y}_i + b_2h\mathbf{k}_1 + b_3h\mathbf{k}_2), \tag{337}$$

$$\mathbf{k}_4 = \mathbf{f}(x_i + a_4h, \mathbf{y}_i + b_4h\mathbf{k}_1 + b_5h\mathbf{k}_2 + b_6h\mathbf{k}_3), \tag{338}$$

$$\mathbf{k}_5 = \mathbf{f}(x_i + a_5h, \mathbf{y}_i + b_7h\mathbf{k}_1 + b_8h\mathbf{k}_2 + b_9h\mathbf{k}_3 + b_{10}h\mathbf{k}_4), \tag{339}$$

$$\mathbf{k}_6 = \mathbf{f}(x_i + a_6h, \mathbf{y}_i + b_{11}h\mathbf{k}_1 + b_{12}h\mathbf{k}_2 + b_{13}h\mathbf{k}_3 + b_{14}h\mathbf{k}_4 + b_{15}h\mathbf{k}_5). \tag{340}$$

Which is the same formulation as earlier except $\mathbf{f}, \{\mathbf{k}_i\}, \mathbf{y}_i$ are all vectors. Then the error is the L2 norm of Eq. 318, or

$$e = ||\mathbf{y}_{i+1} - \hat{\mathbf{y}}_{i+1}||_2. \tag{341}$$

So far we have discussed first order ODEs. Let us now consider the higher, second order ODE

$$m\frac{d^2u}{dt^2} + c\frac{du}{dt} + ku = P(t), \quad u(0) = u_0, \quad \frac{du}{dt}(0) = v_0. \tag{342}$$

This models a spring mass system with spring constant k, mass m, displacement u, damping constant c, external force P, and initial conditions on displacement and velocity $u_0, \quad v_0$. This is visualized in Fig. 3. Now let us redefine some parameters in

$$t \to x, u(t) \to y_1(x), \frac{du}{dt} \to y_2(x), \tag{343}$$

which is called the state space approach. Then Eq. 342 becomes

$$m\frac{dy_2}{dx} + cy_2 + ky_1 = P(x). \tag{344}$$

Rearranging,

$$\frac{dy_2}{dx} = \frac{1}{m}\bigg(-cy_2 - ky_1 + P(x)\bigg) = -\frac{cy_2}{m} - \frac{ky_1}{m} + \frac{P}{m}. \tag{345}$$

By virtue of Eq. 343 we already know that

$$\frac{dy_1}{dx} = y_2; \tag{346}$$

also that the initial conditions become

$$y_1(0) = u_0, y_2(0) = v_0. \tag{347}$$

The new representations that are Eqs. 346 and 345 can be integrated into the matrix equation

$$\begin{Bmatrix} \frac{dy_1}{dx} \\ \frac{dy_2}{dx} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} + \begin{Bmatrix} 0 \\ P/m \end{Bmatrix} \Longleftrightarrow \frac{d\mathbf{y}}{dx} = \underbrace{\mathbf{A}\mathbf{y} + \mathbf{b}}_{\mathbf{f}}. \tag{348}$$

$\mathbf{f} = \frac{d\mathbf{y}}{dx}$ because of Eq. 328. It was our original consideration. Also we can generalize the initial conditions as

$$\mathbf{y}(0) = \mathbf{y}_0 \Longleftrightarrow \begin{Bmatrix} y_1(0) \\ y_2(0) \end{Bmatrix} = \begin{Bmatrix} u_0 \\ v_0 \end{Bmatrix} = \begin{Bmatrix} y_{1,0} \\ y_{2,0} \end{Bmatrix}. \tag{349}$$

We wish to state the problem in terms of physical quantities that are meaningful. In particular we want natural frequency ω and damping ratio ξ. Therefore let

$$\omega^2 = k/m, \quad 2\xi\omega = c/m. \tag{350}$$

Then

$$\begin{Bmatrix} \frac{dy_1}{dx} \\ \frac{dy_2}{dx} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2\xi\omega \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} + \begin{Bmatrix} 0 \\ P/m \end{Bmatrix}, \quad \begin{Bmatrix} y_1(0) \\ y_2(0) \end{Bmatrix} = \begin{Bmatrix} u_0 \\ v_0 \end{Bmatrix}. \tag{351}$$

Also, letting

$$P = 0 \tag{352}$$

permits the system to freely vibrate. Now recalling from Eq. 348 that $\mathbf{f} = \frac{d\mathbf{y}}{dx} = \mathbf{A}\mathbf{y} + \mathbf{b}$, we can use Euler's forward method to say that

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f} = \mathbf{y}_i + h\mathbf{A}\mathbf{y}_i + h\mathbf{b}. \tag{353}$$

We have just considered as an example a spring mass system. As another example let us consider the pendulum in Fig. 3. Here the pendulum is swinging through a viscoous fluid and so it experiences a resisting force proportional to its velocity $cL\dot{\theta}$ as well as the usual tension force T and gravitational force mg.
The sum of the torques (moments) at origin O is equal to the time rate of change of the angular momentums at that same point. That is,

$$\sum M_{0_z} = \dot{H}_{0_z}. \tag{354}$$

Recall that generally speaking torque

$$\tau = M = r \times F = rF \sin\phi \tag{355}$$

where ϕ is the angle between the contributing force and the distance between the body and some axis of interest. It is useful to make $\phi = 90^o$. So for example, since gravity acts straight down, r will be the distance between the ball and the vertical axis so that r is a horizontal space between two vertical lines and

$$M_{mg} = rF \sin\phi = (L\sin\theta)(-mg)\underbrace{\sin\phi}_{\phi=90^o} = -mgL\sin\theta. \tag{356}$$

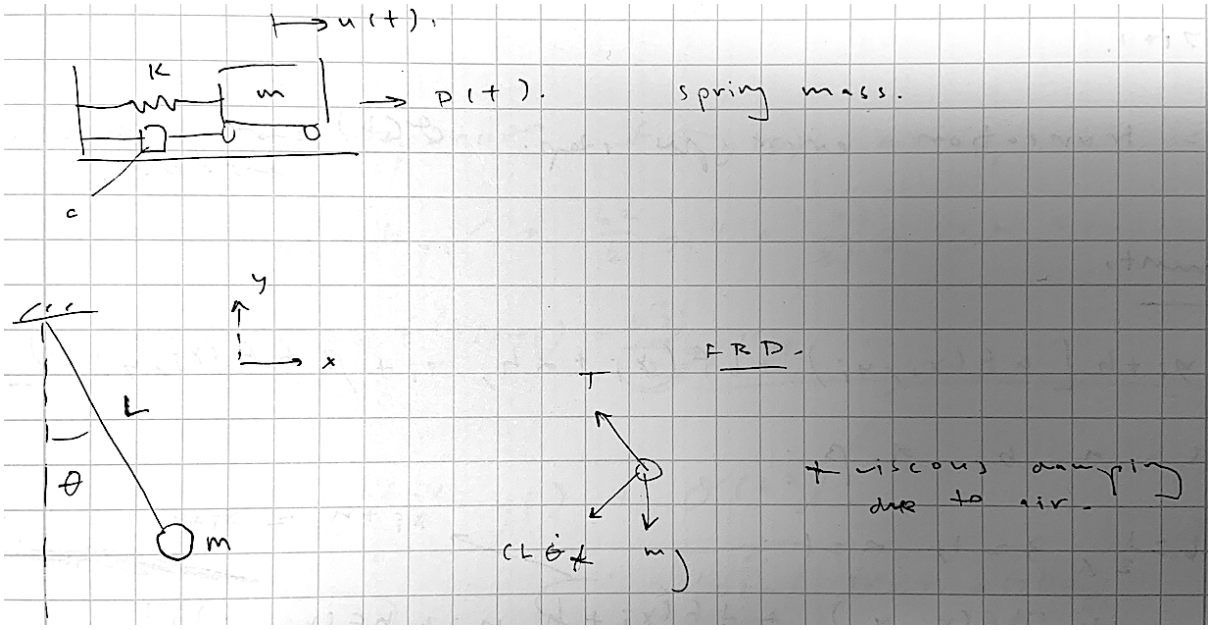

Figure 3: Spring-mass system and pendulum.

Viscous force acts perpendicularly to the wire and the distance between the ball and the origin O with respect to the perpendicular axis is L. So

$$M_{viscous} = -(cl\dot{\theta})L\sin\phi = -cL^2\dot{\theta}. \tag{357}$$

These two equations comprise LHS. Then on the right hand side is the time derivative of the angular momentum, where angular momentum H is linear momentum times radius. Generally speaking

$$\rho = mv = m(L\dot{\theta}) \tag{358}$$

because the velocity of a point on the wire increases as you go further down it. Then

$$H_0 = (mL\dot{\theta})L \rightarrow \dot{H}_0 = \frac{d}{dt}\left[(mL\dot{\theta})L\right] \tag{359}$$

which makes up RHS. Together,

$$-cL^2\dot{\theta} - mgL\sin\theta = mL^2\ddot{\theta} \tag{360}$$

implies

$$mL^2\ddot{\theta} + mgL\sin\theta + cL^2\dot{\theta} = 0 \tag{361}$$

implies

$$\ddot{\theta} + \frac{c}{m}\dot{\theta} + \frac{g}{L}\sin\theta = 0. \tag{362}$$

Eq. 362 can be simplified with the Taylor series

$$\sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \ldots \approx \theta \tag{363}$$

so that now,

$$\ddot{\theta} + \frac{c}{m}\dot{\theta} + \frac{g}{L}\theta = 0. \tag{364}$$

Substituting physical quantities in from Eq. 350,

$$\ddot{\theta} + 2\xi\omega\dot{\theta} + \omega^2\theta = 0. \tag{365}$$

Using the state space approach, we invoke a process similar to Eq. 343, which is

$$x \leftarrow t, \quad \theta \leftarrow y_1, \dot{\theta} \leftarrow y_2. \tag{366}$$

Then

$$\frac{dy_1}{dx} = y_2 \tag{367}$$

and

$$\frac{dy_2}{dx} = \ddot{\theta} = -2\xi\omega\dot{\theta} - \omega^2\theta = -2\xi\omega y_2 - \omega^2 y_1. \tag{368}$$

This information is sufficient to build the matrix equation

$$\begin{Bmatrix} \frac{dy_1}{dx} \\ \frac{dy_2}{dx} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2\xi\omega \end{bmatrix} \begin{Bmatrix} y_1 \\ y_2 \end{Bmatrix} + \begin{Bmatrix} 0 \\ 0 \end{Bmatrix} = \mathbf{A}\mathbf{y} + \mathbf{b} = \mathbf{f} = \frac{d\mathbf{y}}{dx} \quad (\mathbf{b} = \mathbf{0}). \tag{369}$$

Not using the Taylor series approximation and thus leaving the matrix equation system as

$$\frac{dy_1}{dx} = y_2 \tag{370}$$

and

$$\frac{dy_2}{dx} = -2\xi\omega y_2 - \omega^2 \sin y_1 \tag{371}$$

admits a set of nonlinear ODEs. Even one nonlinear equation in a set redefines the whole set as nonlinear.

2.10 Lec 2j State space dynamics and stability

To summarize Fig. 3, the spring mass system can be written as

$$m\ddot{u} + c\dot{u} + ku = P \rightarrow \ddot{u} = -\frac{c}{m}\dot{u} - \frac{k}{m}u + P/m \tag{372}$$

which implies

$$\begin{Bmatrix} \dot{u} \\ \dot{v} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix} \begin{Bmatrix} u \\ v \end{Bmatrix} + \begin{Bmatrix} 0 \\ P/m \end{Bmatrix} \tag{373}$$

provided $\dot{u} = v \rightarrow \ddot{u} = \dot{v}$. Likewise the pendulum is written as

$$\ddot{\theta} + \frac{c}{m}\dot{\theta} + \frac{g}{L}\sin\theta = 0 \tag{374}$$

which implies

$$\begin{Bmatrix} \dot{\theta} \\ \dot{\Omega} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -c/m & -g/L \end{bmatrix} \begin{Bmatrix} \theta \\ \Omega \end{Bmatrix} \tag{375}$$

provided $\dot{\theta} = \Omega \rightarrow \ddot{\theta} = \dot{\Omega}$. The general structure of these equations (especially seen in the spring mass system) is

$$\mathbf{M}\dot{\mathbf{v}} + \mathbf{C}\mathbf{v} + \mathbf{K}\mathbf{u} = \mathbf{P}(t). \tag{376}$$

The conversion of this system to state space is

$$\mathbf{y} = \begin{Bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{Bmatrix} = \begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix}. \tag{377}$$

We understand

$$\dot{\mathbf{u}} = \mathbf{v}; \tag{378}$$

also that, after rearranging Eq. 376,

$$\dot{\mathbf{v}} = \mathbf{M}^{-1}\Big(-\mathbf{K}\mathbf{u} - \mathbf{C}\mathbf{v} + \mathbf{P}\Big) = -\mathbf{M}^{-1}\mathbf{K}\mathbf{u} - \mathbf{M}^{-1}\mathbf{C}\mathbf{v} + \mathbf{M}^{-1}\mathbf{P}. \tag{379}$$

Then

$$\underbrace{\begin{Bmatrix} \dot{\mathbf{u}} \\ \dot{\mathbf{v}} \end{Bmatrix}}_{\mathbf{f}} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\mathbf{M}^{-1}\mathbf{K} & -\mathbf{M}^{-1}\mathbf{C} \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{Bmatrix} \mathbf{u} \\ \mathbf{v} \end{Bmatrix}}_{\mathbf{y}} + \underbrace{\begin{Bmatrix} \mathbf{0} \\ \mathbf{M}^{-1}\mathbf{P} \end{Bmatrix}}_{\mathbf{b}}. \tag{380}$$

Let us apply this to Euler integration. Generally,

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}_i = \mathbf{y}_i + h\mathbf{A}\mathbf{y}_i + h\mathbf{b}_i. \tag{381}$$

This implies

$$\mathbf{y}_{i+1} = \Big(\mathbf{I} + h\mathbf{A}\Big)\mathbf{y}_i + h\mathbf{b}_i. \tag{382}$$

Starting at the first step:

$$\mathbf{y}_1 = \Big(\mathbf{I} + h\mathbf{A}\Big)\mathbf{y}_0 + h\mathbf{b}_0. \tag{383}$$

Then,

$$\mathbf{y}_2 = \Big(\mathbf{I} + h\mathbf{A}\Big)(\mathbf{y}_1) + h\mathbf{b}_1$$

$$= \Big(\mathbf{I} + h\mathbf{A}\Big)(\Big(\mathbf{I} + h\mathbf{A}\Big)\mathbf{y}_0 + h\mathbf{b}_0) + h\mathbf{b}_1$$

implies

$$\mathbf{y}_2 = \Big(\mathbf{I} + h\mathbf{A}\Big)^2\mathbf{y}_0 + \Big(\mathbf{I} + h\mathbf{A}\Big)h\mathbf{b}_0 + h\mathbf{b}_1. \tag{384}$$

In general,

$$\mathbf{y}_n = \Big(\mathbf{I} + h\mathbf{A}\Big)^n\mathbf{y}_0 + \mathbf{g}_n(\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_n). \tag{385}$$

$\mathbf{g}$ is some function of $\mathbf{b}_i$. Matrix

$$\hat{\mathbf{A}} = \Big(\mathbf{I} + h\mathbf{A}\Big) \tag{386}$$

when expressed in

$$\mathbf{y}_n = \hat{\mathbf{A}}^n \mathbf{y}_0 + \mathbf{g} \tag{387}$$

is called the Jordan canonical form. The spectral decomposition

$$\hat{\mathbf{A}} = \mathbf{PJP}^{-1} \tag{388}$$

admits the Jordan normal form $\mathbf{J}$ which is a matrix containing the eigenvalues of $\hat{\mathbf{A}}$ on its diagonal and sometimes the superdiagonal (the parallel line of elements right above the diagonal). Raised to the power n,

$$\hat{\mathbf{A}}^n = \left(\mathbf{PJP}^{-1}\right)\left(\mathbf{PJP}^{-1}\right)\left(\mathbf{PJP}^{-1}\right)\ldots = \mathbf{PJ}^n\mathbf{P}^{-1}. \tag{389}$$

We do not want $\mathbf{J}^n$ to increase to infinity as $n \to \infty$. That is, we want $\mathbf{J}^n$ to be bounded. This is true if the spectral radius of $\hat{\mathbf{A}} = \rho(\hat{\mathbf{A}}) \leq 1$. Spectral radius is defined as the maximum of the absolute values of its eigenvalues. That is,

$$\rho(\hat{\mathbf{A}}) = \max\{|\lambda_1|, |\lambda_2|, \ldots |\lambda_n|\}. \tag{390}$$

We consider an example of using the Jordan canonical form. Let

$$\dot{x}_1 = -3x_1 - 2x_2, \quad \dot{x}_2 = 2x_1 + x_2; \quad x_1(0) = 1, \quad x_2(0) = 0. \tag{391}$$

Then

$$\begin{Bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{Bmatrix} = \underbrace{\begin{bmatrix} -3 & -2 \\ 2 & 1 \end{bmatrix}}_{\mathbf{A}} \begin{Bmatrix} x_1 \\ x_2 \end{Bmatrix}. \tag{392}$$

For a linear system

$$\mathbf{x}' = \mathbf{Ax}, \tag{393}$$

if $\{\lambda, \mathbf{r}\}$ is an eigenpair for $\mathbf{A}$, then the solution of the system is

$$\mathbf{x} = e^{\lambda t}\mathbf{r} \to \mathbf{x}' = \lambda e^{\lambda t}\mathbf{r} \tag{394}$$

because

$$\mathbf{Ax} = \mathbf{A}e^{\lambda t}\mathbf{r} = e^{\lambda t}\mathbf{Ar} = \lambda e^{\lambda t}\mathbf{r} = \mathbf{x}' \tag{395}$$

implies the eigenproblem

$$e^{\lambda t}\mathbf{Ar} = e^{\lambda t}\lambda\mathbf{r} \to \mathbf{Ar} = \lambda\mathbf{r}. \tag{396}$$

So we need to find the eigenpairs of the matrix $\mathbf{A}$. This requires

$$0 = \det(\mathbf{A} - \lambda\mathbf{I}) = \det\begin{bmatrix} -3-\lambda & -2 \\ 2-\lambda & 1 \end{bmatrix} = (-3-\lambda)(1-\lambda) - (2)(-2) = 0. \tag{397}$$

Solving for λ,

$$\lambda^2 + 2\lambda + 1 = 0 \to (\lambda+1)(\lambda+1) = 0 \to \lambda_1, \lambda_2 = -1. \tag{398}$$

The corresponding eigenvector $\mathbf{r}$ can be found in

$$\mathbf{A}\mathbf{r} = \lambda\mathbf{r} \rightarrow \begin{bmatrix} -3 & -2 \\ 2 & 1 \end{bmatrix} \begin{Bmatrix} r_1 \\ r_2 \end{Bmatrix} = \begin{Bmatrix} -r_1 \\ -r_2 \end{Bmatrix} \tag{399}$$

by considering algebraically

$$-3r_1 - 2r_2 = -r_1 \rightarrow -2r_2 = 2r_1 \rightarrow r_2 = -r_1$$

$$\rightarrow \mathbf{r} = \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} \rightarrow \text{normalizing if desired/necessary} \rightarrow \mathbf{r} = \begin{Bmatrix} 1\sqrt{2} \\ -1\sqrt{2} \end{Bmatrix}. \tag{400}$$

Noonberg (2010) pg. 166-167 proves that if $\lambda_1 = \lambda_2$ for $[\mathbf{A}]_{2\text{x}2}$ then the general solution of $\mathbf{A}\mathbf{x} = \mathbf{x}'$ can be written in the form

$$\mathbf{x} = c_1 e^{\lambda t}\mathbf{r} + c_2 e^{\lambda t}(t\mathbf{r} + \mathbf{r}^*), \quad \text{where} \quad (\mathbf{A} - \lambda\mathbf{I})\mathbf{r}^* = \mathbf{r}. \tag{401}$$

This is also called variation of parameters. In this case,

$$\underbrace{\begin{bmatrix} -2 & -2 \\ 2 & 2 \end{bmatrix}}_{\mathbf{A}-\lambda\mathbf{I}} \underbrace{\begin{Bmatrix} 1 \\ -3/2 \end{Bmatrix}}_{\mathbf{r}^*} = \underbrace{\begin{Bmatrix} 1 \\ -1 \end{Bmatrix}}_{\mathbf{r}}. \tag{402}$$

So we can write the general solution as

$$\mathbf{x} = c_1 e^{-t} \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} + c_2 e^{-t} \begin{Bmatrix} t+1 \\ -t-3/2 \end{Bmatrix} = \mathbf{x}(t). \tag{403}$$

That means

$$\mathbf{x}(0) = c_1 \begin{Bmatrix} 1 \\ -1 \end{Bmatrix} + c_2 \begin{Bmatrix} 1 \\ -3/2 \end{Bmatrix} = \begin{Bmatrix} x_1(0) \\ x_2(0) \end{Bmatrix} = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} \tag{404}$$

where $\{x_1(0), x_2(0)\}^{\text{T}}$ comes from Eq. 391. Recall that we want to bound $\mathbf{J}^n$ by causing the spectral radius of $\hat{\mathbf{A}}$ - that is, its maximum eigenvalue - to be less than or equal to 1. This is true if, using Eq. 386 and 390,

$$\max\{|\lambda_1|, |\lambda_2|\} \leq 1, \tag{405}$$

where λ_i are recovered in letting

$$0 = \det(\hat{\mathbf{A}} - \lambda I) = \det(\mathbf{I} + h\mathbf{A} - \lambda I) = \det \begin{bmatrix} 1 - 3h - \lambda & -2h \\ 2h & 1 + h - \lambda \end{bmatrix}$$

$$= (1 - 3h - \lambda)(1 + h - \lambda) - (2h)(-2h)$$

$$= -h^2 - 2h\lambda + 2h - \lambda^2 + 2\lambda - 1 = 0 \rightarrow \lambda_1 = \lambda_2 = 1 - h. \tag{406}$$

To satisfy the enforcement of Eq. 405,

$$|1 - h| \leq 1 \rightarrow -2 \leq h \leq 2. \tag{407}$$

Now recall once again the spring mass system

$$\begin{Bmatrix} \dot{u} \\ \dot{v} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix} \begin{Bmatrix} u \\ v \end{Bmatrix} + \begin{Bmatrix} 0 \\ P/m \end{Bmatrix}. \tag{408}$$

Consider the undamped case. That is, $c = 0$. Remembering also that natural frequency $\omega^2 = k/m$,

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & 0 \end{bmatrix}; \tag{409}$$

eigenvalues are brought out of

$$0 = \det \begin{bmatrix} -\lambda & 1 \\ -\omega^2 & -\lambda \end{bmatrix} = \lambda^2 + \omega^2 = 0 \rightarrow \lambda_1, \lambda_2 = \pm i\omega. \tag{410}$$

Eigenvalues of the Jordan canonical form are in

$$0 = \det(\mathbf{I} + h\mathbf{A} - \lambda\mathbf{I}) = \det \begin{bmatrix} 1-\lambda & h \\ -h\omega^2 & 1-\lambda \end{bmatrix} = (1-\lambda)(1-\lambda) - (-h\omega^2)(h)$$

$$= 1 - 2\lambda + \lambda^2 + h^2\omega^2 = 0 \rightarrow \begin{Bmatrix} \lambda_1 \\ \lambda_2 \end{Bmatrix} = \begin{Bmatrix} i(h\omega - i) = 1 + h\omega i \\ -i(h\omega + i) = 1 - h\omega i \end{Bmatrix}. \tag{411}$$

$\lambda = 1 + ih\omega$ implies

$$|\lambda| = \sqrt{1^2 + i^2 h^2 \omega^2} = \sqrt{1 + h^2\omega^2}. \tag{412}$$

$h^2\omega^2$ is always positive, so $\lambda \geq 1$ always, which means the rule can never be enforced, making $\mathbf{J}^n$ unbounded and $\hat{\mathbf{A}}$ unstable always.
We consider a stiff ODE system as another example. Let

$$\frac{du}{dt} = (\beta - 2)u + (2\beta - 2)v, \tag{413}$$

$$\frac{dv}{dt} = (1 - \beta)u + (1 - 2\beta)v, \tag{414}$$

as well as

$$u(0) = 1, \quad v(0) = 0. \tag{415}$$

The exact solution for $\beta > 2$ is

$$u(t) = 2e^{-t} - e^{-\beta t}, \tag{416}$$

$$v(t) = -e^{-t} + e^{-\beta t}. \tag{417}$$

where $\lambda_1 = -1, \lambda_2 = -\beta$. For a large β, the terms containing β will decay very fast with time. Now, for any explicit (Euler forward) method, one must resolve the faster scale. Otherwise severe instability will be brought into the system. For this problem consider a very general linear set of ODEs with constant coefficients

$$\dot{\mathbf{y}} = \mathbf{A}\mathbf{y} \Longleftrightarrow \begin{Bmatrix} \frac{du}{dt} \\ \frac{dv}{dt} \end{Bmatrix} = \begin{bmatrix} \beta - 2 & 2\beta - 2 \\ 1 - \beta & 1 - 2\beta \end{bmatrix} \begin{Bmatrix} u \\ v \end{Bmatrix}. \tag{418}$$

As played out in Eqs. 393-396, we can assume the solution

$$\mathbf{y} = \boldsymbol{\phi} e^{\lambda t} \tag{419}$$

leading to

$$\dot{\mathbf{y}} = \lambda \boldsymbol{\phi} e^{\lambda t} \tag{420}$$

and then, substituting into the original linear set,

$$\lambda \boldsymbol{\phi} e^{\lambda t} = \mathbf{A} \boldsymbol{\phi} e^{\lambda t} \tag{421}$$

which implies the standard eigenproblem

$$\lambda \boldsymbol{\phi} = \mathbf{A} \boldsymbol{\phi}. \tag{422}$$

For the previous problem,

$$\mathbf{A} = \begin{bmatrix} \beta - 2 & 2\beta - 2 \\ 1 - \beta & 1 - 2\beta \end{bmatrix} \tag{423}$$

and $\lambda_1 = -1, \lambda_2 = -\beta$. For a large $\beta = 100$,

```
% Input file
clear
clc
beta = 100
A = [(beta-2), (2*beta -2); (1-beta), (1-2*beta)];
[Lambda, Phi] = eig(A)
```

```
% Command window
beta =

    100


Lambda =

    0.8944   -0.7071
   -0.4472    0.7071


Phi =

   -1     0
    0  -100
```

The assumed solution

$$\mathbf{y} = \boldsymbol{\phi} e^{\lambda t} \tag{424}$$

for two eigenvalues requires a general solution in the form of

$$\mathbf{y} = c_1\boldsymbol{\phi}_1 e^{\lambda_1 t} + c_2\boldsymbol{\phi}_2 e^{\lambda_2 t}. \tag{425}$$

Now suppose we want to find c_1, c_2 which satisfies the initial conditions

$$y_1(0) = 0, \quad y_2(0) = 0 \leftrightarrow u(0) = 0, \quad v(0) = 0. \tag{426}$$

To do this we have at our disposal Euler's method or one of the Runge Kutta methods. For a large β the second term $\lambda_2 = -\beta$ decays very fast and thus requires treatment to avoid instability.
Evaluation of the eigenvalues of the Jordan canonical form of $\mathbf{A}$ - that is, $\hat{\mathbf{A}}$, brings

$$0 = \det(\underbrace{\mathbf{I} + h\mathbf{A}}_{\hat{\mathbf{A}}} - \lambda\mathbf{I}) \rightarrow \hat{\lambda}_1 = 1 - h, \hat{\lambda}_2 = 1 - \beta h. \tag{427}$$

Then to satisfy

$$\max\{|\lambda_1|, |\lambda_2|\} \leq 1, \tag{428}$$

we determine

$$|1 - h| \leq 1 \rightarrow -2 \leq h \leq 2 \tag{429}$$

and

$$|1 - \beta h| \leq 1 \rightarrow -2 \leq \beta h \leq 2 \rightarrow \frac{-2}{\beta} \leq h \leq \frac{2}{\beta}. \tag{430}$$

Therefore one would need a very small h for a large β; as β grows, $2/\beta$ shrinks and still h must be lesser than $2/\beta$.

Finally let us consider Euler's backward method for a linear system, which is a generalization and reindex of Eq. 287. That is,

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{f}_{i+1}. \tag{431}$$

This implies

$$\mathbf{y}_{i+1} = \mathbf{y}_i + h(\mathbf{A}\mathbf{y}_{i+1} + \mathbf{b}_{i+1}). \tag{432}$$

Rearranging,

$$\mathbf{y}_{i+1} - h\mathbf{A}\mathbf{y}_{i+1} = \Big(\mathbf{I} - h\mathbf{A}\Big)\mathbf{y}_{i+1} = \mathbf{y}_i + h\mathbf{b}_{i+1}. \tag{433}$$

We can isolate

$$\mathbf{y}_{i+1} = \Big(\mathbf{I} - h\mathbf{A}\Big)^{-1}(\mathbf{y}_i + h\mathbf{b}_{i+1}) \tag{434}$$

and let the Jordan canonical form be differently defined for the backward problem as

$$\hat{\mathbf{A}} = \Big(\mathbf{I} - h\mathbf{A}\Big)^{-1}. \tag{435}$$

Similarly to the forward problem, for stability we require that

$$\rho(\hat{\mathbf{A}}) \leq 1. \tag{436}$$

Finding the eigenvalues of this expression even as it is inverted is not difficult. One must only find the eigenvalues of the expression $\mathbf{I} - h\mathbf{A}$ and then invert each eigenvalue individually to receive those of $(\mathbf{I} - h\mathbf{A})^{-1}$.

2.11 Lec 2k Qualitative theory of ODEs

One must look at nonlinear problems with a different perspective. There is no superposition with nonlinear ODEs. Analytical solutions are also difficult or impossible to obtain. We can however use numerical methods. We do this with the goal of gathering information that provides insight into the character of general solutions. Poincaré initiated the qualitative theory of ODEs in the late 19th century. We can use qualitative theory for nonlinear phenomena in solids and fluids, and for control of nonlinear systems. The key issue as far as numerical methods go is stability.
The qualitative theory is a highly geometric approach as opposed to solving many equations. We consider the trajectory of a system for specified initial conditions in the phase space (or, with two dependent variables, a phase plane). A set of trajectories provides a phase portrait.
We can then characterize the nonlinear ODE according to the geometric pattern of the phase portrait. We can attempt to investigate the trajectories as $t \to \infty$.
We begin with a special linear case. Consider the 2nd order linear homogeneous ODE with constant coefficients

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x}. \tag{437}$$

We define the critical points as those in which

$$\frac{d\mathbf{x}}{dt} = 0 \tag{438}$$

and thus

$$\mathbf{A}\mathbf{x} = \mathbf{0}. \tag{439}$$

At the critical (also called equilibrium) points, for $\det \mathbf{A} \neq 0$ (i.e. if we have two linearly independent equations; i.e. if $\mathbf{A}$ is nonsingular; i.e. if $\mathbf{A}$ is invertible), then the only solution is at the origin $\mathbf{x} = \mathbf{0}$. That is the trivial solution.
Now let us classify the solution. Let

$$\mathbf{x}(t) = \mathbf{k}e^{\lambda t}. \tag{440}$$

From here we form the characteristic equation and find roots (eigenvalues). The nature of these eigenvalues determines the type of general solution.
If (I) both eigenvalues are real and of the same sign, then the general solution

$$\mathbf{x}(t) = c_1\mathbf{k}_1 e^{\lambda_1 t} + c_2\mathbf{k}_2 e^{\lambda_2 t}, \quad \lambda_1 \neq \lambda_2 \in \mathcal{R}. \tag{441}$$

If

$$\lambda_1 < \lambda_2 < 0, \tag{442}$$

then all trajectories approach the origin as $t \to \infty$. That is, $\mathbf{x}$ approaches $\mathbf{0}$ because if both λ_1, λ_2 are negative, then both terms $e^{\lambda_1 t}, e^{\lambda_2 t}$ will go to zero. Suppose we define $\mathbf{x}_0$ as the initial condition or the point in space in the phase plane. If $\mathbf{x}$ is on $\mathbf{k}_1$, then $c_2 = 0$ and, assuming still that λ_1 is negative, the trajectory will approach the origin in a straight line along $\mathbf{k}_1$. The converse is true if $\mathbf{x}$ is on $\mathbf{k}_2$: $c_1 = 0$ and the trajectory will approach $\mathbf{0}$ along $\mathbf{k}_2$.

Suppose we rewrote

$$\mathbf{x}(t) = c_1\mathbf{k}_1 e^{\lambda_1 t} + c_2\mathbf{k}_2 e^{\lambda_2 t} \tag{443}$$

as

$$\mathbf{x}(t) = e^{\lambda_2 t}[c_1\mathbf{k}_1 e^{(\lambda_1-\lambda_2)t} + c_2\mathbf{k}_2]. \tag{444}$$

Now as $t \to \infty$, the $c_1\mathbf{k}_1 e^{(\lambda_1-\lambda_2)t}$ term becomes negligible compared to the $c_2\mathbf{k}_2$ term because λ_1 and λ_2 are competing. This means essentially that

$$\lim_{t\to\infty} \mathbf{x}(t) = e^{\lambda_2 t}[c_2\mathbf{k}_2], \tag{445}$$

or

$$\lim_{t\to\infty} \mathbf{x}(t)e^{-\lambda_2 t} = c_2\mathbf{k}_2 \tag{446}$$

provided $c_2 \neq 0$. As time goes to infinity, our solutions will go to the critical or equilibrium point that is the origin. So we see that based on Eq. 446, trajectories near the critical point align with $\mathbf{k}_2$, as in Fig. 4. Here the equilibrium solution $\mathbf{x} = \mathbf{0}$ is stable asymptotically. We also say that the critical point is an improper node.
Now if

$$\lambda_1, \lambda_2 > 0, \tag{447}$$

then the same behavior occurs as in Fig. 4 except that the trajectories are in opposite directions. So they bend outward to infinity from alignment with $\mathbf{k}_2$. The equilibrium solution is unstable and the critical point is still an improper node.

Our first case (I) was that where the two eigenvalues was of the same sign, either positive or negative. Now let us consider if (II) the eigenvalues are real but of opposite sign. The general solution is still

$$\mathbf{x}(t) = c_1\mathbf{k}_1 e^{\lambda_1 t} + c_2\mathbf{k}_2 e^{\lambda_2 t}, \quad \lambda_1 \neq \lambda_2 \in \mathcal{R}. \tag{448}$$

If

$$\lambda_1 > 0 \text{ and } \lambda_2 < 0, \tag{449}$$

then for the initial conditions along $\mathbf{k}_1$ or $\mathbf{k}_2$, the trajectories are still straight along, but on $\mathbf{k}_1$ the trajectories point away from the origin (positive trajectory as $t \to \infty$) while on $\mathbf{k}_2$ the trajectories point toward the origin (negative trajectory as $t \to \infty$). The $c_1\mathbf{k}_1 e^{\lambda_1 t}$ (λ_1 positive) term begins to dominate as $t \to \infty$ while the $c_2\mathbf{k}_2 e^{\lambda_2 t}$ (λ_2 negative) term tends to zero. So, all other points tend asymptotically towards $\mathbf{k}_1$. The equilibrium solution is unstable because all lines tend away from the origin. Moreover the equilibrium point is called a saddle point. This case is shown in Fig. 5.

Now let us consider the case (III) where the two eigenvalues are equal. If (IIIi) we have two independent eigenvectors such that

$$\lambda_1 = \lambda_2 = \lambda, \quad \mathbf{x}(t) = c_1\mathbf{k}_1 e^{\lambda t} + c_2\mathbf{k}_2 e^{\lambda t}, \tag{450}$$

then for $\lambda < 0$, all trajectories approach the origin along a straight line as $t \to \infty$. The equilibrium point in this case is called a proper node. The phase space is asymptotically stable. For $\lambda > 0$ the trajectories approach infinity along the same straight lines. Now

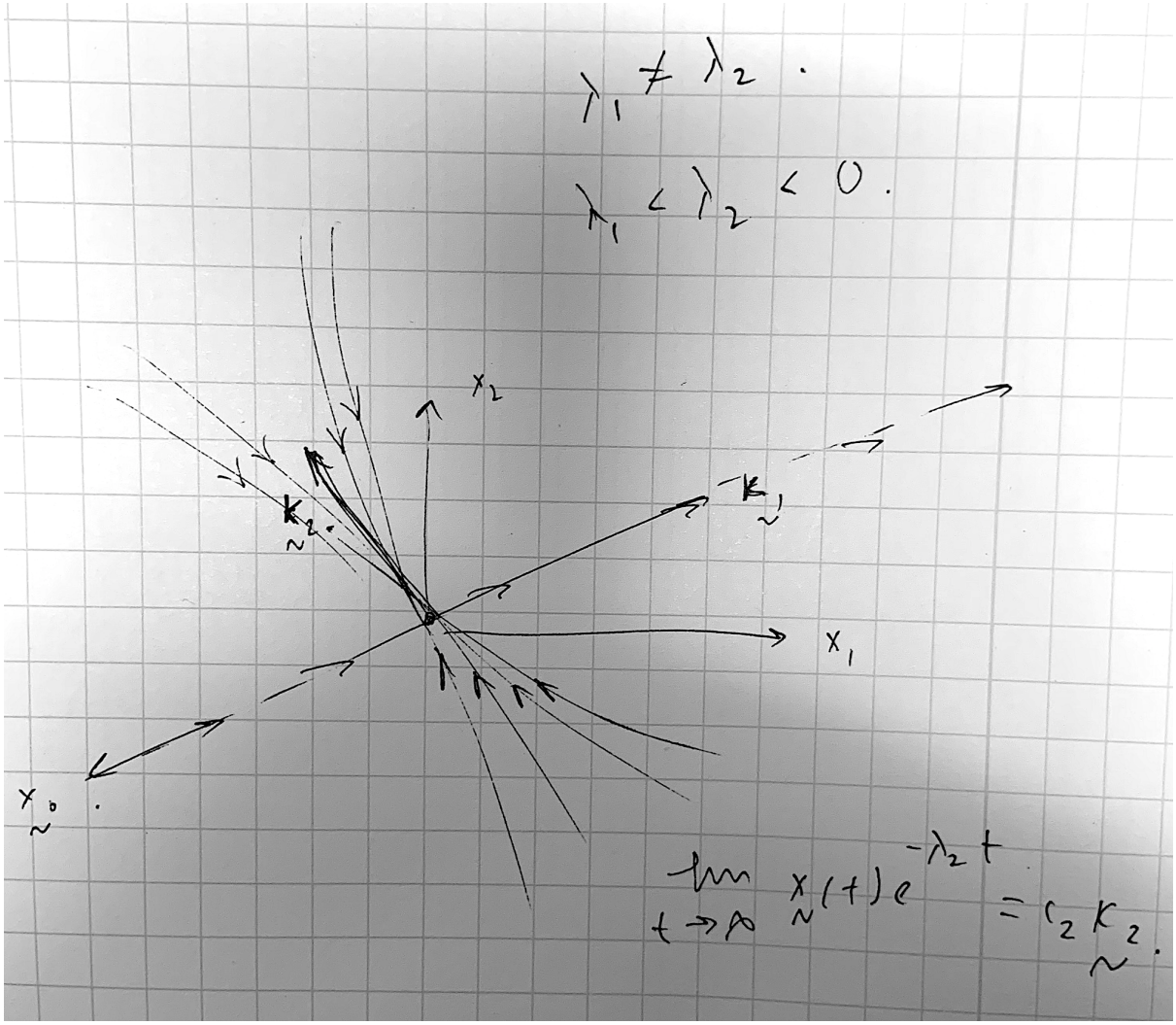

Figure 4: $\lambda_1 < \lambda_2 < 0, \quad \lim_{t\to\infty} \mathbf{x}(t)e^{-\lambda_2 t} = c_2\mathbf{k}_2$

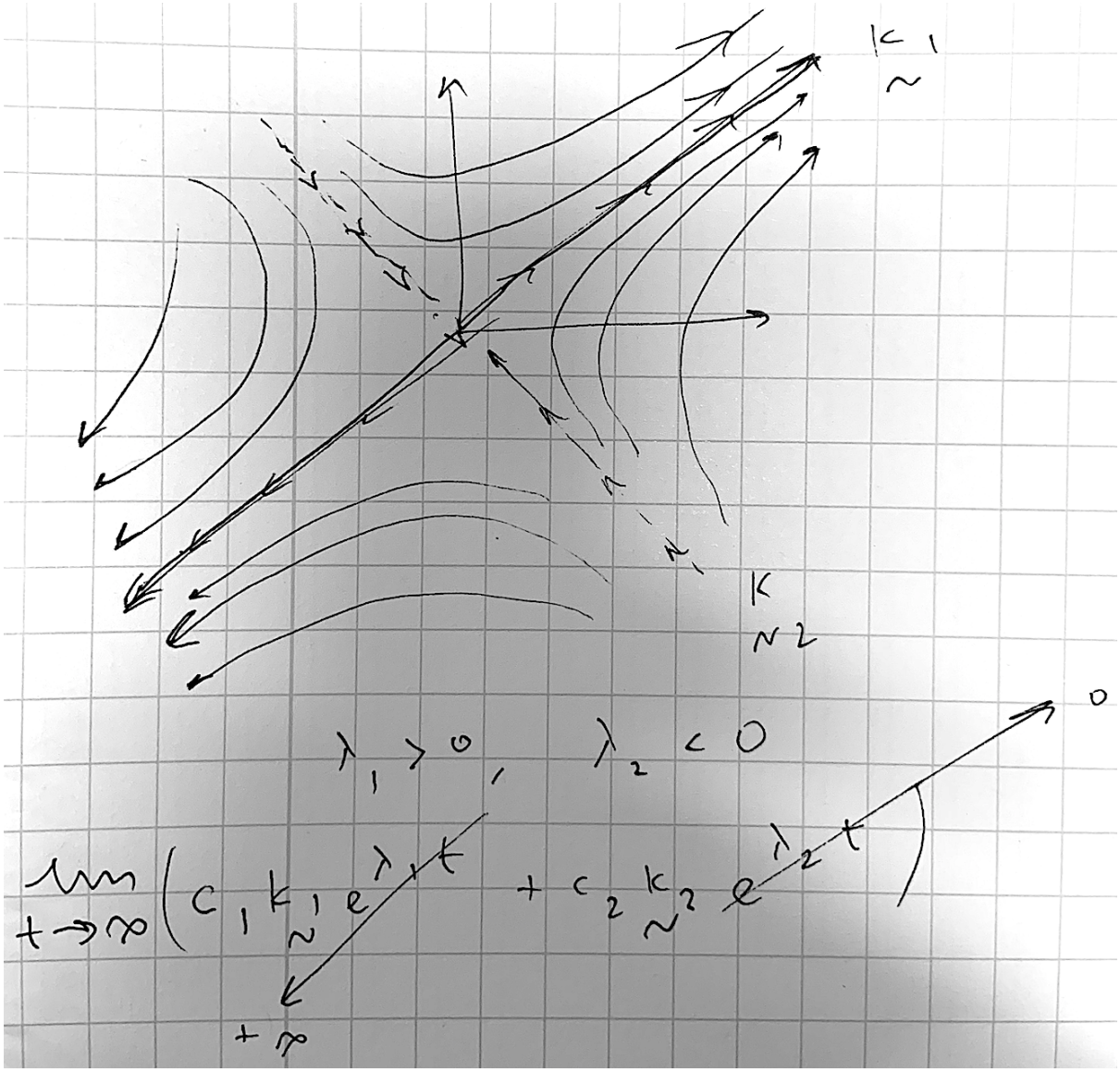

Figure 5: $\lambda_1 > 0, \quad \lambda_2 < 0$

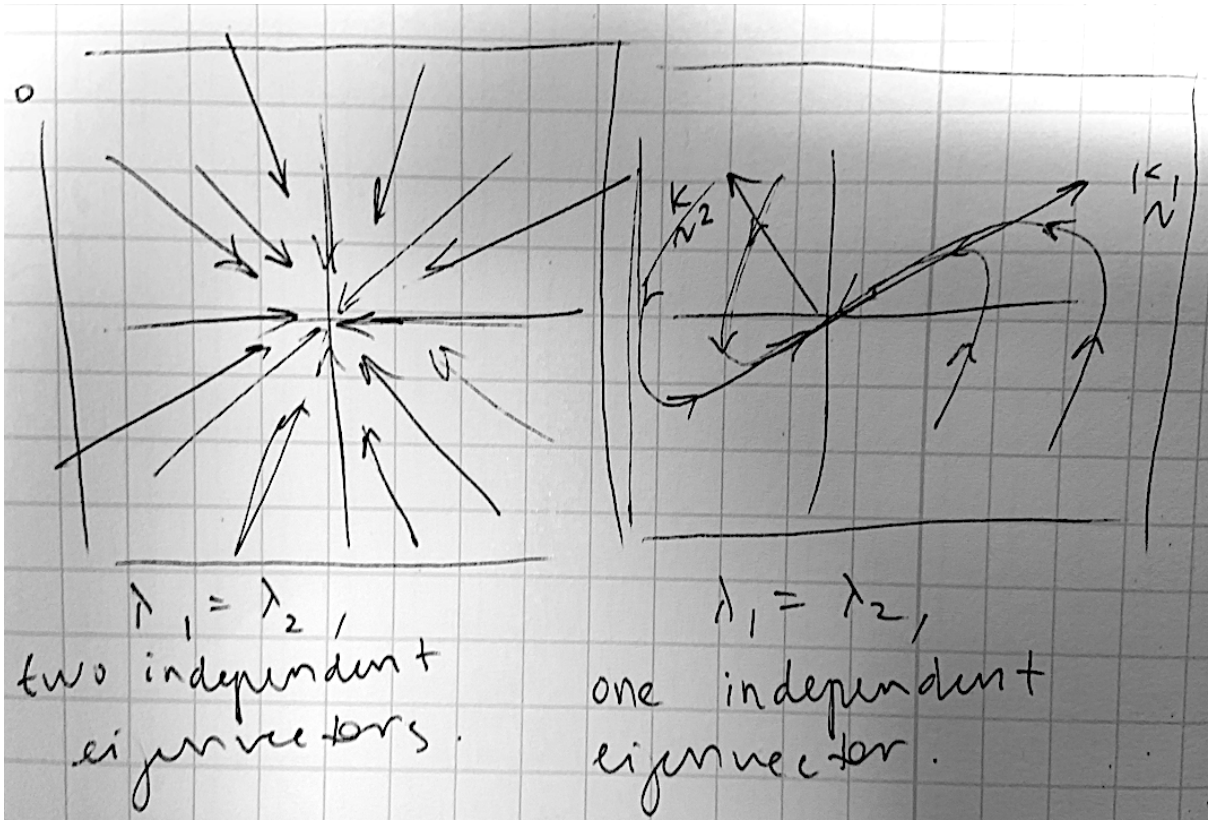

Figure 6: $\lambda_1 = \lambda_2 = \lambda < 0,$ one independent eigenvector

if (IIIii) we have only one independent eigenvector, then the solution is derived through variation of parameters as in Eq. 401 and is written as

$$\lambda_1 = \lambda_2 = \lambda, \quad \mathbf{x}(t) = c_1\mathbf{k}_1 e^{\lambda t} + \underbrace{c_2(\mathbf{k}_1 t e^{\lambda t}}_{\text{dominates, } t\to\infty} + \mathbf{k}_2 e^{\lambda t}). \tag{451}$$

In this case one of the $\mathbf{k}_1$ terms dominate and so all trajectories will tend toward $\mathbf{k}_1$. Fig. 6 portrays the case of $\lambda < 0$ where the system is stable and trajectories tend toward the origin. Conversely though if $\lambda > 0$, the trajectories would tend toward infinitely far from the origin and so the system would be classified as unstable.

Now let us consider the case (IV) where the eigenvalues are complex with a real part. That is,

$$\lambda = \alpha \pm i\beta. \tag{452}$$

If the eigenvalues are complex then they will always be complex conjugate pairs. In this case, in the phase space there exists a spiral pattern where the general solution has some $e^{\alpha t}e^{i\beta t}$ term. The critical point is the spiral point or the center of the spiral. If $\alpha < 0$, the trajectories spiral inward towards the critical point and the system is stable; if $\alpha > 0$, the trajectories spiral outward towards it and the system is unstable.
In the special case (V) of $\alpha = 0$, so that the eigenvalues are strictly complex with no real part, i.e.

$$\lambda = \pm i\beta, \tag{453}$$

then the spiral will neither tend inward nor outward. Actually, the trajectories do not comprise a spiral but a set of closed loops. The system is then stable because the trajectories do not tend to infinity. The equilibrium point is called a center.
We have considered all possible classifications of the critical points, which are: improper node, proper node, saddle point, spiral point, or center.
We have let the critical point be the origin but this does not have to be the case. Let us generalize and call the critical point $\mathbf{x}^*$. As we have seen, the stability of the critical point has to do with how trajectories behave around $\mathbf{x}^*$. If you begin at $t = 0$ within some circle of radius δ centered at $\mathbf{x}^*$, then the trajectory must remain within some circle of radius ϵ for all t. If we can identify some circle then the solution is stable. If we cannot identify any circle - that is, if the trajectories tend toward infinity - then the solution is unstable.
If in addition to satisfying stability,

$$\lim_{t\to\infty} \mathbf{x} = \mathbf{x}^*, \tag{454}$$

then the critical point is called asymptotically stable.
Let us reconsider the damped spring mass system in this context. The same as in Eq. 342 and Eq. 350, it is

$$m\frac{d^2u}{dt^2} + c\frac{du}{dt} + ku = 0 \tag{455}$$

where

$$\omega^2 = k/m, \quad 2\xi\omega = c/m, \tag{456}$$

so that

$$\ddot{u} + 2\xi\omega\dot{u} + \omega^2 u = 0. \tag{457}$$

At the global level, the sign and magnitude of $2\xi\omega$ and ω^2 affects the stability. One can draw a stability diagram with axes $c/m = 2\xi\omega$ and $k/m = \omega^2$ a line drawn where the solution crosses from stability to instability. Static stability in this case is associated with a negative stiffness $k < 0$; dynamic instability is associated with a negative damping $c < 0$. The sensitivity of system behavior to parameter changes is noteworthy. In this case, for example, small changes in ξ influentially pushes the system from stable to asymptotically stable to unstable.

2.12 Lec 2l Autonomous systems

Suppose

$$\frac{dx}{dt} = F(x, y), \quad x(t_0) = x_0, \tag{458}$$

$$\frac{dy}{dt} = G(x, y), \quad y(t_0) = y_0, \tag{459}$$

where F and G are arbitrary functions. Of course we can summarize this to say that

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x}(t_0) = \mathbf{x}_0. \tag{460}$$

This is an autonomous system, which means the independent variable t does not appear anywhere in explicit form in the equation. That is, F and G do not have a direct contribution from the variable t. Note that for constant $\mathbf{A}$, $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$ is autonomous.
Autonomous systems have restrictions. This simplifies the analysis. That is because there is only one trajectory which passes through a point in phase space. In this case the critical points are exactly the $\mathbf{x}$ where

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}. \tag{461}$$

There exist some nonlinear systems that are called an "almost linear system" (ALS). Suppose

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}). \tag{462}$$

Let us find trajectories near the critical point $\mathbf{x}^* = \mathbf{0}$. Although this is assumed to be the origin it does not have to be. If it is not the origin then one can shift the coordinates by introducing the translation $\mathbf{u} = \mathbf{x} - \mathbf{x}^*$.
Now, if (I) it is true that we can rewrite Eq. 462 as

$$\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x} + \mathbf{g}, \tag{463}$$

and if (II) $\mathbf{x}^*$ is an isolated critical point of $\frac{d\mathbf{x}}{dt} = \mathbf{A}\mathbf{x} + \mathbf{g}$, and if (III) $\det \mathbf{A} \neq 0$ (so that $\mathbf{x} = \mathbf{0}$ is the only critical point of $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x}$), and if (IV) $\frac{||\mathbf{g}||}{||\mathbf{x}||} \to 0$ as $\mathbf{x} \to 0$, then the system is "almost linear." In (II), an isolated critical point means that one can draw a circle around $\mathbf{x} = \mathbf{0}$ with no other critical points in that circle. Also, the nonlinear terms that are $\mathbf{g}$ should become smaller and smaller as we approach the critical point. That is basically (IV).
For ALS's, one can say something about the trajectories near the critical point. The behavior of the ALS near the critical point is the same as that of corresponding linear system, *except* for (I) centers, where the eigenvalues are imaginary, and (2) nodes, where the eigenvalues are equal. For (I), a center in the linear system can become a spiral in the ALS. The stability also is compromised. For (II), the node in the linear system to a spiral point in the ALS. However, the stability does not change.
A perfect nonlinear example is the damped pendulum that is Eq. 364, or

$$\ddot{\theta} + \frac{c}{m}\dot{\theta} + \frac{g}{L}\sin\theta = 0. \tag{464}$$

We maintain the claim that is Eq. 354, or

$$\sum M_{0_z} = \dot{H}_{0_z}. \tag{465}$$

We can convert this to a first order system by letting

$$x \leftarrow \theta, y \leftarrow \dot{\theta}, \tag{466}$$

making

$$\dot{x} = y, \tag{467}$$

$$\dot{y} + \frac{c}{m}y + \frac{g}{L}\sin x = 0 \quad \longrightarrow \quad \dot{y} = -\frac{c}{m}y - \frac{g}{L}\sin x. \tag{468}$$

The critical points are

$$0 = \dot{x} = y, \tag{469}$$

$$0 = \dot{y} = -\frac{c}{m}y - \frac{g}{L}\sin x. \tag{470}$$

Therefore

$$y = 0, \;\; \sin x = 0 \;\; \rightarrow \;\; x = \pm n\pi, \;\; n = 0,\ 1,\ 2,\ \ldots\,. \tag{471}$$

The physical meaning of $x = \pm n\pi$ is related to the pendulum diagram. if $x = 0$, the pendulum hangs straight down; if $x = \pi$, the pendulum shoots straight up; if $x = 2\pi$, the pendulum has traversed all the way to its original position and hangs straight down; etc. These are the equilibrium positions. However, this does not necessarily imply stability. For instance, if $x = \pi$ and the mass is directly above the support (really any n odd), then this is not a stable position. But, if $x = 0$ and the mass is directly below the support (really any n even), then this is a stable position.

Given a nonlinear system we attempt to convert it to an ALS. For the pendulum, let

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots\,. \tag{472}$$

Still

$$\dot{x} = y \tag{473}$$

but now, substituting Eq. 472 into Eq. 468,

$$\dot{y} = -\frac{c}{m}y - \frac{g}{L}(x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots). \tag{474}$$

The conditions of ALS require some equation in the form of

$$\mathbf{A}\begin{Bmatrix} x \\ y \end{Bmatrix} + \mathbf{g} = \mathbf{A}\mathbf{x} + \mathbf{g} = \dot{\mathbf{x}} = \begin{Bmatrix} \dot{x} \\ \dot{y} \end{Bmatrix}. \tag{475}$$

In Eq. 474 we can partition the part of the equation dependent only on x and y from the nonlinear part. That is,

$$\dot{x} = \begin{Bmatrix} 0 & 1 \end{Bmatrix}\begin{Bmatrix} x \\ y \end{Bmatrix} + 0, \tag{476}$$

$$\dot{y} = -\frac{g}{L}x - \frac{c}{m}y + \frac{g}{L}(\frac{x^3}{3!} - \frac{x^5}{5!} + \ldots) \tag{477}$$

$$\rightarrow \dot{y} = \begin{Bmatrix} -\frac{g}{L} & -\frac{c}{m} \end{Bmatrix}\begin{Bmatrix} x \\ y \end{Bmatrix} + \frac{g}{L}(\frac{x^3}{3!} - \frac{x^5}{5!} + \ldots). \tag{478}$$

Joined together as a system,

$$\underbrace{\begin{Bmatrix} \dot{x} \\ \dot{y} \end{Bmatrix}}_{\dot{\mathbf{x}}} = \underbrace{\begin{bmatrix} 0 & 1 \\ -\frac{g}{L} & -\frac{c}{m} \end{bmatrix}}_{\mathbf{A}}\underbrace{\begin{Bmatrix} x \\ y \end{Bmatrix}}_{\mathbf{x}} + \underbrace{\begin{Bmatrix} 0 \\ \frac{g}{L}(\frac{x^3}{3!} - \frac{x^5}{5!} + \ldots) \end{Bmatrix}}_{\mathbf{g}}. \tag{479}$$

To confirm condition (IV), which is $\frac{||\mathbf{g}||}{||\mathbf{x}||} \rightarrow 0$ as $\mathbf{x} \rightarrow \mathbf{0}$, we graph y against g_y to see that $y \gg g_y$ near $\mathbf{0}$. Therefore the condition is satisfied.

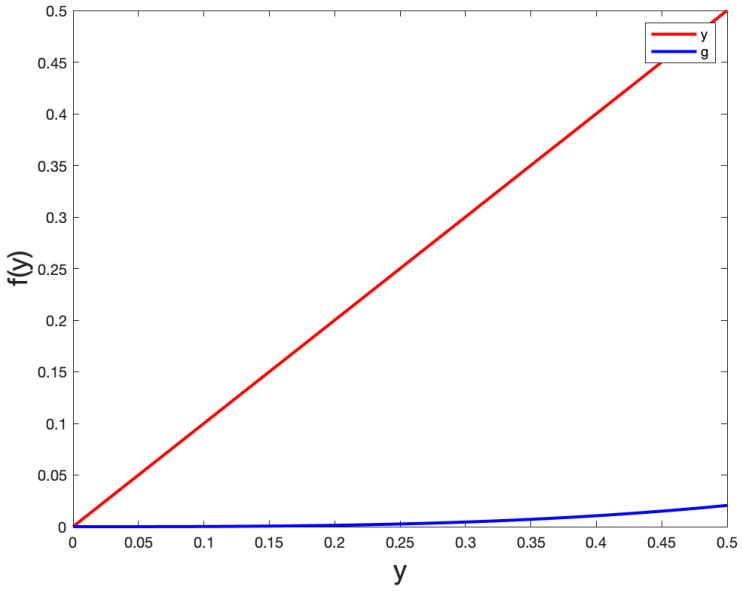

Figure 7: Checking (IV) of ALS: $\frac{||\mathbf{g}||}{||\mathbf{x}||} \to 0$ as $\mathbf{x} \to \mathbf{0}$

```
clear
clc

x = 0:0.01:0.5
g = 0

numiters=10
for iter = 1:numiters
    iter
    g = g + (-1)^(iter-1) * x.^(2*iter+1)/(factorial(2*iter+1))
end

a = plot(x,x)
hold on
b = plot(x,g)
title('')
xlabel('y','FontSize',18)
ylabel('f(y)','FontSize',18)
a.LineWidth = 2
a.Color = 'red'
b.LineWidth = 2
b.Color = 'blue'
legend('y', 'g')
```

Shifting the origin by 2π does not change the behavior of the solution. However, let us

shift the origin by pi on x. We let

$$u = x - \pi \leftarrow u + \pi = x, \quad v = y - 0 \leftarrow v + 0 = y. \tag{480}$$

Then

$$\dot{u} = v, \quad \dot{v} = -\frac{c}{m}v - \frac{g}{L}\sin(u + \pi) = -\frac{c}{m}v + \frac{g}{L}\sin(u). \tag{481}$$

This fosters a different behavior. Whereas the origin at $0, 2\pi$ for $\xi < 1$ exhibits a stable spiral, the origin at π for $\xi < 1$ exhibits an unstable saddle. Then the complete phase portrait of the pendulum is a continuous smooth transition between adjacent saddles and spirals.

2.13 Lec 2m Nonlinear ODEs

For many autonomous linear systems a single critical point $\mathbf{x}^* = \mathbf{0}$ is asymptotically stable. That is, the trajectory through any point tends eventually to $\mathbf{x}^*$ as $t \to \infty$. About this we say that the basin of attraction is the entire phase space. The basin of attraction is essentially the set of points which are attracted to the critical point ("the attractor").
For other, more general nonlinear systems though, there can be other attractors besides just the critical points. At the next level of complexity there is not just one solution but there are periodic solutions at every period T. That is,

$$\mathbf{x}(t + T) = \mathbf{x}(t). \tag{482}$$

An example of such a system is the nonlinear second order Van der Pol oscillator

$$\ddot{u} + \mu(u^2 - 1)\dot{u} + u = 0 \tag{483}$$

where μ is some constant. In an analogy to the pendulum, the $\mu(u^2 - 1)$ term can be thought of as a damp proportional to the velocity $\dot{u}$. The u^2 term implies nonlinearity. If in this system $\mu = 0$ then there is no damp and the equation reduces to

$$\ddot{u} + u = 0 \tag{484}$$

where the solution

$$u(t) = c_1 e^{it} + c_2 e^{-it} \tag{485}$$

has strictly complex roots. In this case therefore the solution trajectory is a circle about the origin. That is to say $\mathbf{x} = \mathbf{0}$ is the only critical point.
Now if $\mu >\neq 0$ then there exists nonlinearity. If (I) $\mu > 0$ then there is damping with coefficient $\mu(u^2 - 1)$. If (Ii) $u > 1$, then $(u^2 - 1)$ is positive and so the entire coefficient is positive. A positive damp will diminish the system by way of an inward spiral. However if (Iii) $u < 1$ then the coefficient is negative and the solution trajectory will grow through an outward spiral.
Converting to phase space form, we let

$$\dot{x} = y, \quad \dot{y} = -\mu(x^2 - 1)y - x. \tag{486}$$

Let us examine the linear portion, which is done by approximating the nonlinear portion x^2 to zero, i.e. $x^2 \approx 0$, and let us do so near the critical point. This is

$$\dot{x} = y, \quad \dot{y} = \mu y - x, \tag{487}$$

or

$$\begin{Bmatrix} \dot{x} \\ \dot{y} \end{Bmatrix} = \begin{bmatrix} 0 & 1 \\ -1 & \mu \end{bmatrix} \begin{Bmatrix} x \\ y \end{Bmatrix} = \mathbf{A}\mathbf{x}. \tag{488}$$

The eigenvalues of $\mathbf{A}$ are $\lambda = \mu \pm \frac{\sqrt{\mu^2-4}}{2}$. Because the sign of the eigenvalues affects the state space trajectory it is then valuable to partition the system behavior into the cases (I) $\mu^2 \geq 4$ and $\mu^2 < 4$.

Another nonlinear example, one more complicated, is the Lorenz model. It is based on the Rayleigh Bérnard problem where a fluid layer is heated from below such that the temperature at the lower layer T_l is greater than that of the upper layer T_u. So $T_l > T_u$, but if these two are approximately close, then there is simple heat transfer from the hotter end to the colder end. However, if there is a large difference such that $T_l \gg T_u$, then there is a stability issue. If there is a critical temperature difference ΔT_{cr} and if

$$T_l - T_u > \Delta T_{cr}, \tag{489}$$

then fluid motion begins due to buoyancy. Heating the bottom fluid, it attempts to rise. The colder fluid on the top begins to sink. This transfer of position occurs cyclically because as the hotter fluid goes up, it is no longer being heated, becomes cooler, and then goes back down.

The system must satisfy mass balance, which is

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} = 0, \tag{490}$$

the balance of momentum, which is the set of Navier-Stokes equations, or Unfinished

3 Mod3 Fourier analysis and integral transforms

3.1 Lec 3a Fourier series

In Sec 2 we explored power series and Frobenius series. Now we look at trigonometric series to represent periodic functions. A periodic function is defined by

$$f(t+T) = f(t+T+T) = \ldots = f(t+nT) = f(t) \tag{491}$$

with period T for all t. The smallest period is the fundamental period. In this example that is T. That is to say $2T$ is not the fundamental period. Periodic functions include

$$f(t) = C \tag{492}$$

where C is a constant, and where f has no fundamental period;

$$g(t) = \cos \omega t = \cos(\omega t + 2\pi) = \cos(\omega t + 4\pi) = \ldots = \cos(\omega t + 2n\pi) \tag{493}$$

where $n = 1, 2, \ldots$ and where g has fundamental period

$$T = \frac{2\pi}{\omega} \tag{494}$$

as

$$g = \cos\Big(\omega t\Big) = \cos\Big(\omega \underbrace{(t + \frac{2\pi}{\omega})}_{t+T}\Big) = \cos(\omega t + 2\pi). \tag{495}$$

These first two examples are continuous functions, but even functions with discontinuous slopes or completely discontinuous functions can be periodic still. However it is easier to work with smooth functions.

Suppose the function $f(t)$ has period 2π. Let us write this particular periodic function f as a sum of periodic functions, such that

$$f(t) = \frac{a_0}{2} + \sum_n (a_n \cos nt + b_n \sin nt) = \frac{a_0}{2} + a_n \cos t + b_n \sin t + a_n \cos 2t + b_n \sin 2t + \ldots). \tag{496}$$

The fundamental period of each term on the right hand side of Eq. 496 is

$$T = \frac{2\pi}{n} \tag{497}$$

because

$$a_n \cos nt = a_n \cos n(t + \frac{2\pi}{n}) = a_n \cos(nt + 2\pi) \tag{498}$$

and the same is true for $b_n \sin \omega t$. The goal is to find the coefficients $a_n = a_1, a_2, \ldots,$ $b_n = b_1, b_2, \ldots$. First of all, integrating Eq. 496 through a single period with bounds $-\pi, \pi$,

$$\int_{-\pi}^{\pi} f(t)dt = \int_{-\pi}^{\pi} \frac{a_0}{2} dt + \int_{-\pi}^{\pi} \Big(\sum_n (a_n \cos nt + b_n \sin nt) \Big) dt. \tag{499}$$

Now, assuming the series converges uniformly, the summation and integration in Eq. 499 can be interchanged so that

$$\begin{aligned}
\int_{-\pi}^{\pi} f(t)dt &= \int_{-\pi}^{\pi} \frac{a_0}{2} dt + \sum_n \int_{-\pi}^{\pi} (a_n \cos nt + b_n \sin nt)dt \\
&= \frac{a_0}{2}(\pi + \pi) + \sum_n \int_{-\pi}^{\pi} (a_n \cos nt + b_n \sin nt)dt \\
&= a_0\pi + \sum_n (a_n \frac{1}{n} \sin nt \Big|_{-\pi}^{\pi} - b_n \frac{1}{n} \cos nt \Big|_{-\pi}^{\pi}) \\
&= a_0\pi = \int_{-\pi}^{\pi} f(t)dt \Longrightarrow a_0 = \frac{1}{\pi} \int_{-\pi}^{\pi} f(t)dt.
\end{aligned} \tag{500}$$

Here because of the nature of the trigonometric functions, there exists just as much in the positive regime as in the negative regime, so the entire second term cancels.

a_0 is the first term in the series. To obtain the remaining terms, multiply Eq. 499 by $\cos mt$ so that

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos mtdt + \int_{-\pi}^{\pi}\left(\sum_n (a_n \cos nt + b_n \sin nt)\right)\cos mtdt. \quad (501)$$

Again assuming uniform convergence,

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos mtdt + \sum_n \int_{-\pi}^{\pi} (a_n \cos nt\cos mt + b_n \sin nt\cos mt)dt \quad (502)$$

Trigonometric identities state

$$\cos nt\cos mt = \frac{1}{2}(\cos(n+m)t + \cos(n-m)t)$$

$$\sin nt\cos mt = \frac{1}{2}(\sin(n+m)t + \sin(n-m)t). \quad (503)$$

Substituting Eq. 503 into the right hand side of Eq. 502,

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos mtdt$$
$$+\sum_n \Bigg(\underbrace{\int_{-\pi}^{\pi} \frac{a_n}{2}\cos((n+m)t)dt}_{\mathbf{I.}} + \underbrace{\int_{-\pi}^{\pi} \frac{a_n}{2}\cos((n-m)t)dt}_{\mathbf{II.}}$$
$$+\underbrace{\int_{-\pi}^{\pi} \frac{b_n}{2}\sin((n+m)t)dt}_{\mathbf{III.}} + \underbrace{\int_{-\pi}^{\pi} \frac{b_n}{2}\sin((n-m)t)dt}_{\mathbf{IV.}}\Bigg). \quad (504)$$

Considering Eq. 504: if $n \neq m$ so that $n+m=\xi, n-m=\eta$,

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos mtdt$$
$$+\sum_n \Bigg(\underbrace{\int_{-\pi}^{\pi} \frac{a_n}{2}\cos(\xi t)dt}_{\mathbf{I.}} + \underbrace{\int_{-\pi}^{\pi} \frac{a_n}{2}\cos(\eta t)dt}_{\mathbf{II.}}$$
$$+\underbrace{\int_{-\pi}^{\pi} \frac{b_n}{2}\sin(\xi t)dt}_{\mathbf{III.}} + \underbrace{\int_{-\pi}^{\pi} \frac{b_n}{2}\sin(\eta t)dt}_{\mathbf{IV.}}\Bigg) = 0+0+0+0+0 \quad (505)$$

again because the curves oscillate evenly between the positive and negative number space. Now, if $n=m$,

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = \int_{-\pi}^{\pi} \frac{a_0}{2}\cos mtdt$$

$$+\sum_m \Big(\underbrace{\int_{-\pi}^{\pi} \frac{a_m}{2}\cos(2mt)dt}_{\mathbf{I.}} + \underbrace{\int_{-\pi}^{\pi} \frac{a_m}{2}\cos(0t)dt}_{\mathbf{II.}}$$

$$+ \underbrace{\int_{-\pi}^{\pi} \frac{b_m}{2}\sin(2mt)dt}_{\mathbf{III.}} + \underbrace{\int_{-\pi}^{\pi} \frac{b_m}{2}\sin(0t)dt}_{\mathbf{IV.}} \Big) = 0 + 0 + a_m\pi + 0 + 0 \tag{506}$$

since in **II**, $\cos 0 = 1$, so $\int_{-\pi}^{\pi} dt = 2\pi$. Therefore,

$$\int_{-\pi}^{\pi} f(t)\cos mtdt = a_m\pi$$

$$\Longrightarrow a_m = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\cos mtdt. \tag{507}$$

To find b_n, the same procedure that developed Eq. 501 could be done for $\sin mt$. That is, multiply 499 by $\sin mt$. Then,

$$b_m = \frac{1}{\pi}\int_{-\pi}^{\pi} f(t)\sin mtdt. \tag{508}$$

Eqs. 507 and 508 are so called Euler formulas. These are all the coefficients in Eq. 496 ($f(t) = \frac{a_0}{2} + \sum_n (a_n \cos nt + b_n \sin nt)$), which is called the Fourier series corresponding to f. a_m, b_m are called the Fourier coefficients of f.

Now, recall that f has period 2π. However, we want to generalize the Fourier series for any T. So, instead of integrating from $-\pi$ to π and multiplying by

$$\cos mt = \cos m\frac{\pi}{\pi}t,\ \sin mt = \sin m\frac{\pi}{\pi}t,$$

we integrate from $-p$ to p and multiply by

$$\cos m\frac{\pi}{p}t,\ \sin m\frac{\pi}{p}t.$$

Then, The Fourier series representation of any f with period $2p$ is

$$f(t) = f(t+nT) = \frac{a_0}{2} + \sum_n (a_n \cos\frac{n\pi t}{p} + b_n \sin\frac{n\pi t}{p}), \tag{509}$$

where

$$a_n = \frac{1}{p}\int_{-p}^{p} f(t)\cos\frac{n\pi t}{p}dt, \tag{510}$$

$$b_n = \frac{1}{p}\int_{-p}^{p} f(t)\sin\frac{n\pi t}{p}dt, \tag{511}$$

$$T = 2p. \tag{512}$$

3.2 Lec 3b Orthogonality

An important part of the proof in Lec 3.1 was that trigonometric functions sin and cos were symmetric over the period $-\pi$ to π. This is called the property of orthogonality. Formally, in continuous form, if a set of functions $\phi_i = \phi_i(x)$ has the property

$$\int_a^b \phi_m \phi_n dx \quad \begin{cases} = 0, & m \neq n \\ \neq 0, & m = n, \end{cases} \tag{513}$$

then we say these functions form an orthogonal set. Moreover, if

$$\int_a^b \phi_m^2 dx = 1, \tag{514}$$

then this set is orthonormal. Then we can generalize and say

$$\int_a^b \phi_m \phi_n dx = \delta_{mn}. \tag{515}$$

It is easy to convert any orthogonal set into an orthonormal one by normalizing each ϕ_i.

In discrete form, an orthogonal set

$$\boldsymbol{\phi}_m^T \boldsymbol{\phi}_n \quad \begin{cases} = 0, & m \neq n \\ \neq 0, & m = n. \end{cases} \tag{516}$$

An orthonormal set

$$\boldsymbol{\phi}_m^T \boldsymbol{\phi}_n = 1. \tag{517}$$

This means

$$\boldsymbol{\phi}_m \cdot \boldsymbol{\phi}_n = \delta_{mn}. \tag{518}$$

Take for example the Cartesian coordinate system $\mathbf{e}_i$, which is orthonormal. We know

$$\mathbf{e}_1^T \mathbf{e}_1 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = 1 + 0 + 0 = 1; \quad \mathbf{e}_3^T \mathbf{e}_2 = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = 0. \tag{519}$$

A set of vectors is orthogonal with respect to the matrix $\mathbf{A}$ if

$$\boldsymbol{\phi}_m^T \mathbf{A} \boldsymbol{\phi}_n \quad \begin{cases} = 0, & m \neq n \\ \neq 0, & m = n. \end{cases} \tag{520}$$

Now, recall that to find Fourier coefficients a_n or b_n of f with period $T = (-p, p)$, we integrate Eq. 509, which is

$$f(t) = f(t + nT) = \frac{a_0}{2} + \sum_n (a_n \cos \frac{n\pi t}{p} + b_n \sin \frac{n\pi t}{p}),$$

over $(-p, p)$ and multiply by $\cos(m\pi t/p)$ to find a_m or $\sin(m\pi t/p)$ to find b_m. We then discussed that the trigonometric functions are symmetrical (orthogonal) over that period

because they oscillate back and forth between the positive and negative regimes. However, we discussed that if $n = m$, then the trigonometric identities Eq. 503 transform one of the $\int_{-p}^{p} \cos(n-m)dt$ terms into $\frac{1}{2}\int_{-p}^{p} 1dt = p$. Altogether,

$$\int_{-p}^{p} \cos\left(\frac{m\pi t}{p}\right)\cos\left(\frac{n\pi t}{p}\right)dt = \begin{cases} 0, & m \neq n \\ p, & m = n. \end{cases} \tag{521}$$

In a completely analogous way, we also find this to be true of $\sin(*)$, in that

$$\int_{-p}^{p} \sin\left(\frac{m\pi t}{p}\right)\sin\left(\frac{n\pi t}{p}\right)dt = \begin{cases} 0, & m \neq n \\ p, & m = n. \end{cases} \tag{522}$$

However, for the product $\sin(*)\cos(*)$,

$$\int_{-p}^{p} \sin\left(\frac{m\pi t}{p}\right)\cos\left(\frac{n\pi t}{p}\right)dt = 0. \tag{523}$$

Again this is because of the trigonometric identity Eq. 503 ($\sin nt \cos mt = \frac{1}{2}(\sin(n+m)t + \sin(n-m)t)$). Then integration is possible. Again, these ideas helped us form the Euler formulas for the coefficients of the Fourier series of f, i.e., Eqs. 509-511.

3.3 Lec 3c Dirichlet conditions

For a function f to satisfy the Dirichlet (DEER-ish-lay) conditions,

- f must be bounded;
- f must be periodic;
- f cannot have infinite local minima and maxima at one period; and
- f cannot be discontinuous at infinite points within one period.

If f satisfies the Dirichlet conditions, then

- The Fourier series of f converges to f wherever f is continuous; and
- Wherever f is discontinuous, its Fourier series converges to the average of its right and left hand limits.

This is all true for Fourier series representation of f with period $T = 2p$ (Eqs. 509-511),

$$f(t) = f(t+nT) = \frac{a_0}{2} + \sum_n (a_n \cos\frac{n\pi t}{p} + b_n \sin\frac{n\pi t}{p}), \tag{524}$$

where

$$a_n = \frac{1}{p}\int_{-p}^{p} f(t)\cos\frac{n\pi t}{p}dt,$$

$$b_n = \frac{1}{p}\int_{-p}^{p} f(t)\sin\frac{n\pi t}{p}dt.$$

As an example, consider a function with period $T = 2\pi$

$$f(x) = \begin{cases} -k, & -\pi < x < 0, \\ k, & 0 < x < \pi. \end{cases} \tag{525}$$

Notice that the function is discontinuous at $x = n\pi$. However, the function is only discontinuous at a finite number of points within a single period. So, the function satisfies the Dirichlet conditions. Of course, this is also a periodic function. To solve, we split f into its two intervals, making

$$\begin{aligned} a_n &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos\frac{n\pi x}{\pi}dx \\ &= \frac{1}{\pi}\left(\int_{-\pi}^{0} -k\cos(nx)dx + \int_{0}^{\pi} k\cos(nx)dx\right) \\ &= \frac{-k}{n\pi}\sin(nx)\Big|_{-\pi}^{0} + \frac{k}{n\pi}\sin(nx)\Big|_{0}^{\pi} = 0 + 0 = 0 \end{aligned} \tag{526}$$

and

$$\begin{aligned} b_n &= \frac{1}{\pi}\left(\int_{-\pi}^{0} -k\sin(nx)dx + \int_{0}^{\pi} k\sin(nx)dx\right) \\ &= \frac{k}{n\pi}\cos(nx)\Big|_{-\pi}^{0} - \frac{k}{n\pi}\cos(nx)\Big|_{0}^{\pi} \\ &= \frac{k}{n\pi}(\cos 0 - \underbrace{\cos(-n\pi) - \cos(n\pi)}_{\cos n\pi = \cos -n\pi} + \cos(0)) = \frac{2k}{n\pi}(1 - \underbrace{\cos n\pi}_{\cos n\pi = \cos -n\pi}) = b_n. \end{aligned} \tag{527}$$

For b_n, if n is even so that $(1 - \cos 0\pi) = (1 - \cos 2\pi) = \ldots = 0$, then $b_n = 0$. But, if n is odd so that $(1 - \cos\pi) = (1 - \cos 3\pi) = \ldots = 2$, then $b_n = 2(2k/n\pi)$. Altogether,

$$b_n = \begin{cases} 4k/n\pi, & n \text{ odd}, \\ 0, & n \text{ even}. \end{cases} \tag{528}$$

Then, from 524, the Fourier series

$$f(x) = \frac{4k}{\pi}\sum_{n \text{ odd}} \frac{\sin nx}{n}. \tag{529}$$

Because of the Dirichlet theorem, we know that at points of discontinuity, the Fourier series converges to the average between the left and right limits, which is zero, as those limits are k and $-k$. This is shown in Fig. 8.

As another example, which currently needs further explanation, consider

$$f(x) = \begin{cases} x, & -2 < x \le 0, \\ x, & 0 < x \le 2, \end{cases} \qquad T = 4. \tag{530}$$

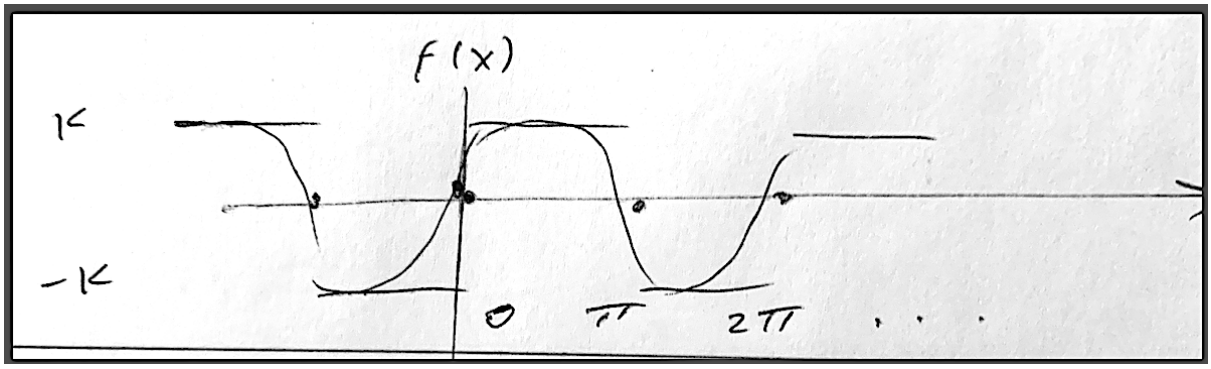

Figure 8: $f(x) = \frac{4k}{\pi} \sum_{n \text{ odd}} \frac{\sin nx}{n}$.

This is continuous everywhere, and is periodic, so, certainly it satisfies the Dirichlet conditions. Then

$$a_n = \frac{1}{2}\int_{-2}^{0} -x \cos\frac{n\pi x}{2} dx + \frac{1}{2}\int_{0}^{2} x \cos\frac{n\pi x}{2} dx. \tag{531}$$

Using integration by parts

$$\int udv = uv - \int vdu; \quad u = x, \quad dv = \cos\frac{n\pi}{2} x dx,$$

we obtain

$$a_n = \frac{4}{n^2\pi}(\cos n\pi - 1) = \begin{cases} -\frac{8}{n^2\pi^2}, & n \text{ odd}, \\ 0, & n \text{ even}. \end{cases} \tag{532}$$

Using the analogous integration by parts technique on b_n, but for $\sin(*)$, we find

$$b_n = 0 \quad \forall \quad n. \tag{533}$$

Then

$$f(x) = \sum_{n \text{ odd}} (-\frac{8}{n^2\pi^2} \cos\frac{n\pi x}{2}). \tag{534}$$

Notice that in Eq. 534 there are only cos terms. However, in the result derived from the first example Eq. 529, there are only sin terms. This is predictable based on inspection: The function $\cos\omega t$ is itself even, as it is symmetrical along the y axis, and the function $\sin\omega t$ is odd, as it is symmetrical along the line $y = x$. This extrapolates to more complex functions: if f is even, then the cos terms will prevail; if f is odd, then the sin terms will prevail. This is obvious thinking about it geometrically. If you are taking the integral along the interval $(-p, p)$, and the function is even, that is, symmetric along the y axis, then the area under the curve of the left half is exactly that of the right half because they are mirror images. That is,

$$\int_{-p}^{p} g(x)dx = 2\int_{0}^{p} g(x)dx, \quad g \text{ even}. \tag{535}$$

On the other hand, taking the same integral but of an odd function, the area under the curve of the left half will be the exact inverse (negative) of that of the right half, because of the axis along which the two halves are symmetrical. That is,

$$\int_{-p}^{p} h(x)dx = 0, \quad h \text{ odd.} \tag{536}$$

Product rules for even g and odd h are

$$g_1 g_2 \text{ is even;} \;\; h_1 h_2 \text{ is even;} \;\; g_1 h_2 \text{ is odd.} \tag{537}$$

We can apply Eqs. 535- 537 to the Fourier series representation of f, Eqs. 509-511. If f is even,

$$f(t) = \frac{a_0}{2} + \sum_n (a_n \cos\frac{n\pi t}{p} + b_n \sin\frac{n\pi t}{p}),$$

$$a_n = \frac{1}{p}\int_{-p}^{p} \underbrace{f(t)}_{\text{even}} \underbrace{\cos\frac{n\pi t}{p}}_{\text{even}} dt = \frac{2}{p}\int_0^p f(t)\cos\frac{n\pi t}{p}dt,$$

$$b_n = \frac{1}{p}\int_{-p}^{p} \underbrace{f(t)}_{\text{even}} \underbrace{\sin\frac{n\pi t}{p}}_{\text{odd}} dt = 0$$

$$\implies f(t) = \frac{a_0}{2} + \sum_n a_n \cos\frac{n\pi t}{p}, \quad a_n = \frac{2}{p}\int_0^p f(t)\cos\frac{n\pi t}{p}dt, \quad f \text{ even;} \tag{538}$$

if f is odd,

$$a_n = \frac{1}{p}\int_{-p}^{p} \underbrace{f(t)}_{\text{odd}} \underbrace{\cos\frac{n\pi t}{p}}_{\text{even}} dt = 0,$$

$$b_n = \frac{1}{p}\int_{-p}^{p} \underbrace{f(t)}_{\text{odd}} \underbrace{\sin\frac{n\pi t}{p}}_{\text{odd}} dt = \frac{2}{p}\int_0^p \underbrace{f(t)}_{\text{odd}} \underbrace{\sin\frac{n\pi t}{p}}_{\text{odd}} dt$$

$$\implies f(t) = \sum_n b_n \sin\frac{n\pi t}{p}, \quad b_n = \frac{2}{p}\int_0^p f(t)\sin\frac{n\pi t}{p}dt, \quad f \text{ odd.} \tag{539}$$

These are called the Fourier cosine and Fourier sine functions.

Now, suppose f is not periodic but is instead defined only over one interval. If this is the case, we can apply either a Fourier sine or Fourier cosine extension to this function (depending on if it is even or odd; if neither, then either extension works). Here we only evaluate the interval in which the function exists. These extensions are called half range expansions.

3.4 Lec 3d Fourier integrals

The standard form of the Fourier series of f with period $T = 2p$, once again (Eqs. 509-511),

$$f(t) = f(t + nT) = \frac{a_0}{2} + \sum_n (a_n \cos\frac{n\pi t}{p} + b_n \sin\frac{n\pi t}{p}),$$

where

$$a_n = \frac{1}{p}\int_{-p}^{p} f(t)\cos\frac{n\pi t}{p}dt,$$
$$b_n = \frac{1}{p}\int_{-p}^{p} f(t)\sin\frac{n\pi t}{p}dt.$$

Let us convert the standard form of this series into a complex exponential form. Specifically, let

$$\cos\theta = \frac{e^{i\theta} + e^{-i\theta}}{2}, \quad \sin\theta = \frac{e^{i\theta} - e^{-i\theta}}{2i}. \tag{540}$$

Applying Eq. 540 to Eq. 509, with $\theta = 2\pi t/p$,

$$f(t) = c_n e^{in\pi t/p} \tag{541}$$

where

$$c_n = \frac{1}{2p}\int_{-p}^{p} f(t)e^{-in\pi t/p}dt = \frac{1}{2p}\int_{-p}^{p} f(t)e^{-i\omega_n t}dt. \tag{542}$$

The correspondence between the real and complex coefficients are

$$c_0 = \frac{a_0}{2}, \quad c_n = \frac{a_n - ib_n}{2}, \quad c_{-n} = \frac{a_n + ib_n}{2}. \tag{543}$$

Here n in c_n is an index that travels negatively as well as positively. It is not practical but it is a good way to generalize the Fourier series. The derivation is in Greenberg.

Eq. 542 is a representation of f in the time domain. On the other hand, Eq. 543 can be made into a plot of $\mathrm{Re}(c_n)$ vs. ω_n, where

$$\omega_n = \frac{n\pi}{p}. \tag{544}$$

Then c_n represents the spectrum of f. Explanation needs work

Now, suppose f is not periodic. Then we do not have a frequency domain representation that involves only discrete frequencies. This is because we cannot obtain a Fourier series representation in the first place. Therefore, we must generalize this function f and allow the frequency representation to assume continuous values. Generalizing Eqs. 541 ($f(t) = c_n e^{in\pi t/p}$) and Eq. 542 ($c_n = \frac{1}{2p}\int_{-p}^{p} f(t)e^{-i\omega_n t}$), with frequency $\omega = n\pi/p$, we let

$$f(t) = \int_{-\infty}^{\infty} \mathcal{C}(\omega)e^{i\omega t}d\omega, \tag{545}$$

where complex coefficients

$$\mathcal{C}(\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt. \tag{546}$$

This is called the Fourier integral representation. The Fourier integral is basically a natural extension of the idea of a complex esponential Fourier series into functions f that are not periodic. In physical problems, time t corresponds to circular frequency ω. On the other hand, space x corresponds to wave number k. Basically, the analog of t is x and the analog of ω is k.

In general, if the Dirichlet conditions are satisfied and if the integral $\int_{-\infty}^{\infty} f(t)dt$ exists in the first place, then the Fourier integral that is Eq. 545 with $\mathcal{C}$ given in Eq. 546 actually gives the value of f wherever the function is continuous. Also, the Fourier integral converges to the average of the left and right hand limits of f where the function is discontinuous. And this is completly analoguos to the Dirichlet theorem itself In Lec 3.3.

Now, many functions satisfy the Dirichlet conditions. However, the condition that $\int_{-\infty}^{\infty} f(t)dt$ must exist is more exclusive. Basically, for this to be true, the function must decay as $t \to \pm\infty$. If the function does not decay sufficiently fast, then the integral is not bounded, and so the area under the curve does not exist, and so a Fourier representation is not possible.

However, while these conditions are sufficient, they are not necessary. There do exist some functions that possess a Fourier representation despite not being periodic. For example, consider

$$f(t) = \begin{cases} 0, & t < 0 \\ e^{-\alpha t}, & t > 0 \end{cases} \quad \text{for } \alpha > 0. \tag{547}$$

The Dirichlet conditions are satisfied. However, this function is not periodic. Still, using Eq. 546, we can write the Fourier integral representation of f

$$\begin{aligned}
\mathcal{C}(\omega) &= \frac{1}{2\pi}\int_0^{\infty} e^{-\alpha t}e^{-i\omega t}dt = \frac{1}{2\pi}\int_0^{\infty} e^{-(\alpha+i\omega)t}dt \\
&= \frac{1}{-2\pi(a+i\omega)}\lim_{T\to\infty}[e^{-(\alpha+i\omega)t}]\Big|_0^T \\
&= \frac{1}{-2\pi(a+i\omega)}(\lim_{T\to\infty}[e^{-(\alpha+i\omega)T}] - \lim_{T\to\infty}[e^0]) \\
&= \frac{1}{2\pi(a+i\omega)}(1 - \lim_{T\to\infty}[e^{-\alpha t}e^{-i\omega T}]) = \frac{1}{2\pi(a+i\omega)} = \mathcal{C}(\omega).
\end{aligned} \tag{548}$$

With $\mathcal{C}$, we can now use Eq. 545 to obtain

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{1}{2\pi(a+i\omega)}d\omega. \tag{549}$$

Lastly, recall the Fourier-cosine and Fourier-sine integral representations in the real form Eqs. 538 and 539,

$$f(t) = \sum_n b_n \sin\frac{n\pi t}{p}, \quad b_n = \frac{2}{p}\int_0^p f(t)\sin\frac{n\pi t}{p}dt, \quad f \text{ odd.}$$

$$f(t) = \frac{a_0}{2} + \sum_n a_n \cos\frac{n\pi t}{p}, \quad a_n = \frac{2}{p}\int_0^p f(t)\cos\frac{n\pi t}{p}dt, \quad f \text{ even.}$$

The representation is completely analogous in the frequency domain. If $\omega = n\pi/p$ and

$$f(t) = \int_0^\infty [\mathcal{A}(\omega)\cos\omega t + \mathcal{B}(\omega)\sin\omega t]d\omega, \tag{550}$$

where

$$\mathcal{A}(\omega) = \frac{1}{\pi}\int_{-\infty}^\infty f(t)\cos\omega t dt, \quad \mathcal{B}(\omega) = \frac{1}{\pi}\int_{-\infty}^\infty f(t)\sin\omega t dt, \tag{551}$$

then the corresponding Fourier cosine integral for f even is

$$f(t) = \int_0^\infty \mathcal{A}(\omega)\cos\omega t d\omega, \quad \mathcal{A}(\omega) = \frac{2}{\pi}\int_0^\infty f(t)\cos\omega t dt; \tag{552}$$

the corresponding Fourier sine integral for f odd is

$$f(t) = \int_0^\infty \mathcal{B}(\omega)\sin\omega t d\omega, \quad \mathcal{B}(\omega) = \frac{2}{\pi}\int_0^\infty f(t)\sin\omega t dt. \tag{553}$$

3.5 Lec 3e Fourier transforms

The purpose of Fourier analysis in the complex representation (Eq. 541-542) is to judge f in the domain of frequency and in the domain of time. We have essentially imposed transformations between these two domains, so that

$$f(t) = \frac{1}{2\pi}\int_{-\infty}^\infty F(\omega)e^{i\omega t}d\omega = \mathcal{F}^{-1}(F(\omega)), \tag{554}$$

$$F(\omega) = \int_{-\infty}^\infty f(t)e^{-i\omega t}dt = \mathcal{F}(f(t)). \tag{555}$$

Script $\mathcal{F}$ is called the Fourier transform. The Fourier transform of the time domain takes you to the frequency domain. The inverse Fourier transform takes you from the frequency domain back to the time domain.

Some useful examples are

$$f(t) = \begin{cases} 0, & t < 0 \\ e^{-\alpha t}, & t > 0 \end{cases} \Longrightarrow F(\omega) = \mathcal{F}(f(t)) = \frac{1}{\alpha + i\omega}, \tag{556}$$

$$f(t) = \begin{cases} e^{\alpha t}, & t \le 0 \\ e^{-\alpha t}, & t > 0 \end{cases} \Longrightarrow F(\omega) = \mathcal{F}(f(t)) = \frac{2\alpha}{\alpha + i\omega}. \tag{557}$$

These are from tables in Erdelyi (1954) and Greenberg (1998).

Besides for tabulating various F, $\mathcal{F}$ has useful mathematical properties. Assuming

$$F(\omega) = \mathcal{F}(f(t)), \quad f(t) = \mathcal{F}^{-1}(F(\omega)), \tag{558}$$

linearity holds, which means

$$\mathcal{F}(a_1 f_1 + a_2 f_2) = a_1 \mathcal{F}(f_1(t)) + a_2 \mathcal{F}(f_2(t)). \tag{559}$$

Symmetry holds, which means

$$\mathcal{F}(F(t)) = -2\pi f(\omega). \tag{560}$$

Here, while normally $\mathcal{F}$ transforms $f(t)$ into $F = F(\omega)$, it is also possible for a function to be given in terms of its frequency ω such that $f = f(\omega)$ and then to impose on it $\mathcal{F}$ to obtain $F(t)$. Eq. 560 describes that reverse relationship. As an extension of linearity, time scaling holds:

$$\mathcal{F}(F(\alpha t)) = \frac{1}{\alpha} F(\frac{\omega}{\alpha}); \tag{561}$$

time shifting holds:

$$\mathcal{F}(f(t - t_0)) = e^{-i\omega t} F(\omega); \tag{562}$$

and frequency shifting holds:

$$\mathcal{F}^{-1}(F(\omega - \omega_0)) = e^{i\omega_0 t} f(t). \tag{563}$$

Next, suppose we wish to impose Fourier transform $\mathcal{F}$ on f. This is possible if

- f is continuous everywhere;
- f' is piece wise continuous everywhere; and
- $\int_{-\infty}^{\infty} f(t)dt, \int_{-\infty}^{\infty} f'(t)dt$ exist.

Then the Fourier transform of the time derivative (Eq. 555)

$$\mathcal{F}(f'(t)) = \int_{-\infty}^{\infty} f'(t) e^{-i\omega t} dt; \tag{564}$$

integrating by parts with $u = e^{-i\omega t}, dv = f'(t)dt, \int udv = uv - \int vdu$, we obtain

$$\mathcal{F}(f'(t)) = i\omega F(\omega). \tag{565}$$

If this is repeated for higher order derivatives,

$$\mathcal{F}(f^{(n)}(t)) = (i\omega)^n F(\omega). \tag{566}$$

This rule is significant becuse it allows us to convert differential equations with constant coefficients in the time domain to algebraic equations in the frequency domain. The reverse operation from the frequency domain to the time domain is

$$\mathcal{F}^{-1}(F'(\omega)) = -itf(t), \tag{567}$$

$$\mathcal{F}^{-1}(F^{(n)}(\omega)) = -(it)^n f(t). \tag{568}$$

From here we introduce the convolution operation. Given functions $f(t)$ and $g(t)$, then

$$\text{convolution}(f,g) = (f*g)(t) := \int_a^{t-a} f(\tau)g(t-\tau)d\tau. \tag{569}$$

In the special case of $a = 0$, which is called unilateral convolution,

$$(f*g)(t) := \int_0^t \overbrace{f(\tau)}^{\boldsymbol{\Psi(\tau)}} \underbrace{g(t-\tau)}_{\text{fundamental solution}} d\tau. \tag{570}$$

For $a = -\infty$, which is called bilateral convolution,

$$(f*g)(t) := \int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau. \tag{571}$$

Time convolution is related to the product of Fourier transforms. Given $F(\omega)$ and $G(\omega)$, the product

$$F(\omega)G(\omega) = \mathcal{F}(\int_{-\infty}^{\infty} f(\tau)g(t-\tau)d\tau) \tag{572}$$

which is the exact form of a bilateral convolution in Eq. 571

3.6 Lec 3f Generalized functions

3.7 Lec 3g Laplace transforms

3.8 Lec 3h Integral transform summary

3.9 Lec 3i Boundary value problems

4 Mod4 PDEs

4.1 Lec 4a PDE introduction

An ODE has one independent variable (time or space) and n dependent variables. An example is

$$EI\frac{d^4w}{dx^4} = p(x). \tag{573}$$

On the other hand, a PDE has m independent variables (time and space) as well as any number n of dependent variables. An example is

$$EI\frac{\partial^4 w}{\partial x^4} + \rho A\frac{\partial^2 w}{\partial t^2} = p(x,t) \tag{574}$$

where $w = w(x,t)$ if $p = p(x,t)$. Consider Fig. 9, which is (1) a beam with base (depth) b, height h, and length L and then (2) a plate with lengths L_x and L_y and height h.

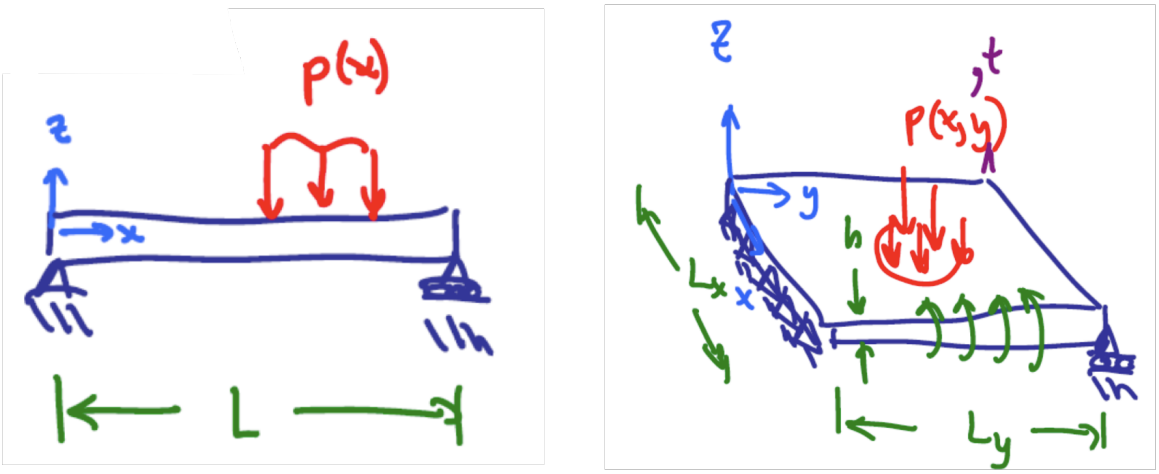

Figure 9: Beam/plate schemes

We assume that for the beam, $L \gg h$ and $L \gg b$. For the plate, $L_x \gg h$ and $L_y \gg h$. The PDEs which chracterize the solutions are

$$\text{Beam: } EI\frac{\partial^4 w}{\partial x^4} + \rho I\frac{\partial^2 w}{\partial t^2} = p(x,t), \tag{575}$$

$$\text{Plate: } D\left(\frac{\partial^4 w}{\partial x^4} + 2\frac{\partial^4 w}{\partial x^2 \partial y^2} + \frac{\partial^4 w}{\partial y^4}\right) + \rho h\frac{\partial^2 w}{\partial t^2} = p(x,y,t), \tag{576}$$

where

$$D = \frac{Eh^3}{12(1-\nu^2)}, \tag{577}$$

E is elastic modulus, ν is Poisson's ratio, I is the cross sectional bending moment, ρ is the material density(?), and A is cross sectional area(?).

Associated with these problems are of course initial and boundary conditions. These are 2nd order PDEs with independent variables x, y, t, etc. and dependent variable $u(x,y)$. Their form is often

$$A_{11}(x,y)\frac{\partial^2 u}{\partial x^2} + 2A_{12}(x,y)\frac{\partial^2 u}{\partial x \partial y} + A_{22}(x,y)\frac{\partial^2 u}{\partial y^2} + g(x,y,u,\frac{\partial u}{\partial x},\frac{\partial u}{\partial y}) = 0. \tag{578}$$

To determine homogeneity, g is decomposed into

$$g = g_1(x,y) + g_2(\dots). \tag{579}$$

If $g_1 = 0$, the equation is homogeneous. If $g_1 \neq 0$, the equation is nonhomogeneous. The linearity of the function also depends on g. If g is linear in u, $\partial u/\partial x$, $\partial u/\partial y$, then the equation is linear. If g is not linear in these ways, then the equation is not linear.

Eq. 578 can be generalized to

$$A_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + g = 0 \tag{580}$$

where $i, j \Rightarrow 1, 2$ and $x_1 \Leftarrow x,\ x_2 \Leftarrow y$.

Now, second order tensor

$$A_{ij} \Longleftrightarrow [\mathbf{A}] = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \tag{581}$$

has determinant

$$\det \mathbf{A} = A_{11}A_{22} - A_{12}A_{21} = A_{11}A_{22} - A_{12}^2 \tag{582}$$

because $\mathbf{A}$ is always symmetric.

If $\det \mathbf{A} > 0$, Eq. 580 is an elliptic PDE. An elliptic PDE is typically a steady state process not dependent on time. Laplace's equation

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \tag{583}$$

follows from $A_{11} = 1,\ A_{12} = 0, A_{22} = 1$, meaning $\mathbf{A} = \mathbf{I} \Rightarrow \det \mathbf{A} = 1 > 0$. In index notation,

$$u_{,ii} = 0. \tag{584}$$

In another notation,

$$\nabla^2 u = 0. \tag{585}$$

Laplace's equation has a few applications. Without time, Laplace's equation is the steady state heat conduction equation, where u is temperature.

If $\det \mathbf{A} = 0$, Eq. 580 is a parabolic PDE. A parabolic PDE is typically a diffusive system that is dependent on time, so that in 1D, $x_1 \Leftarrow x, x_2 \Leftarrow t$. Then the 1D heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2} \tag{586}$$

follows from $g = -\partial u/\partial t,\ A_{11} = \kappa, A_{12} = A_{22} = 0$. Therefore, $\det \mathbf{A} = 0$. κ is the thermal diffusivity coefficient.

In 2D, $x_1 \Leftarrow x,\ x_2 \Leftarrow y,\ x_3 \Leftarrow t$, and the 2D heat equation

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2} + \kappa \frac{\partial^2 u}{\partial y^2}. \tag{587}$$

This means

$$[\mathbf{A}] = \begin{bmatrix} \kappa & 0 & 0 \\ 0 & \kappa & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{588}$$

$\det \mathbf{A} = 0$.

Now, if $\det \mathbf{A} < 0$, Eq. 580 is called a hyperbolic PDE. Its solutions are like waves, and they usually involve space and time. The wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{589}$$

follows from $A_{11} = c^2,\ A_{12} = 0,\ A_{22} = -1$. Then, $\det \mathbf{A} = -1 < 0$.

4.2 Lec 4b Hyperbolic PDEs

Recall the 2nd order PDE Eq. 580, which is

$$A_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + g = 0. \tag{590}$$

Let us consider the different forms of PDEs in the order of hyperbolic, parabolic, and then elliptic. So, first are hyperbolic PDEs. An example of a hyperbolic PDE is the vibration of a string. We assume the string is homogeneous, elastic and has no resistance to bending. Tension in the string is sufficiently large so that the weight of the mass is negligible in comparison. Small horizontal movement is neglected.

We examine a section Δx of the string in its deformed configuration, which is the red line. (The reference configuration is the blue line.) The tension is always tangent to the string. Since horizontal movement is neglected, the tensile forces in the x direction on both sides shall be equal. That is,

$$T_1 \cos\alpha = T_2 \cos\beta = T_x, \tag{591}$$

where T_i are the forces on the left and right side, and α and β are the corresponding angles between the tensile force and the normal axis.

Now, assuming vertical motion, we can thus apply Newton's second law, which is

$$F = \frac{d}{dt}(mv). \tag{592}$$

We can also define force as

$$F = \rho A \Delta x \frac{\partial^2 u}{\partial t^2} \tag{593}$$

because $\rho A = mA/V = m/L$ is mass per unit length, and Δx is the length of interest, giving only mass. In addition, $\partial^2 u/\partial t^2$ is acceleration. Together, we get force. Another representation of force is

$$-T_1 \sin\alpha + T_2 \sin\beta = F \tag{594}$$

because of the sum of the tensile forces in the vertical direction. The signage of both terms depends on the drawing. Then, substituting,

$$-T_1 \sin\alpha + T_2 \sin\beta = \rho A \Delta x \frac{\partial^2 u}{\partial t^2} \tag{595}$$

implies

$$\frac{-T_1 \sin\alpha + T_2 \sin\beta}{T_x} = \frac{\rho A}{T_x}\Delta x \frac{\partial^2 u}{\partial t^2} \tag{596}$$

implies

$$\frac{-T_1 \sin\alpha}{T_1 \cos\alpha} + \frac{T_2 \sin\beta}{T_2 \cos\beta} = \frac{\rho A}{T_x}\Delta x \frac{\partial^2 u}{\partial t^2} \tag{597}$$

implies

$$\tan\beta - \tan\alpha = \frac{\rho A}{T_x}\Delta x \frac{\partial^2 u}{\partial t^2}. \tag{598}$$

Tangents inform slope, as in the unit circle. We let

$$\tan\beta = \frac{\partial u}{\partial x}|_{x+\Delta x}, \qquad \tan\alpha = \frac{\partial u}{\partial x}|_{x}. \tag{599}$$

Then

$$\frac{\partial u}{\partial x}|_{x+\Delta x} - \frac{\partial u}{\partial x}|_{x} = \frac{\rho A}{T_x}\Delta x \frac{\partial^2 u}{\partial t^2} \tag{600}$$

implies

$$\left(\frac{\partial u}{\partial x}|_{x+\Delta x} - \frac{\partial u}{\partial x}|_{x}\right)\frac{1}{\Delta x} = \frac{\rho A}{T_x}\frac{\partial^2 u}{\partial t^2}. \tag{601}$$

Now, the left hand side takes the form of the definition of a spatial derivative. Therefore,

$$\frac{\partial}{\partial x}\left(\frac{\partial u}{\partial x}\right) = \frac{\rho A}{T_x}\frac{\partial^2 u}{\partial t^2} \tag{602}$$

which implies

$$\frac{\partial^2 u}{\partial x^2} = \frac{\rho A}{T_x}\frac{\partial^2 u}{\partial t^2}. \tag{603}$$

Tensile force $T_x = \sigma_x A$ because $\sigma_x := \text{force/area} = T_x/A$. Substituting,

$$\frac{\partial^2 u}{\partial x^2} = \frac{\rho A}{\sigma_x A}\frac{\partial^2 u}{\partial t^2} = \frac{\rho}{\sigma_x}\frac{\partial^2 u}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} \tag{604}$$

implies the prototypical 1D wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \tag{605}$$

where $c^2 = \sigma_x/\rho$ is the wave speed.

We want to trace the vertical vibration of the string after we pluck it. We impose the boundary conditions that the horizontal displacement does not change at either, which are

$$u(0,t) = 0, \quad u(l,t) = 0, \qquad t \geq 0. \tag{606}$$

Also, we want to start the string as stationary, expressed in initial conditions as

$$u(x,0) = f_1(x), \quad \dot{u}(x,0) = f_2(x), \quad 0 \geq x \geq l. \tag{607}$$

To solve the PDE Eq. 605 we must find the solution to each derivative term separately and then plug it back into the governing equation. To do this the solution strategy is the method of separation of variables. Let the solution be some

$$u(x,t) = F(x)G(t). \tag{608}$$

Then, derivatives

$$\frac{\partial^2 u}{\partial x^2} = G\frac{d^2F}{dx^2} = GF'', \qquad \frac{\partial^2 u}{\partial t^2} = F\frac{d^2G}{dt^2} = F\ddot{G}. \tag{609}$$

Substituting Eq. 609 into Eq. 605,

$$F\ddot{G} = c^2 GF'' \tag{610}$$

implies

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F} = \kappa \tag{611}$$

which is represented as

$$g(t) = f(x) = \text{constant}. \tag{612}$$

Then,

$$\ddot{G} = \kappa c^2 G, \qquad F'' = \kappa F \tag{613}$$

implies

$$\ddot{G} - \kappa c^2 G = 0, \qquad F'' - \kappa F = 0. \tag{614}$$

Eqs. 614 are a pair of second order linear homogeneous ODEs. κ is still unknown though. First, recall the boundary conditions Eq. 606, as well as the general form of the solution Eq. 608. Substituting,

$$u(0,t) = F(0)G(t) = 0, \qquad u(l,t) = F(l)G(t) = 0 \tag{615}$$

means

$$F(0) = F(l) = 0. \tag{616}$$

Now we look to find κ. Consider Eq. 614, particularly $F'' - \kappa F = 0$.

If $\kappa = 0$,

$$F'' = 0 \Longrightarrow F(x) = C_1 x + C_2. \tag{617}$$

Then

$$0 = F(0) = C_2 \Longrightarrow F(x) = C_1 x. \tag{618}$$

Then

$$0 = F(l) = C_1 l \Longrightarrow C_1 = 0 \Longrightarrow F(x) = 0 \tag{619}$$

is the uninteresting trivial solution. Therefore we let $\kappa \neq 0$.

If $\kappa = k^2 > 0$, from Eq. 614 we receive

$$F'' = k^2 F \Longrightarrow F(x) = C_1 e^{kx} + C_2 e^{-kx}. \tag{620}$$

Then

$$0 = F(0) = C_1 + C_2, \tag{621}$$

$$0 = F(l) = C_1 e^{kl} + C_2 e^{-kl}, \tag{622}$$

again admitting only the uninteresting trivial solution

$$F(x) = 0. \tag{623}$$

Now, if $\kappa = -k^2 < 0$, then from Eq. 614 and from Euler's formula we receive

$$F'' = k^2 F \Longrightarrow F(x) = C_1 \cos kx + C_2 \sin kx. \tag{624}$$

Then

$$0 = F(0) = C_1 \implies F(x) = C_2 \sin kx, \tag{625}$$

$$0 = F(l) = C_2 \sin kl. \tag{626}$$

Eq. 626 means that either $C_2 = 0$ or that $\sin kl = 0$. The first result is the trivial solution, but the second result is more interesting. So, we are content with $\kappa = -k^2$. If $kl = n\pi$,

$$k_n = \frac{n\pi}{l}, \qquad \kappa_n = -\frac{n^2\pi^2}{l^2}. \tag{627}$$

Now, Eq. 614 ($F'' - \kappa F = 0$) happens to be able to be turned into a Sturm Liouville eigenproblem, as in Sec. 2.6. This means that κ_n are characteristic eigenvalues and

$$F_n(x) = \sin \frac{n\pi x}{l} \tag{628}$$

are characteristic eigenfunctions. In general, the eigenvalues are associated with natural frequencies and the eigenfunctions are associated with mode shapes.

Thus far the spatial part F has been considered in Eq. 614. Now let us consider the temporal part G. Fully,

$$\ddot{G} - c^2 \kappa G = 0. \tag{629}$$

Substituting in κ_n,

$$\ddot{G}_n + c^2 \frac{n^2\pi^2}{l^2} G = \omega_n^2 G = 0, \tag{630}$$

where $\omega_n = cn\pi/l$ is the circular frequency of the string. The form of Eq. 630 along with Euler's formula admit the general form of the solution

$$G_n(t) = C_1^* \cos \omega_n t + C_2^* \sin \omega_n t. \tag{631}$$

Note that this function is not goverened by a Sturm Liouville eigenproblem in contrast to F because it is an initial value problem, not a boundary value problem.

Altogether, the form of the overall solution to the combined initial/boundary value problem is

$$u_n(x,t) = F_n(x) G_n(t) = \left(\sin \frac{n\pi x}{l} \right) \left(C_{1n}^* \cos \omega_n t + C_{2n}^* \sin \omega_n t \right), \tag{632}$$

where

$$u(x,t) = \sum_n u_n(x,t). \tag{633}$$

4.3 Lec 4c String initial boundary value problem solutions

Recall the orthogonality property Eq. 522 in Sec. 3.2, which is

$$\int_0^l \sin\left(\frac{m\pi x}{l}\right) \sin\left(\frac{n\pi x}{l}\right) dx = \begin{cases} 0, & m \neq n \\ l/2, & m = n \end{cases} \qquad = \frac{l}{2}\delta_{mn}. \tag{634}$$

Also, we assume the separable solution

$$u(x,t) = \sum_n \left(\sin\frac{n\pi x}{l}\right)\left(C_{1n}^* \cos\omega_n t + C_{2n}^* \sin\omega_n t\right), \tag{635}$$

as in Eqs. 632-633. Recall that eigenfunctions $F_n(x)$ are mode shapes and that eigenvalues ω_n are natural frequencies. However, C_{in}^* are not known. To find them, we turn to the intial conditions 607, which are $u(x,0) = f_1(x),\ \dot{u}(x,0) = f_2(x)$. Substituting in the assumed separable solution to the position IC,

$$u(x,0) = \sum_n C_{1n} \sin\frac{n\pi x}{l} = f_1(x), \tag{636}$$

which is a Fourier sine series, as in Eq. 539. By virtue of a Fourier sine function, this means that the coefficients

$$C_{1n} = \frac{2}{l}\int_0^l f_1(x) \sin\frac{n\pi x}{l} dx. \tag{637}$$

In terms of the velocity IC,

$$\frac{\partial u}{\partial t}(x,0) = \sum_n C_{2n}\omega_n \sin\frac{n\pi x}{l} dx = f_2(x), \tag{638}$$

where

$$C_{2n} = \frac{2}{\omega_n l}\int_0^l f_2(x) \sin\frac{n\pi x}{l} dx. \tag{639}$$

Provided ICs $f_1,\ f_2$, it is possible to approximate C_{in} numerically.

An example of initial conditions and boundary conditions stated explicitly to solve for the PDE are

$$u(x,0) = \begin{cases} 2u_0 x/l, & 0 < x < l/2 \\ 2u_0(l-x)/l & l/2 < x < l \end{cases} = f_1(x), \qquad \dot{u}(x,0) = 0 = f_2(x), \tag{640}$$

$$u(0,t) = u(l,t) = 0. \tag{641}$$

The spatial boundary conditions (BCs) and temporal initial conditions (ICs) are visualized in Fig. 10.

Two functions on each half of the string define separate BCs. Dividing Eq. 637 and Eq. 639 into the two intervals,

$$C_{1n} = \frac{2}{l}\left(\int_0^{l/2} \frac{2u_0}{l} x \sin\frac{n\pi x}{l} dx + \int_{l/2}^{l} \frac{2u_0}{l}(l-x)\sin\frac{n\pi x}{l} dx\right) \tag{642}$$

and

$$C_{2n} = 0 \quad (f_2 = 0). \tag{643}$$

Figure 10: Example string IBVP.

Integrating,

$$C_{1n} = \frac{2v_0}{n^2\pi^2}\sin\frac{n\pi}{2} = \begin{cases} 8(-1)^{(n-1)/2}v_0/n^2\pi^2, & n = 1,3,5,\ldots \\ 0, & n = 0,2,4,\ldots\,. \end{cases} \tag{644}$$

Substituting this result into Eq. 632,

$$u(x,t) = \sum_{n=1,3,5,\ldots} (-1)^{(n-1)/2}\frac{8v_0}{n^2\pi^2}\cos\frac{n\pi ct}{l}\sin\frac{n\pi x}{l} \tag{645}$$

$$= \sum_{n=1,3,5,\ldots} 2A_n \sin\frac{n\pi x}{l}\cos\frac{n\pi ct}{l}, \tag{646}$$

where $A_n = 4(-1)^{(n-1)/2}v_0/n^2\pi^2$. To simplify further, consider the trig identity

$$\sin\alpha\cos\beta = \frac{1}{2}[\sin(\alpha+\beta)+\sin(\alpha-\beta)], \quad \alpha = \frac{n\pi x}{l}, \beta = \frac{n\pi ct}{l}. \tag{647}$$

Substituting,

$$u(x,t) = \sum_{n=1,3,5,\ldots} A_n\left[\sin(\frac{n\pi x}{l}+\frac{n\pi ct}{l})+\sin(\frac{n\pi x}{l}-\frac{n\pi ct}{l})\right] \tag{648}$$

$$= \sum_{n=1,3,5,\ldots} A_n\left[\sin\frac{n\pi}{l}(x+ct)+\sin\frac{n\pi}{l}(x-ct)\right] = u(x,t). \tag{649}$$

Eq. 649 has two terms. The first is a tent propagating to the left. The second is a tent propagating to the right. Note that c is a constant.

4.4 Lec 4d d'Alembert solutions

Once again, consider the 1d wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2\frac{\partial^2 u}{\partial x^2}, \tag{650}$$

or

$$u_{,tt} = c^2 u_{,xx}. \tag{651}$$

Now, let us introduce new variables

$$\eta = x + ct, \quad \xi = x - ct. \tag{652}$$

The chain rule admits the first derivative with respect to x

$$u_{,x} = \frac{\partial u}{\partial x} = \frac{\partial u}{\partial \eta}\frac{\partial \eta}{\partial x} + \frac{\partial u}{\partial \xi}\frac{\partial \xi}{\partial x} = \frac{\partial u}{\partial \eta}(1) + \frac{\partial u}{\partial \xi}(1) = u_{,\eta} + u_{,\xi}. \tag{653}$$

Similarly, Second derivative

$$\begin{aligned} u_{,xx} &= \frac{\partial(\partial u/\partial \eta)}{\partial x} + \frac{\partial(\partial u/\partial \xi)}{\partial x} \\ &= \frac{\partial(\partial u/\partial \eta)}{\partial \eta}\frac{\partial \eta}{\partial x} + \frac{\partial(\partial u/\partial \eta)}{\partial \xi}\frac{\partial \xi}{\partial x} + \frac{\partial(\partial u/\partial \xi)}{\partial \eta}\frac{\partial \eta}{\partial x} + \frac{\partial(\partial u/\partial \xi)}{\partial \xi}\frac{\partial \xi}{\partial x} \\ &= u_{,\eta\eta} + u_{,\eta\xi} + u_{,\xi\eta} + u_{,\xi\xi} = u_{,\eta\eta} + 2u_{,\eta\xi} + u_{,\xi\xi} = u_{,xx} = \frac{\partial^2 u}{\partial x^2}. \end{aligned} \tag{654}$$

On the other hand, first and second derivatives with respect to t are

$$u_{,t} = \frac{\partial u}{\partial t} = u_{,\eta}\eta_{,t} + u_{,\xi}\xi_{,t} = u_{,\eta} - u_{,\xi} \tag{655}$$

and

$$\begin{aligned} u_{,tt} &= (u_{,\eta})_{,\eta}\eta_{,t} + (u_{,\eta})_{,\xi}\xi_{,t} + (u_{,\xi})_{,\eta}\eta_{,t} + (u_{,\xi})_{,\xi}\xi_{,t} \\ &= u_{\eta\eta} - 2u_{,\eta\xi} + u_{\xi\xi} = u_{,tt}. \end{aligned} \tag{656}$$

Substituting Eq. 654 and Eq. 656 into the 1d wave equation,

$$u_{\eta\eta} - 2u_{,\eta\xi} + u_{\xi\xi} = c^2(u_{,\eta\eta} + 2u_{,\eta\xi} + u_{,\xi\xi}). \tag{657}$$

From this representation we conclude that

$$(2c^2 + 2)u_{,\eta\xi} = 0 \Rightarrow u_{,\eta\xi} = 0 \tag{658}$$

if c is a real constant, and it is.

Now, let us represent the partial derivative of u with respect to η as some function

$$h(\eta) = u_{,\eta}. \tag{659}$$

Integrating this function with respect to η,

$$u = u(\eta, \xi) = \int u_{,\eta} d\eta = \int h(\eta) d\eta + \Psi(\xi) = \Phi(\eta) + \Psi(\xi). \tag{660}$$

Here, Φ is the antiderivative of h and Ψ is some scalar function which may have some ξ dependence since the integral is being taken without respect to ξ. It is essentially an integration constant, like how $\int f'(x)dx = f(x) + C$. Substituting in the definitions of ξ and η,

$$u = u(x, t) = \Phi(x + ct) + \Psi(x - ct). \tag{661}$$

This is the d'Alembert solution. Functions Φ and Ψ are to be determined specifically by the ICs of the problem. Suppose we have the special case of ICs

$$\dot{u}(x,0) = 0, \quad u(x,0) = f(x). \tag{662}$$

Physically, this means that the string is released from rest and that the reference configuration has the exact same shape as the curve $f(x)$. Using Eq. 661 and the chain rule, the partial derivative of u with respect to t

$$u_{,t}(x,t) = c\Phi_{,\eta} - c\Psi_{,\xi}. \tag{663}$$

Initially, the velocity IC is

$$\begin{aligned} 0 = u_{,t}(x,0) = c\Phi_{,\eta} - c\Psi_{,\xi} \implies 0 = \Phi_{,\eta} - \Psi_{,\xi} \\ \implies \Phi_{,\eta} = \Psi_{,\xi} \implies \Phi = \Psi + K. \end{aligned} \tag{664}$$

The displacement IC is

$$f(x) = u(x,0) = \Phi + \Psi = (\Psi + K) + \Psi = 2\Psi + K = f(x) \tag{665}$$

which implies

$$\Psi = \frac{1}{2}[f(x) - K], \qquad \Phi = \frac{1}{2}[f(x) + K]. \tag{666}$$

Substituting this result into Eq. 661,

$$u(x,t) = \frac{1}{2}[f(x+ct) + \cancel{K} + f(x-ct) - \cancel{K}] = \frac{1}{2}[f(x+ct) + f(x-ct)]. \tag{667}$$

4.5 Lec 4e Heat diffusion introduction

A few assumptions of heat diffusion follow from experimentation:

- **I**: Heat flows from hot to cold;
- **II**: Rate of heat flow is proportional to (1) the magnitude of the cross sectional area through which it flows and (2) the temperature gradient in the direction normal to that cross sectional area;
- **III**: The quantity of heat gained or lost is proportional to (1) the mass of the body and (2) the change in temperature.

Henceforth, u is temperature. A conceptual interpretation of the energy balance equation is

$$\underbrace{\text{rate of heat stored}}_{\mathbf{1}} = \underbrace{\text{rate of heat produced}}_{\mathbf{2}} + \underbrace{\text{rate of heat flowing in}}_{\mathbf{3}}. \tag{668}$$

For **1**, the rate of heat stored ΔQ is defined according to assumption **III**, which is basically

$$\Delta Q = c_v \Delta m \Delta u. \tag{669}$$

Mass change

$$\Delta m = \rho \Delta V = \rho \Delta x \Delta y \Delta z, \tag{670}$$

meaning

$$\Delta Q = \rho c_v \Delta x \Delta y \Delta z \Delta u. \tag{671}$$

The time rate of change of heat change

$$\frac{\Delta Q}{\Delta t} = \rho c_v \Delta x \Delta y \Delta z \frac{\Delta u}{\Delta t}. \tag{672}$$

c_v is called the specific heat, and it is a rating of the relationship between heat gain and temperature gain.

For **2**, the rate of heat produced within the body is some

$$\Phi(x, y, z, t) \Delta x \Delta y \Delta z, \tag{673}$$

where Φ is the heat rate per unit volume and $\Delta V = \Delta x \Delta y \Delta z$ is the volume of the body.

For **3**, the rate of heat flowing into the surface invokes assumption **II**. In the direction x:

$$\begin{aligned} &\text{heat flow through surfaces in } x = k \Delta y \Delta z \frac{\partial u}{\partial x}|_{x+\Delta x} - k \Delta y \Delta z \frac{\partial u}{\partial x}|_x \\ &= k \Delta y \Delta z \left(\frac{\partial u}{\partial x}|_{x+\Delta x} - \frac{\partial u}{\partial x}|_x \right), \end{aligned} \tag{674}$$

where k is a rating of flow magnitude and is called the thermal conductivity. In the direction y,

$$\text{heat flow through surfaces in } y = k \Delta x \Delta z \left(\frac{\partial u}{\partial y}|_{y+\Delta y} - \frac{\partial u}{\partial y}|_y \right). \tag{675}$$

In the direction z,

$$\text{heat flow through surfaces in } z = k \Delta x \Delta y \left(\frac{\partial u}{\partial z}|_{z+\Delta z} - \frac{\partial u}{\partial z}|_z \right). \tag{676}$$

Therefore, the total energy balance equation

$$\begin{aligned} &\rho c_v \Delta x \Delta y \Delta z \frac{\Delta u}{\Delta t} = \Phi(x, y, z, t) \Delta x \Delta y \Delta z \\ &+ k \Delta y \Delta z \left(\frac{\partial u}{\partial x}|_{x+\Delta x} - \frac{\partial u}{\partial x}|_x \right) + k \Delta x \Delta z \left(\frac{\partial u}{\partial y}|_{y+\Delta y} - \frac{\partial u}{\partial y}|_y \right) + k \Delta x \Delta y \left(\frac{\partial u}{\partial z}|_{z+\Delta z} - \frac{\partial u}{\partial z}|_z \right). \end{aligned} \tag{677}$$

This implies

$$\begin{aligned} \rho c_v \frac{\Delta u}{\Delta t} &= \Phi(x, y, z, t) + \frac{k}{\Delta x} \left(\frac{\partial u}{\partial x}|_{x+\Delta x} - \frac{\partial u}{\partial x}|_x \right) \\ &+ \frac{k}{\Delta y} \left(\frac{\partial u}{\partial y}|_{y+\Delta y} - \frac{\partial u}{\partial y}|_y \right) + \frac{k}{\Delta z} \left(\frac{\partial u}{\partial z}|_{z+\Delta z} - \frac{\partial u}{\partial z}|_z \right). \end{aligned} \tag{678}$$

Now, just like in Eq. 601, certain expressions take on the form of the definition of the derivative. Substituting appropriately,

$$\rho c_v \frac{\partial u}{\partial t} = \Phi(x,y,z,t) + k\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}\right). \tag{679}$$

Isolating the time evolution of temperature,

$$\frac{\partial u}{\partial t} = \kappa \nabla^2 u + \Phi(x,y,z,t), \tag{680}$$

where $\kappa = k/\rho c_v$ is called the thermal diffusivity.

Now, recall the 2nd order PDE Eq. 580, which is

$$A_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + g = 0. \tag{681}$$

The different forms of PDEs are hyperbolic ($\det \mathbf{A} < 0$), parabolic ($\det \mathbf{A} = 0$), and elliptic ($\det \mathbf{A} > 0$). So far we have discussed at length hyperbolic PDEs. Now we consider the parabolic PDEs that is the heat equation. We know it is parabolic because if $x_1 \Leftarrow x,\ x_2 \Leftarrow y,\ x_3 \Leftarrow z,\ x_4 \Leftarrow t$, and if $g = -\partial u/\partial t + \Phi$, then

$$A_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial u}{\partial t} \tag{682}$$

implies

$$[\mathbf{A}] = \begin{bmatrix} \kappa & 0 & 0 & 0 \\ 0 & \kappa & 0 & 0 \\ 0 & 0 & \kappa & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \det \mathbf{A} = \kappa^3(0) + 0 + 0 + \ldots + 0 = 0. \tag{683}$$

Suppose we simplify the model in that $\Phi = 0$. Then,

$$\frac{\partial u}{\partial t} = \dot{u} = \nabla^2 u \tag{684}$$

is a homogeneous PDE in three directions. Focusing on one direction only, the 1d heat equation is

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2},$$

which was articulated first in Eq. 586.

The strictest simplification actually changes the nature of the PDE from parabolic to elliptic. That simplification is one of steady heat flow, or

$$\nabla^2 u = 0 \quad (\dot{u} = 0), \tag{685}$$

meaning

$$A_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + \not{g} = 0 \tag{686}$$

implies

$$[\mathbf{A}] = \mathbf{I} \iff \delta_{ij} \Rightarrow \det \mathbf{A} = 1 > 0. \tag{687}$$

This leads to

$$\delta_{ij} \frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_i \partial x_i} \iff \nabla^2 u = 0. \tag{688}$$

Now, let us consider heat flow in a long, slender bar of length l positioned in the direction x. We assume the cross section is constant, that the bar has uniform material properties, and that it is perfectly insulated in the lateral directions, meaning there is only heat flow in the x direction and not in the other directions. Based on these assumptions, we can reduce the heat equation to 1d, which is, again,

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2},$$

which is Eq. 586. In conjunction with the PDE several boundary conditions can be imposed. The homogeneous zero temp I/BCs are

$$u(0,t) = u(l,t) = 0, \quad t > 0, \tag{689}$$

$$u(x,0) = f(t), \quad 0 < x < l. \tag{690}$$

Other common BCs at $x = 0$ are

$$-\kappa \frac{\partial u}{\partial x}(0,t) = \hat{\mathbf{q}}, \quad u(0,t) = \hat{u}, \tag{691}$$

or

$$-\kappa \frac{\partial u}{\partial x}(0,t) = h[u(0,t) - \hat{u}_{\text{amb}}], \tag{692}$$

where h is some heat convection/transfer coefficient and $\hat{u}_{\text{amb}}$ is the ambient temperature of the surrounding fluid, provided that the bar is surrounded by a fluid environment.

The way to solve this is just like that in the string, which is the separation of variables. We assume the decomposition

$$u(x,t) = F(x)G(t). \tag{693}$$

Then the procedure is completely analogous to Sec. 4.2, starting at Eq. 608.

4.6 Lec 4f Heat diffusion in rod

Let us continue the derivation of the solution to the 1d heat equation Eq. 586, or

$$\frac{\partial u}{\partial t} = \kappa \frac{\partial^2 u}{\partial x^2}.$$

which characterizes a rod. Throughout, notice the analogies to the string. Suppose the I/BCs

$$u(0,t) = u(l,t) = 0, \quad u(x,0) = f(x) \tag{694}$$

hold for this system. Assuming the decomposition

$$u(x,t) = F(x)G(t), \tag{695}$$

this can be substituted into Eq. 586 to receive

$$F\dot{G} = \kappa F''G \tag{696}$$

which implies

$$\frac{\dot{G}}{\kappa G} = \frac{\dot{F}}{F} = \beta = \text{constant}, \tag{697}$$

where β is not yet known. Then,

$$F'' - \beta F = 0, \tag{698}$$

$$\dot{G} - \beta\kappa G = 0 \tag{699}$$

are the two separated ODEs that can be solved separately and then integrated back into the governing equation. Starting with the spatial part F, the only nontrivial solution follows from $\beta = -\mu^2$, so that

$$F'' = \beta F = -\mu^2 F \Longrightarrow F(x) = C_1 \cos\mu x + C_2 \sin\mu x. \tag{700}$$

If BCs are

$$F(0) = F(l) = 0, \tag{701}$$

then

$$0 = F(0) = C_1 \Longrightarrow F(x) = C_2 \sin\mu x, \tag{702}$$

$$0 = F(l) = C_2 \sin\mu l \Longrightarrow \sin\mu l = 0 \Longrightarrow \mu l = n\pi. \Longrightarrow \mu = \frac{n\pi}{l}. \tag{703}$$

Generalizing,

$$F_n = C_2 \sin\frac{n\pi x}{l}. \tag{704}$$

For the temporal part G, assume the same $\beta = -\mu^2$, so that

$$\dot{G}_n + \mu_n^2 \kappa G_n = 0. \tag{705}$$

This means

$$G_n(t) = C_3 e^{-\mu_n^2 \kappa t} = C_3 e^{-n^2\pi^2\kappa t/l^2}. \tag{706}$$

Reintegrating the two parts back into the governing equation,

$$u_n(x,t) = F_n(x)G_n(t) = b_n \sin\frac{n\pi x}{l} e^{-n^2\pi^2\kappa t/l^2}, \tag{707}$$

and

$$u(x,t) = \sum_n u_n(x,t). \tag{708}$$

Then, if ICs are

$$u(x,0) = f(x), \tag{709}$$

then

$$u(x,0) = \sum_n b_n \sin \frac{n\pi x}{l}, \tag{710}$$

which is a Fourier sine series Eq. 539. By virtue of this definition, coefficients

$$b_n = \frac{2}{l} \int_0^l f(x) \sin \frac{n\pi x}{l} dx. \tag{711}$$

An example is the transient heat flow in a bar.

4.7 Lec 4g Vibrating membrane

4.8 Lec 4h Transform approaches