MODULE 1

Topics: Vectors space, subspace, span

I. Vector spaces:

General setting: We need

\( V = \) a set: its elements will be the vectors \( x, y, f, u \), etc.
\( F = \) a scalar field: its elements are the numbers \( \alpha, \beta \), etc.

+ a rule to add elements in \( V \)
• a rule to multiply elements in \( V \) with numbers in \( F \).

\( V, F, + \) and \( \cdot \) can be quite general within the abstract framework of vector spaces.

IN THIS COURSE a vector in \( V \) is generally

i) an \( n \)-tuple of real or complex numbers

\[
x = (x_1, \ldots, x_n)
\]

or

ii) a function defined on a given set \( D \).

In this case it is common to denote the vector by \( f \).

The scalar field \( F \) is generally the set of real or complex numbers.

+ is the component wise addition of \( n \)-tuples of numbers

\[
x + y = (x_1 + y_1, \ldots, x_n + y_n)
\]

or the pointwise addition of functions

\[
(f + g)(t) = f(t) + g(t)
\]

• is the usual multiplication of an \( n \)-tuple with a scalar

\[
x = (x_1, \ldots, x_n)
\]

or the pointwise multiplication of a function

\[
(\alpha f)(t) = \alpha f(t).
\]
Hence nothing special or unusual is happening in this regard.

**Definition:** If for any \(x, y \in V\) and any \(\alpha \in F\)

\[x + y \in V \]

\[\alpha x \in V\]

then \(V\) is a vector space (over \(F\)).

We say that \(V\) is closed under vector addition and scalar multiplication.

**Examples:**

i) All \(n\)-tuples of real numbers form the vector space \(\mathbb{R}_n\) over the real numbers \(\mathbb{R}\).

ii) All \(n\)-tuples of complex numbers form the vector space \(\mathbb{C}_n\) over the complex numbers \(\mathbb{C}\).

iii) All continuous real valued functions on a set \(D\) form a vector space over \(\mathbb{R}\).

iv) All \(k\)-times continuously differentiable functions on a set \(D\) form a vector space over \(\mathbb{R}\).

**Convenient notation:**

\(C^k(a, b)\) denotes the vector space of \(k\)-times continuously differentiable functions defined on the open interval \((a, b)\).

\(C^k[a, b]\) denotes the vector space of \(k\)-times continuously differentiable functions defined on the closed interval \([a, b]\).

v) All real valued \(n\)-tuples of the form

\[x = (1, x_2, \ldots, x_n)\]

do not form a vector space over \(\mathbb{R}\) because \(0 \cdot x\) is the zero vector and has a 0 and not a 1 in the first component.

vi) Finally, let \(V\) be the set of all \(m \times n\) real matrices. Let + denote the usual matrix addition and \(\cdot\) the multiplication of a matrix with a scalar, then \(V\) is closed under vector addition and scalar multiplication. Hence \(V\) is a vector space and the vectors here are the \(m \times n\) matrices.

When the vectors, scalars and functions are real valued then \(V\) is a real vector space.

Most of our applications will involve real vector spaces.
When complex numbers and functions arise then $V$ is called a complex vector space.

**Subspaces:**

**Definition:** Let $M$ be a subset of $V$. If $M$ itself is closed with respect to the vector addition and scalar multiplication defined for $V$ then $M$ is a subspace of $V$.

**Examples** (in all cases $F = \mathbb{R}$):

i) $M = \{x : x = (x_1, 0, x_3)\}$ is a subspace of $\mathbb{R}_3$

ii) $M = \{f \in C^1[0, 1] : f(0) = 0\}$ is a subspace of $C^0[0, 1]$.

iii) $M = \{\text{all functions in } C^0[0, 1] \text{ which you can integrate analytically}\}$ form a subspace of $C^0[0, 1]$.

iv) $M = \{\text{all functions in } C^0[0, 1] \text{ which you cannot integrate analytically}\}$ do not form a subspace of $C^0[0, 1]$ because a subspace has to contain 0 (i.e., $f \equiv 0$) and you know how to integrate the zero function.

v) $\mathbb{R}_2$ is not a subspace of $\mathbb{R}_3$ because $\mathbb{R}_2$ is not a subset of $\mathbb{R}_3$. On the other hand, $M = \{x : x = (x_1, x_2, 0)\}$ is a subspace of $\mathbb{R}_3$.

vi) $M = \{\text{all polynomials of degree } < N\}$ is a subspace of $C^k(-\infty, \infty)$ for any integer $k > 0$.

vii) Let $\{x_1, x_2, x_3, \ldots, x_K\}$ be $K$ given vectors in a vector space $V$.

Let $M$ be the set of all linear combinations of these vectors, i.e., $m \in M$ if

$$m = \sum_{j=1}^{K} \alpha_j x_j \quad \text{for } \{\alpha_j\} \subset F.$$ 

Then $M$ is a subspace of $V$.

The previous example fits this setting if the vector $x_j$ is identified with the function $f(t) = t^{j-1}$ for $j = 1, \ldots, K$ where $K = N + 1$.

The last example will now be discussed at greater length.

**Definition:** Let $V$ be a vector space, let $\{x_1, \ldots, x_n\}$ be a set of $n$ vectors in $V$. Then the span of these vectors is the subspace of all their linear combinations, i.e.,

$$\text{span}\{x_1, \ldots, x_n\} = \left\{ x : x = \sum_{j=1}^{n} \alpha_j x_j \right\} \quad \text{for } \alpha_j \in F.$$
For example, if \( \hat{e}_1 = (1, 0, 0) \), \( \hat{e}_2 = (0, 1, 0) \) and \( \hat{e}_3 = (0, 0, 1) \) then \( \text{span}\{\hat{e}_1, \hat{e}_2, \hat{e}_3\} = \mathbb{R}^3 \). If \( x \) is a given vector in \( \mathbb{R}^3 \) then \( \text{span}\{x\} \) is the straight line through 0 with direction \( x \). If \( x_1 = (1, 2, 3) \) and \( x_2 = (2, -1, 4) \), then

\[
x = \alpha_1 x_1 + \alpha_2 x_2
\]

for arbitrary \( \alpha_1 \) and \( \alpha_2 \) is just the parametric representation of the plane

\[
11x + 2y - 5z = 0
\]

so that \( \text{span}\{x_1, x_2\} \) is this plane in \( \mathbb{R}^3 \).

We note that if any \( x_k \) is a linear combination of the remaining vectors \( \{x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n\} \) then

\[
\text{span} \{x_i\}_{i=1}^n = \text{span} \{x_i\}_{i=1, i \neq k}^n.
\]
Module 1 - Homework

1) In each case assume that \( F = \mathbb{R} \). Prove or disprove that \( \mathcal{M} \) is a subspace of \( V \).
   
i) \( V = \mathbb{R}^n \)
   \[ \mathcal{M} = \left\{ x : \sum_{j=1}^{n} x_j = 0 \right\} \]

   ii) \( V = \mathbb{R}^n \)
   \[ \mathcal{M} = \left\{ x : \sum_{j=1}^{k} jx_j = 0 \right\} \]
   for some given \( k < n \).

   iii) \( V = \mathbb{R}^3 \)
   \[ \mathcal{M} = \left\{ x : x_1x_2x_3 = 0 \right\} \]

   iv) \( V = \mathbb{R}^3 \)
   \[ \mathcal{M} = \left\{ x : e^{x_1^2+x_2^2+x_3^2} = 1 \right\} \]

   v) \( V = \mathbb{R}^3 \)
   \[ \mathcal{M} = \left\{ x : |x_1| = |x_2| \right\} \]

2) Let \( V = C^0[-\pi, \pi] \). Let \( \mathcal{M} \) be the subspace given by
   \[ \mathcal{M} = \text{span}\{1, \cos t, \cos 2t, \ldots, \cos Nt, \sin t, \sin 2t, \ldots, \sin Nt\} \]

   For given \( f \in V \) define
   \[ Pf(t) = \sum_{j=0}^{N} \alpha_j \cos jt + \sum_{j=1}^{N} \beta_j \sin jt \]
   where
   \[ \alpha_j = \frac{\int_{-\pi}^{\pi} f(t) \cos jt \, dt}{\int_{-\pi}^{\pi} \cos^2 jt \, dt} , \quad \beta_j = \frac{\int_{-\pi}^{\pi} f(t) \sin jt \, dt}{\int_{-\pi}^{\pi} \sin^2 jt \, dt} \]

   Compute \( Pf(t) \) when
   
i) \( f(t) = t \)
   
   ii) \( f(t) = t^2 \)
   
   iii) \( f(t) = \sin 5t \)
   
   iv) \( f(t) = e^t \).
MODULE 2

Topics: Linear independence, basis and dimension

We have seen that if in a set of vectors one vector is a linear combination of the remaining vectors in the set then the span of the set is unchanged if that vector is deleted from the set. If no one vector can be expressed as a combination of the remaining ones then the vectors are said to be linearly independent. We make this concept formal with:

**Definition:** The vectors \( \{x_1, x_2, \ldots, x_n\} \in V \) are linearly independent if

\[
\sum_{j=1}^{n} \alpha_j x_j = 0
\]

has only the trivial solution \( \alpha_1 = \alpha_2 = \alpha_3 = \cdots = \alpha_n = 0 \).

We note that this condition says precisely that no one vector can be expressed as a combination of the remaining vectors. For were there a non-zero coefficient \( \alpha_k \) then \( x_k \) can be expressed as a linear combination of the remaining vectors. In this case the vectors are said to be linearly dependent.

**Examples:**

1) Given \( k \) vectors \( \{x_i\}_{i=1}^{k} \) with \( x_i \in \mathbb{R}_n \) then they are linearly dependent if the linear system

\[
\alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_k x_k = 0,
\]

which can be written as

\[
A\vec{\alpha} = 0,
\]

has a nontrivial solution. Here \( A \) is the \( n \times k \) matrix whose \( j \)th column is the vector \( x_j \) and \( \vec{\alpha} = (\alpha_1, \ldots, \alpha_k) \).

2) The \( k + 1 \) vectors \( \{t^j\}_{j=0}^{k} \) are linearly independent in \( C^n(-\infty, \infty) \) for any \( n \) because if

\[
\sum_{j=0}^{k} \alpha_j t^j \equiv 0 \quad \text{for all } t
\]

then evaluating the polynomial and its derivatives at \( t = 0 \) shows that all coefficients vanish.
3) The functions \( \sin t, \cos t \) and \( \cos(3 - t) \) are linearly dependent in \( C^k[a, b] \) for all \( k \) because they are \( k \) times continuously differentiable and

\[
\cos(3 - t) = \cos 3 \cos t + \sin 3 \sin t.
\]

The first example shows that a check for linear independence in \( \mathbb{R}_n \) or \( \mathbb{C}_n \) reduces to solving a linear system of equations

\[
A\vec{\alpha} = 0
\]

which either has or does not have a nontrivial solution. The test for linear dependence in a function space seems more ad-hoc. Two consistent approaches to obtain a partial answer in this case are as follows.

Let \( \{f_1, \ldots, f_n\} \) be \( n \) given functions in the (function) vector space \( V = C^{n-1}(a, b) \). We consider the arbitrary linear combination

\[
H(t) \equiv \alpha_1 f_1(t) + \alpha_2 f_2(t) + \cdots + \alpha_n f_n(t).
\]

If \( H(t) \equiv 0 \) then the derivatives \( H^{(j)}(t) \equiv 0 \) for \( j = 0, 1, \ldots, n-1 \). We can write these equations in matrix form

\[
W(t)\vec{\alpha} = \begin{pmatrix}
    f_1(t) & f_2(t) & f_n(t) \\
    f_1'(t) & f_2'(t) & f_n'(t) \\
    \vdots & \vdots & \vdots \\
    f_1^{(n-1)}(t) & f_2^{(n-1)}(t) & f_n^{(n-1)}(t)
\end{pmatrix}
\begin{pmatrix}
    \alpha_1 \\
    \alpha_2 \\
    \vdots \\
    \alpha_n
\end{pmatrix} = 0.
\]

If the matrix \( W \) is non-singular at one point \( t \) in the interval then necessarily \( \vec{\alpha} = (\alpha_1, \ldots, \alpha_n) = (0, \ldots, 0) \) and the functions are linearly independent. However, a singular \( W \) at a point (even zero everywhere on \( (a, b) \)) does not in general imply linear dependence. We note that in the context of ordinary differential equations the determinant of \( W \) is known as the Wronskian of the functions \( \{f_i\} \). Hence if the Wronskian is not zero at a point then the functions are linearly independent.

The second approach is to evaluate \( H(t) \) at \( n \) distinct points \( \{t_i\} \). If \( H(t_i) = 0 \) for all \( i \) implies \( \vec{\alpha} = 0 \) then the functions \( \{f_j\} \) are necessarily linearly independent. Written in matrix form we obtain

\[
A\vec{\alpha} = 0
\]
where $A_{ij} = f_j(t_i)$. Hence if $A$ is not singular then we have linear independence. As in the other test, a singular $A$ does not guarantee linear dependence.

**Definition:** Let $\{x_1, \ldots, x_n\}$ be a set of linearly independent vectors in the vector space $V$ such that

$$\text{span}\{x_1, \ldots, x_n\} = V$$

Then $\{x_1, \ldots, x_n\}$ is a basis of $V$.

**Definition:** The number of elements in a basis of $V$ is the dimension of $V$. If there are infinitely many linearly independent elements in $V$ then $V$ is infinite-dimensional.

**Theorem:** Let $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ be bases of the vector space $V$ then $m = n$, i.e., the dimension of the vector space is uniquely defined.

**Proof:** Suppose that $m < n$. Since $\{x_j\}$ is a basis we have

$$y_j = \sum_{i=1}^{m} \alpha_{ij} x_i \quad \text{for } j = 1, 2, \ldots, n.$$ 

Since the matrix $A = (\alpha_{ij})$ has fewer rows than columns, Gaussian elimination shows that there is a non-zero solution $\beta = (\beta_1, \ldots, \beta_n)$ of the non-square system

$$A\beta = 0.$$

But then

$$\sum_{j=1}^{n} \beta_j y_j = \sum_{i=1}^{m} \left( \sum_{j=1}^{n} \alpha_{ij} \beta_j \right) x_i = 0$$

which contradicts the linear independence of $\{y_j\}$.

**Examples:**

i) If $\{x_1, \ldots, x_k\}$ are linearly independent then these vectors are a basis of $\text{span}\{x_1, \ldots, x_k\}$ which has dimension $k$. In particular, the unit vectors $\{\hat{e}_i\}$, $1 \leq i \leq n$, where $\hat{e}_i = (0, 0, \ldots, 1, 0, \ldots, 0)$ with a 1 in the $i$th coordinate, is a (particularly convenient) basis of $\mathbb{R}_n$ or $\mathbb{C}_n$.

ii) The vectors $x = (1, 2, 3)$ and $x = (2, -1, 4)$ are linearly independent because one is not a scalar multiple of the other; hence they form a basis for the plane $11x + 2y - 5z = 0$, i.e. for the subspace of all vectors in $\mathbb{R}_3$ whose components satisfy the equation of the plane.
iii) The vectors $x_i = t^i$, $i = 0, \ldots, N$ form a basis for the subspace of all polynomials of degree $\leq N$ in $C^k(-\infty, \infty)$ for arbitrary $k$. Since $N$ can be any integer, the space $C^k(-\infty, \infty)$ contains countably many linearly independent elements and hence has infinite dimension.

iv) Any set of $n$ linearly independent vectors in $\mathbb{R}_n$ is a basis of $\mathbb{R}_n$.

The norm of a vector:

The norm of a vector $x$, denoted by $\|x\|$, is a real valued function which describes the “size” of the vector. To be admissible as a norm we require the following properties:

i) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.

ii) $\|\alpha x\| = |\alpha|\|x\|$, $\alpha \in F$.

iii) $\|x + y\| \leq \|x\| + \|y\|$ (the triangle inequality).

Certain norms are more useful than others. Below are examples of some commonly used norms:

Examples:

1) Setting: $V = \mathbb{R}_n$ and $F = \mathbb{R}$ or $V = \mathbb{C}_n$ and $F = \mathbb{C}$

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

   ii) $\|x\|_2 = \left(\sum_{j=1}^n |x_j|^2\right)^{1/2}$

   iii) $\|x\|_1 = \sum_{j=1}^n |x_j|$

   iv) Let $C$ be any non-singular $n \times n$ matrix then

   \[ \|x\|_C = \|Cx\|_\infty \]

2) Setting: $V = C^0[a, b]$, $F = \mathbb{R}$

   v) $\|f\|_\infty = \max_{a \leq t \leq b} \|f(t)\|$

   vi) $\|f\|_2 = \left(\int_a^b |f(t)|^2 dt\right)^{1/2}$

   vii) $\|f\|_1 = \int_a^b |f(t)| dt$

To show that 1-i) satisfies the conditions for a norm consider:

\[ \|x\|_\infty > 0 \quad \text{for } x \neq 0 \quad \text{and} \quad \|0\|_\infty = 0 \quad \text{by inspection} \]
\[ \|\alpha x\|_\infty = \max_i |\alpha x_i| = |\alpha| \max_i |x_i| = |\alpha|\|x\|_\infty \]
\[ \|x + y\|_\infty = \max_i |x_i + y_i| \leq \max_i \{ |x_i| + |y_i| \} \]
\[ \leq \max_i |x_i| + \max_i |y_i| = \|x\|_\infty + \|y\|_\infty. \]

The verification of the norm properties for 1-iii) and 1-iv) as well as for 2-v) and 2-vii) is also straightforward. However, the triangle inequalities for 1-ii) and 2-vi) are not obvious and will only be considered after we have introduced inner products.

**Examples:**

i) \( \|x\|_2 \) for \( x \in \mathbb{R}^3 \) is just the Euclidean length of \( x \).

ii) \( \|x\|_1 \) is nicknamed the taxicab (or Manhattan norm) of a vector in \( \mathbb{R}^2 \).

iii) \( \|f\|_2 \) is related to the root mean square of \( f \) defined by \( \text{rms}(f) = \|f\|_2 / \sqrt{b-a} \) which is used to describe, e.g., alternating electric current. To give an illustration:

The voltage of a 110V household current is modeled by

\[ E(t) = E_0 \cos(\omega t - \alpha). \]

Let us consider this function as an element of \( C^0[0,T] \) where \( T = 2\pi/\omega \) is one period. Then

\[ \|E\|_2 = \left(\|E\|_2 / \sqrt{T}\right) \sqrt{T} = 110\sqrt{T} \]

where the term in parentheses is the root mean square of the voltage i.e., 110V. Since also, by direct computation, \( \|E\|_2 = E_0 \sqrt{T/2} \) we see that \( E_0 = \sqrt{2} \times 110 \) and hence that the peak voltage is given by

\[ \|E\|_\infty = E_0 = \sqrt{2} \times 110. \]
Module 2 - Homework

1) Let \( x_1 = (i, 3, 1, i) \), \( x_2 = (1, 1, 2, 2) \), \( x_3 = (-1, i, A, 2) \). Prove or disprove: There is a (complex) number \( A \) such that the vectors \( \{x_1, x_2, x_3\} \) are linearly dependent.

2) Plot the set of vectors in \( \mathbb{R}_2 \) for which
   i) \( \|x\|_1 = 1 \)
   ii) \( \|x\|_2 = 1 \)
   iii) \( \|x\|_\infty = 1 \).

3) For a given vector \( x = (x_1, \ldots, x_n) \in \mathbb{C}_n \) with \( |x_i| \leq 1 \) show that
   \[
   f(p) \equiv \|x\|_p = \left( \sum_{j=1}^{n} |x_j|^p \right)^{1/p}
   \]
   is a decreasing function of \( p \) for \( p \in [1, \infty) \). Show that this result is consistent with your plots of homework problem 2.

4) Find an element \( x \) in span\( \{1, t\} \subset C^0_0(0, 1) \) such that
   \[
   \|x\|_1 = 1 \quad \|x\|_2 = 1.
   \]
   Is this element uniquely determined or are there many such elements?

5) Show that the functions \( e^{\alpha t} \) and \( e^{\beta t} \) are linearly independent in \( C^0_0(-5, 5) \) for \( \alpha \neq \beta \).

6) Prove or disprove: The functions \( f_1 = 1 + t \), \( f_2 = 1 + t + t^2 \) and \( f_3 = 1 - t^2 \) are linearly dependent in \( C^0_0(-\infty, \infty) \).

7) Let \( f(t) = \max\{0, t^3\} \) and \( g(t) = \min\{0, t^3\} \).
   i) Show that these functions are linearly independent in \( C^0_0(-\infty, \infty) \).
   ii) Show that the Wronskian of these two functions is always zero.

8) Let \( f(t) = t(1 - t) \) and \( g(t) = t^2(1 - t) \). Let \( H(t) \) denote a linear combination of \( f \) and \( g \). Show that \( H(0) = H(1) = 0 \) but that the functions are linearly independent.
Topics: Inner products

The inner product of two vectors: $x, y \in V$, denoted by $\langle x, y \rangle$ is (in general) a complex valued function which has the following four properties:

i) $\langle x, y \rangle = \overline{\langle y, x \rangle}$

ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$

iii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$ where $z$ is any other vector in $V$

iv) $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

We note from these properties that:

$\langle 0, y \rangle = 0$ for all $y$

$\langle x, \alpha y \rangle = \langle \alpha x, y \rangle = \alpha \langle x, y \rangle = \overline{\alpha} \langle x, y \rangle$

and

$\langle x, y + z \rangle = \overline{\langle y + z, x \rangle} = \overline{\langle y, x \rangle + \langle z, x \rangle} = \langle x, y \rangle + \langle x, z \rangle$

For real vector spaces the complex conjugation has no effect so that the inner product is linear in the first and second argument, i.e.

$\langle x + \alpha y, z \rangle = \langle x, z \rangle + \alpha \langle y, z \rangle$

$\langle x, y + \alpha z \rangle = \langle x, y \rangle + \alpha \langle x, z \rangle$.

For complex scalars we have what is sometimes called anti-linearity in the second argument.

$\langle x, y + \alpha z \rangle = \langle x, y \rangle + \overline{\alpha} \langle x, z \rangle$.

Examples:

1) Setting: $V = \mathbb{R}_n$ and $F = \mathbb{R}$

   i) $\langle x, y \rangle = \sum_{j=1}^{n} x_j y_j$
       This is the familiar dot product and we shall often use the more familiar notation $x \cdot y$ instead of $\langle x, y \rangle$.

   ii) $\langle x, y \rangle = \sum_{j=1}^{n} x_j y_j w_j$ where $w_i > 0$ for $i = 1, \ldots, n$.  

Note that this inner product can be written as
\[ \langle x, y \rangle = W x \cdot y \]
where \( W \) is the diagonal matrix \( W = \text{diag}\{w_1, \ldots, w_n\} \).

iii) \( \langle x, y \rangle = C x \cdot y \) where \( C \) is a positive definite symmetric matrix. The proof that this defines an inner product will be deferred until we have discussed matrices and their eigenvalues.

2) Setting: \( V = \mathbb{C}_n \) and \( F = \mathbb{C} \)
\[ \langle x, y \rangle = \sum_{j=1}^{n} x_j \overline{y}_j \]
This is the complex dot product and likewise is more commonly denoted by \( x \cdot y \). Note that the order of the arguments now matters because the components of \( y \) are conjugated.

3) Setting: \( V = C^0[D], F = \mathbb{C} \)
\[ \langle f, g \rangle = \int_D f(x) \overline{g(x)} w(x) dx \]
where the so-called weight function \( w \) is continuous and positive except at isolated points of the given (possibly multi-dimensional) domain \( D \). For real valued functions the conjugation has no effect and \( \langle f, g \rangle = \langle g, f \rangle \).

In general when checking whether a function defined on pairs of vectors is an inner product properties ii) and iii) are easy to establish. Properties i) and iv) may require more work. For example, let us define on \( \mathbb{R}_2 \) the function
\[ \langle x, y \rangle = A x \cdot y \]
where \( A \) is the matrix
\[ A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \]
Then
\[ \langle x, y \rangle = x_1 y_1 + x_2 y_1 + x_2 y_2 \]
while
\[ \langle y, x \rangle = Ayx = y_1 x_1 + y_2 x_1 + x_2 y_2. \]
Hence $\langle x, y \rangle \neq \langle y, x \rangle$ for all $x$ and $y$ (take, for example, $x = (1, 0)$ and $y = (1, 1)$) and property i) does not hold. Looking ahead, if for vectors $x, y \in \mathbb{R}^n$ we require that $\langle x, y \rangle = Ax \cdot y = \langle y, x \rangle = Ay \cdot x$ for a given real $n \times n$ matrix $A$ then $A$ must be a symmetric matrix, i.e., $A = A^T$. However, $A = A^T$ is not sufficient to make $\langle x, y \rangle$ an inner product. For example, if $A = \text{diag}\{1, -1\}$ then $A$ is symmetric but $\langle x, x \rangle = Ax \cdot x = 0$ for the non-zero vector $x = (1, 1)$. Thus, property iv) does not hold. It turns out that this matrix $A$ is not positive definite.

As a final example consider the following function defined on $C^0[-2, 2]$

$$\langle f, g \rangle = \int_{-2}^{2} f(t)g(t)w(t)dt$$

where $w(t) = \max\{0, t^2 - 1\}$. We see that properties i)–iii) hold but that for the non-zero function $f(t) = \max\{0, 1 - t^2\}$ we obtain $\langle f, f \rangle = 0$. Clearly, the weight function $w$ may not be zero over any interval.

The following theorem shows that inner products satisfy an important inequality which is known as Schwarz’s inequality.

**Theorem:** Let $V$ be a vector space over $\mathbb{C}$ with inner product $\langle x, y \rangle$. Then for any $x, y \in V$ we have

$$|\langle x, y \rangle|^2 \leq \langle x, x \rangle \langle y, y \rangle.$$ 

**Proof:** Let $x$ and $y$ be arbitrary in $V$. If $y = 0$ then the inequality is trivially true. Hence let us assume that $y \neq 0$. Next let us choose $\theta$ such that for $\hat{x} = e^{i\theta}x$ the inner product $\langle \hat{x}, y \rangle$ is real. Then from the properties of the inner product we see that

$$g(\lambda) = \langle \hat{x} - \lambda y, \hat{x} - \lambda y \rangle = \langle x, x \rangle - 2\lambda \langle \hat{x}, y \rangle + \lambda^2 \langle y, y \rangle \geq 0$$

for all real $\lambda$. In particular, the minimum of $g$ is $\geq 0$. This minimum is achieved at

$$\lambda_\theta = \langle \hat{x}, y \rangle / \langle y, y \rangle$$

so that

$$g(\lambda_\theta) = \langle \hat{x}, \hat{x} \rangle - \langle \hat{x}, y \rangle^2 / \langle y, y \rangle \geq 0$$

from which we obtain $|\langle \hat{x}, y \rangle| \leq \langle \hat{x}, \hat{x} \rangle \langle y, y \rangle$. Finally, we observe that this inequality remains unchanged when $\hat{x}$ is replaced by $x$ since the phase factor $e^{i\theta}$ drops out.
Two illustrations:

i) It is usually shown in a first course on vectors in $\mathbb{R}^2$ that

$$x \cdot y = \|x\|_2 \|y\|_2 \cos \theta$$

where $\theta$ is the angle between the vectors $x$ and $y$. Since $|\cos \theta| \leq 1$ it follows that

$$|\langle x, y \rangle| \equiv |x \cdot y| \leq \|x\|_2 \|y\|_2.$$

ii) Let $D$ be the triangle with vertices $(0, 0)$, $(1, 0)$, $(0, 1)$. Let $w$ be a continuous function positive function in $C^0[D]$. For any $f, g \in C^0[D]$ define the inner product

$$\langle f, g \rangle = \int_D f(x, y)g(x, y)w(x, y)dx \, dy$$

Then for $\epsilon > 0$ we obtain from Schwarz’s inequality

$$|\langle f, g \rangle|^2 = \langle \sqrt{\epsilon} f, \sqrt{\epsilon} g \rangle = \langle \sqrt{\epsilon} f, \sqrt{\epsilon} g \rangle = \epsilon \int_D f(x, y)^2 w(x, y)dx \, dy \cdot \frac{1}{\epsilon} \int_D g(x, y)^2 w(x, y)dx \, dy$$

**Theorem:** Let $V$ be a vector space with inner product $\langle \cdot, \cdot \rangle$. Then

$$\|x\| = \langle x, x \rangle^{1/2}$$

is a norm on $V$.

**Proof.** Properties i) and ii) of the norm are a direct consequence of properties ii) and iv) of the inner product. To establish the triangle inequality we observe that

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \langle x, x \rangle + 2 \text{Re} \langle x, y \rangle + \langle y, y \rangle \leq \|x\|^2 + 2|\langle x, y \rangle| + \|y\|^2.$$

By Schwarz’s inequality $|\langle x, y \rangle| \leq \|x\| \|y\|$ so that $\|x + y\|^2 \leq (\|x\| + \|y\|)^2$ which establishes the triangle inequality and hence that $\langle x, x \rangle^{1/2} = \|x\|$ is a norm on $V$. For example,

$$\|f\| = \left( \int_D f(x, y)^2 w(x, y)dx \, dy \right)^{1/2}$$

defines a norm on the vector space $C^0[D]$ provided $w$ is a positive weight function.
**Definition:** Two vectors $x, y$ in a vector space $V$ with inner product $\langle x, y \rangle$ are orthogonal if $\langle x, y \rangle = 0$.

**Examples:**

i) In $\mathbb{R}^2$ with inner product $x \cdot y$ two vectors are orthogonal if they are perpendicular to each other because $x \cdot y = \|x\|\|y\| \cos \theta$.

ii) Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$ and define $\langle x, y \rangle = Ax \cdot y$. Take for granted for the time being that $\langle x, y \rangle$ is an inner product. Let $x = (1, 0)$, then $y = (1, -2)$ satisfies $\langle x, y \rangle = 0$ and hence is orthogonal to $x$. Note that orthogonality in this case says nothing about the angle between the vectors.

iii) Let $\langle f, g \rangle = \int_{-\pi}^{\pi} f(t)g(t)dt$ be the inner product on $C^0(-\pi, \pi)$ then the functions $\{\cos nt, \sin nt\}$ are all mutually orthogonal.

iv) The functions $t^{2k}$ and $t^{2n+1}$ for any non-negative integers $k$ and $n$ are also mutually orthogonal with respect to the inner product of iii).

We have repeatedly introduced norms and inner products expressed in terms of integrals for continuous functions defined on some set $D$. However, integrals remain defined for much more general functions, for example, functions with certain discontinuities. In fact, even the notion of the integral can be extended beyond the concept of the Riemann integral familiar to us from calculus. We shall introduce and use routinely the following vector (i.e., function) space.

**Definition:** $L_2(D) = \{\text{all functions defined on } D \text{ such that } \int_D |f(x)|^2dx < \infty\}$. We observe that if $f \in L_2(D)$ then $\alpha f \in L_2(D)$. If $f, g \in L_2(D)$ then it follows from

$$\int_D |f(x) + g(x)|^2dx = \int_D (|f(x)|^2 + 2|f(x)g(x)| + |g(x)|^2)dx$$

and the algebraic-geometric mean inequality

$$2|f(x)g(x)| \leq |f(x)|^2 + |g(x)|^2$$

that

$$\int_D |f(x) + g(x)|^2dx \leq 2 \left( \int_D |f(x)|^2dx + \int_D |g(x)|^2dx \right) < \infty$$

so that $L_2(D)$ is closed under vector addition and scalar multiplication and hence a vector space.
We observe that $C^k[D]$ for any $k > 0$ is a subspace of $L_2(D)$ provided that $D$ is a bounded set. Finally, we note that

$$\langle f, g \rangle = \int_D f(x)\overline{g(x)}\,dx \quad \text{and} \quad \|f\| = (\langle f, f \rangle)^{1/2}$$

define the inner product and norm usually associated with $L_2(D)$.

In general, the functions will be real and the conjugation can be ignored. On occasion the inner product and norm are modified by including a weight function $w(x) > 0$ in the integral. In this course the integral will remain to be the Riemann integral. In a more abstract setting it should be the Lebesgue integral.
Module 3 - Homework

1) \( V = \mathbb{C}_n, \ F = \mathbb{C}, \ \langle x, y \rangle = x \cdot y. \)

Let \( x = (x_1, \ldots, x_n) \) where \( x_j = j; \)

\( y = (y_1, \ldots, y_n) \) where \( y_j = (1+i)j \) and \( i^2 = -1. \)

Compute \( \langle x, y \rangle, \langle y, x \rangle, \langle x, x \rangle, \langle y, y \rangle. \)

2) i) Let \( A \) be an \( m \times n \) real matrix and \( A^T \) its transpose. Show that

\[ Ax \cdot y = x \cdot A^T y \quad \text{for all } x \in \mathbb{R}_n \quad \text{and } y \in \mathbb{R}_m. \]

ii) Let \( A \) be an \( m \times n \) complex matrix and \( A^* \) its conjugate transpose (i.e., \( A^* = \overline{A^T} \)). Show that

\[ Ax \cdot y = x \cdot A^* y \quad \text{for all } x \in \mathbb{C}_n \quad \text{and } y \in \mathbb{C}_m. \]

3) Let

\[ A = \begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix} \]

and

\[ \langle x, y \rangle = Ax \cdot y \]

for \( x, y \in \mathbb{R}_2. \) Show that \( \langle , \rangle \) defines an inner product on \( \mathbb{R}_2. \)

4) Prove or disprove:

\[ \langle f, g \rangle = \int_0^1 f(t)g(t)dt \]

is an inner product on \( C^0[-1, 1]. \)

5) Show that if \( w \) is a positive continuous function on \([a, b]\) then

\[ \left| \int_a^b f(t)w(t)dt \right| \leq \left( \int_a^b w(t)dt \int_a^b f(t)^2w(t)dt \right)^{1/2} \]

and in particular that

\[ \int_a^b f(t)dt \leq \sqrt{b-a} \sqrt{\int_a^b f(t)^2dt}. \]
Topics: Orthogonal projections

**Definition:** Let $V$ be an inner product space over $F$. Let $M$ be a subspace of $V$. Given an element $y \in V$ then the orthogonal projection of $y$ onto $M$ is the vector $Py \in M$ which satisfies

$$y = Py + v$$

where $v$ is orthogonal to every element $m \in M$ (in short: $v$ is orthogonal to $M$).

The orthogonal projection, if it exists, is uniquely defined because if

$$y = Py_1 + v_1$$

and

$$y = Py_2 + v_2$$

then by subtracting we find that

$$Py_1 - Py_2 = -(v_1 - v_2).$$

Since $Py_1 - Py_2$ is an element of $M$ and $(v_1 - v_2)$ is orthogonal to $M$ it follows from

$$\langle Py_1 - Py_2, Py_1 - Py_2 \rangle = -\langle v_1 - v_2, Py_1 - Py_2 \rangle = 0$$

that $Py_1 = Py_2$.

The existence of orthogonal projections onto infinite dimensional subspaces $M$ is complicated. Hence we shall consider only the case where

$$M = \text{span}\{x_1, \ldots, x_n\}$$

where, in addition, the vectors $\{x_1, \ldots, x_n\}$ are assumed to be linearly independent so that the dimension of $M$ is $n$. In this case $Py$ must have the form

$$Py = \sum_{j=1}^{n} \alpha_j x_j.$$
for some properly chosen scalars \( \{\alpha_j\} \). From the definition of the orthogonal projection
now follows that
\[
\langle y, m \rangle = \left\langle \sum_{j=1}^{n} \alpha_j x_j, m \right\rangle + \langle v, m \rangle = \sum_{j=1}^{n} \alpha_j \langle x_j, m \rangle
\]
for arbitrary \( m \in M \). In particular, it has to be true for \( m = x_i \) for each \( i \), and if it is true for each \( x_i \) then it is true for any linear combination of the \( \{x_j\} \), i.e., it is true for all \( m \in M \). Hence the orthogonal projection is
\[
Py = \sum_{j=1}^{n} \alpha_j x_j
\]
where the \( n \) coefficients \( \{\alpha_j\} \) are determined from the \( n \) equations
\[
\langle y, x_i \rangle = \sum_{j=1}^{n} \alpha_j \langle x_j, x_i \rangle, \quad i = 1, \ldots, n.
\]
In other words, the coefficients \( \{\alpha_j\} \) are found from the matrix system
\[
A\vec{\alpha} = b, \quad \vec{\alpha} = (\alpha_1, \ldots, \alpha_n)
\]
where
\[
A_{ij} = \langle x_j, x_i \rangle
\]
and
\[
b_i = \langle y, x_i \rangle.
\]
The question now arises: Does the orthogonal projection always exist? Or equivalently, can I always solve the linear system
\[
A\vec{\alpha} = b.
\]
The solution \( \vec{\alpha} \) exists and is unique whenever \( A \) is invertible, or what is the same, whenever
\[
A\vec{\beta} = 0
\]
has only the zero solution \( \vec{\beta} = (0, \ldots, 0) \). Suppose that there is a non-zero solution \( \vec{\beta} = (\beta_1, \ldots, \beta_n) \). If we set
\[
w = \sum_{j=1}^{n} \beta_j x_j
\]
then we see by expanding the inner product that

$$\langle w, w \rangle = A\vec{\beta} \cdot \vec{\beta}.$$ 

But this implies that $w = 0$ which contradicts that the $\{x_j\}$ are linearly independent. Hence $A\vec{\beta} = 0$ cannot have a non-zero solution. $A$ is invertible and the orthogonal projection is computable.

**Examples:**

1) Let $M = \text{span}\{(1, 2, 3), (3, 2, 1)\} \in \mathbb{R}^3$. We see that $M$ is the plane in $\mathbb{R}^3$ (through the origin, of course) given algebraically by

$$x_1 - 2x_2 + x_3 = 0.$$ 

Then the projection of the unit vector $\hat{e}_1 = (1, 0, 0)$ onto $M$ is the vector

$$P\hat{e}_1 = \alpha_1 (1, 2, 3) + \alpha_2 (3, 2, 1)$$

where $\alpha_1$ and $\alpha_2$ are found from

$$\begin{pmatrix} \langle x_1, x_1 \rangle & \langle x_2, x_1 \rangle \\ \langle x_1, x_2 \rangle & \langle x_2, x_2 \rangle \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 14 & 10 \\ 10 & 14 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}.$$ 

It follows that

$$P\hat{e}_1 = -\frac{1}{6} (1, 2, 3) + \frac{1}{3} (3, 2, 1) = (\frac{5}{6}, \frac{1}{3}, -\frac{1}{6}).$$

We note that in this case $v = \hat{e}_1 - P\hat{e}_1 = (\frac{1}{6}, -\frac{1}{3}, \frac{1}{6})$ which is perpendicular to the plane $M$ as required.

2) Find the orthogonal projection of the function $f(t) = t^3$ onto

$$M = \text{span}\{1, t, t^2\}$$

when $V = C^0[-1, 1]$ and $\langle f, g \rangle = \int_{-1}^1 f(t)g(t)dt$. In other words, find the projection of $t^3$ onto the subspace of polynomials of degree $\leq 2$.

Answer:

$$P(t^3) = \alpha_0 1 + \alpha_1 t + \alpha_2 t^2$$

21
where

\[
\begin{pmatrix}
\langle 1, 1 \rangle & \langle t, 1 \rangle & \langle t^2, 1 \rangle \\
\langle 1, t \rangle & \langle t, t \rangle & \langle t^2, t \rangle \\
\langle 1, t^2 \rangle & \langle t, t^2 \rangle & \langle t^2, t^2 \rangle
\end{pmatrix}
\begin{pmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2
\end{pmatrix}
= \begin{pmatrix}
\langle t^3, 1 \rangle \\
\langle t^3, t \rangle \\
\langle t^3, t^2 \rangle
\end{pmatrix}.
\]

Carrying out the integrations we find the algebraic system

\[
\begin{pmatrix}
2 & 0 & 2/3 \\
0 & 2/3 & 0 \\
2/3 & 0 & 2/5
\end{pmatrix}
\begin{pmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2
\end{pmatrix}
= \begin{pmatrix}
0 \\
2/5 \\
0
\end{pmatrix}
\]

from which we obtain the orthogonal projection

\[Pt^3 = \frac{3}{5} t\]

The question now arises. Why should one be interested in orthogonal projections?
Module 4 - Homework

1) Find the orthogonal projection with respect to the dot product of the vector \( y = (0, 2) \) onto

\[ M = \text{span}\{(1, 1)\} \]

and draw a picture that makes clear what is

i) \( M \)

ii) \( y \)

iii) \( Py \)

iv) \( v \).

2) Let \( A = \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \) and define for \( \mathbb{R}_2 \)

\[ \langle x, y \rangle = Ax \cdot y. \]

i) Show that \( \langle \cdot, \cdot \rangle \) is an inner product on \( \mathbb{R}_2 \).

ii) Find all vectors which are orthogonal to the vector \((0, 1)\) with respect to this inner product.

iii) Compute the orthogonal projection of the vector \((0, 2)\) onto \( M = \text{span}\{(1, 1)\} \).

Draw a picture as in Problem 1.

3) Let \( V = C^0[0, 2\pi] \), \( \langle f, g \rangle = \int_0^{2\pi} f(t)g(t)dt \). Find the orthogonal projection of \( f(t) \equiv t \) onto \( M = \text{span}\{1, \cos t, \cos 2t\} \).

4) Find the orthogonal projection \( Pf \) in \( L_2(0, 1) \) of \( f(t) \equiv \cos(t - 4) \) onto

\[ M = \text{span}\{\cos t, \sin t\}. \]

Compute

\[ v = f - Pf. \]

5) Let

\[ M_1 = \text{span}\{(1, 2, 1, 2), (1, 1, 2, 2)\} \]

\[ M_2 = \text{span}\{(1, 2, 1, 2), (1, 1, 2, 2), (1, 0, 3, 2)\}. \]

Let \( y = (1, 0, 0, 0) \).

i) With respect to the dot product find the orthogonal projections of \( y \) onto \( M_1 \) and \( M_2 \).
ii) Show that $M_1 = M_2$.

6) Suppose a pole of height 5$m$ stands vertically on a hillside. The elevation of the ground relative to the base of the pole is $+6m$ at a distance of 50$m$ to the east and $-17m$ at a distance 200$m$ south from the pole. Suppose the sun is in the southwest and makes an angle of $\pi/6$ with the pole. Assume that the hillside can be approximated by a plane. Find the vector which describes the shadow of the pole on the ground. What is its length?
It is a common process in the application of mathematics to approximate a given vector by a vector in a specified subspace. For example, a linear system like
\[ Ax = b \]
which does not have a solution, may be approximated by a linear system
\[ Ax = b' \]
for which there does exist a solution and where the vector \( b' \) is chosen so that it is “close” to \( b \). This situation will be discussed later in connection with the least squares solution of linear systems. Another common example is the approximation of a given function \( f \) defined on some interval, say \((-\pi, \pi)\), in terms of a trigonometric sum like
\[ f(t) \sim \alpha_0 + \sum_{j=1}^{N} \alpha_j \cos jt + \sum_{j=1}^{N} \beta_j \sin jt \]
which leads to the concept of Fourier series. As we shall discover, \( b' \) and the trigonometric sum will be a “best” approximation to \( b \) and \( f \), respectively.

**Definition:** Given a vector space \( V \) with norm \( \| \| \) and a subspace \( M \subset V \) the best approximation of a given vector \( x \in V \) in the subspace \( M \) is a vector \( \hat{m} \) which satisfies
\[ \| x - \hat{m} \| \leq \| x - m \| \quad \text{for all } m \in M. \]

Note that the best approximation in this definition is tied to a norm. Changing the norm will usually change the element \( \hat{m} \). The choice of norm is dictated by the application or the desire to compute easily the best approximation.

To illustrate that the best approximation can be easy or hard to find depending on the choice of norm consider the following simply stated problem:

Find the best approximation to the vector \((1, 2, 3)\) in the subspace
\[ M = \text{span}\{(3, 2, 1)\} \subset \mathbb{R}^3. \]

i) when \( \| x \| = \| x \|_\infty \)
ii) when $\|x\| = \|x\|_1$

iii) when $\|x\| = \|x\|_2$.

**Answer:**

i) Since $m = \alpha (3, 2, 1)$ for $\alpha \in (-\infty, \infty)$ the problem is to find an $\alpha$ which minimizes the expression

$$\|(1, 2, 3) - \alpha (3, 2, 1)\|_\infty = \max\{|1 - 3\alpha|, |2 - 2\alpha|, |3 - \alpha|\} \equiv f(\alpha).$$

If one plots $f(\alpha)$ vs. $\alpha$ one finds that it has a minimum at $\hat{\alpha} = 1$ with $f(1) = 2$. Note that this $\hat{\alpha}$ cannot be found with calculus because the function is not differentiable. Hence the best approximation in $M$ in this norm is $\hat{m} = (3, 2, 1)$.

ii) $\|(1, 2, 3) - \alpha (3, 2, 1)\|_1 = |1 - 3\alpha| + |2 - 2\alpha| + |3 - \alpha| \equiv f(\alpha)$. The function $f$ is piecewise linear and constant on the interval $[1/3, 1]$ where it assumes its minimum of $f(1/3) = 4$. Hence the best approximation is not unique. $\hat{m}$ may be chosen to be $\alpha (3, 2, 1)$ for any $\alpha \in [1/3, 1]$.

iii) Since the Euclidean norm involves square roots it is usually advantageous to minimize the square of the norm rather than the norm itself. Thus,

$$\|(1, 2, 3) - \alpha (3, 2, 1)\|_2^2 = (1 - 3\alpha)^2 + (2 - 2\alpha)^2 + (3 - \alpha)^2 \equiv f(\alpha)$$

is minimized where $f'(\alpha) = 0$. A simple calculation shows that $\alpha = 5/7$ so that $\hat{m} = 5/7 (3, 2, 1)$. If we set $v = x - \hat{m}$ then we find by direct calculation that $v \cdot (3, 2, 1) = 0$, so that $v$ is orthogonal to $M$. Hence $\hat{m}$ is the orthogonal projection $Px$ of $x$ onto $M$. According to Module 4 we can calculate the projection as

$$Px = \alpha (3, 2, 1)$$

where

$$\alpha = \frac{(1, 2, 3) \cdot (3, 2, 1)}{(3, 2, 1) \cdot (3, 2, 1)} = \frac{5}{7}$$

which shows that the best approximation is obtainable also without calculus in this case.

The next theorem shows that if the norm is derived from an inner product then the best approximation always is the orthogonal projection.
Theorem: Let $V$ be a vector space (real or complex) with inner product $\langle \cdot , \cdot \rangle$ and norm $\| \cdot \| = (\langle \cdot , \cdot \rangle)^{1/2}$. Let $M = \text{span}\{x_1, \ldots, x_n\} \subset V$ be a subspace of dimension $n$. Given $y \in V$ then $\hat{m} \in M$ is the best approximation to $y$ in $M$ if and only if $\hat{m}$ is the orthogonal projection $Py$ of $y$ onto $M$.

Proof. Let us show first that $Py$ is a best approximation, i.e., that

$$\|y - Py\| \leq \|y - m\| \quad \text{for all } m \in M.$$ 

Let $m$ be arbitrary in $M$. Then $m = Py + (m - Py)$ and

$$\|y - m\|^2 = \langle y - Py - (m - Py), y - Py - (m - Py) \rangle$$

$$= \langle y - Py, y - Py \rangle - \langle m - Py, y - Py \rangle$$

$$- \langle y - Py, m - Py \rangle + \langle m - Py, m - Py \rangle.$$ 

But by definition of the orthogonal projection $y - Py$ is orthogonal to $M$, while $m - Py \in M$. Hence the two middle terms on the right drop out and thus $\|y - m\|^2 = \|y - Py\|^2 + \|m - Py\|^2$ so that

$$\|y - Py\| \leq \|y - m\| \quad \text{for all } m \in M.$$ 

Conversely, suppose that $\hat{m}$ is the best approximation to $y$ in $M$. For arbitrary but fixed $m \in M$ and real $t$ define

$$g(t) = \langle y - (\hat{m} + tm), y - (\hat{m} + tm) \rangle.$$ 

Since $\hat{m} + tm$ is an element of $M$ and $\hat{m}$ is the best approximation it follows that $g$ has a minimum at $t = 0$. Hence necessarily $g'(0) = 0$. Differentiation shows that

$$g'(0) = -[\langle m, y - \hat{m} \rangle + \langle y - \hat{m}, m \rangle] = -2\Re \langle m, y - \hat{m} \rangle = 0.$$ 

Since $m$ is arbitrary in $M$ this implies that $\langle m, y - \hat{m} \rangle = 0$ so that $\hat{m}$ satisfies the definition of an orthogonal projection. Its uniqueness guarantees that $\hat{m} = Py$. Looking back at the examples of Module 4 we see that the function $Pt^3 \equiv \frac{3}{5}t$ is the best approximation to the function $f(t) \equiv t^3$ in the sense that

$$\|t^3 - \frac{3}{5}t\|^2_2 = \int_{-1}^{1} (t^3 - \frac{3}{5}t)^2 dt \leq \int_{-1}^{1} (t^3 - P_2(t))^2 dt$$

for any other polynomial $P_2(t)$ of degree $\leq 2$. In other words, $Pt^3 \equiv \frac{3}{5}t$ is the best polynomial approximation of degree $\leq 2$ to the function $f(t) \equiv t^3$ in the mean square sense.
Module 5 - Homework

1) Let \( y = (1, 2, 3, 4) \) and \( M = \text{span}\{(4, 3, 2, 1)\} \). Find the best approximation \( \hat{m} \in M \) to \( y \)
   
   i) in the \( \| \cdot \|_1 \) norm
   
   ii) in the \( \| \cdot \|_2 \) norm
   
   iii) in the \( \| \cdot \|_\infty \) norm.

2) Let \( M \) be the plane in \( \mathbb{R}_3 \) given by
   
   \[ 3x - 2y + z = 0 \]
   
   Find the orthogonal projection of the unit vector \( \hat{e}_1 \) onto \( M \) when
   
   i) the inner product is the dot product
   
   ii) the inner product is
   
   \[ \langle x, y \rangle = Ax \cdot y \]
   
   where
   
   \[ A = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix} \]
   
   (You may assume at this point without further checking that \( \langle x, y \rangle \) is indeed an inner product).

3) Compute the best approximation in the \( L_2(-\pi, \pi) \) sense of the function

   \[ H(t) = \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases} \]

   in terms of the functions \( \{\sin nt\}_{n=1}^N \) and \( \{\cos nt\}_{n=0}^N \) where \( N > 0 \) is some integer.

4) Let \( 0 = t_0 < t_1 < \ldots < t_N = 1 \) where \( t_i = i\Delta t \) and \( \Delta t = 1/N \). For each \( i = 0, 1, \ldots, N \) define on \([0,1]\) the function

   \[ \phi_i(t) = \begin{cases} \frac{(t-t_{i-1})}{\Delta t} & t \in [t_{i-1}, t_i) \\ \frac{(t_{i+1}-t)}{\Delta t} & t \in [t_i, t_{i+1}) \\ 0 & \text{otherwise} \end{cases} \]

   i) For \( N = 4 \) plot \( \phi_0(t) \) and \( \phi_3(t) \)
Let
\[ M = \text{span}\{\phi_0(t), \phi_1(t), \phi_2(t), \phi_3(t), \phi_4(t)\} \]

We shall consider \( M \) as a subspace of \( L_2[0, 1] \) with inner product
\[ \langle f, g \rangle = \int_0^1 f(t)g(t)dt \]

ii) Compute and plot the orthogonal projection \( Pf \) of the function \( f(t) \equiv 1 + t \) onto \( M \).

iii) Compute and plot the orthogonal projection \( Pf \) of the function \( f(t) \equiv t^2 \) onto \( M \).
MODULE 6

Topics: Gram-Schmidt orthogonalization process

We begin by observing that if the vectors \( \{x_j\}_{j=1}^N \) are mutually orthogonal in an inner product space \( V \) then they are necessarily linearly independent. For suppose that \( \langle x_i, x_j \rangle = 0 \) for \( i \neq j \) and

\[
\sum_{j=1}^{N} \alpha_j x_j = 0
\]

then taking the inner product of this equation with \( x_k \) shows that \( \alpha_k \langle x_k, x_k \rangle = 0 \) so that \( \alpha_k = 0 \). Hence only zero coefficients are possible. We shall now observe the simplification which arises when we compute the orthogonal projection of a vector \( y \) in a subspace with an orthogonal basis. Hence assume that

\[
M = \text{span}\{x_1, \ldots, x_N\} \subset V
\]

and \( \langle x_i, x_j \rangle = 0 \) for \( i \neq j \). Let \( y \) be a given vector in \( V \). According to Module 4 the orthogonal projection \( Py \) in \( M \) is given as

\[
Py = \sum_{j=1}^{N} \alpha_j x_j
\]

where

\[
A\vec{\alpha} = b,
\]

with

\[
A_{ij} = \langle x_j, x_i \rangle
\]

and

\[
b_i = \langle y, x_i \rangle,
\]

has to be solved to obtain the coefficients \( \{\alpha_j\} \). But because the \( \{x_j\} \) are all orthogonal the matrix \( A \) is diagonal so that

\[
\alpha_i = \frac{\langle y, x_i \rangle}{\langle x_i, x_i \rangle}
\]

**Example.** Let \( f \in L_2(-\pi, \pi) \) and let \( M = \text{span}\{1, \cos t, \ldots, \cos Nt, \sin t, \ldots, \sin Nt\} \).
It is straightforward to verify that
\[ \langle \cos mt, \sin nt \rangle = \int_{-\pi}^{\pi} \cos mt \sin nt \, dt = 0 \quad \text{for all } m, n \]
and that
\[ \langle \cos mt, \cos nt \rangle = \langle \sin mt, \sin nt \rangle = 0 \quad \text{for } m \neq n. \]
Hence all the elements spanning \( M \) are mutually orthogonal and therefore linearly independent and thus a basis of \( M \). The orthogonal projection of \( f \) onto \( M \) is then given by
\[ Pf = \sum_{n=0}^{N} \alpha_n \cos nt + \sum_{n=1}^{N} \beta_n \sin nt \]
where the coefficients are found explicitly as
\[ \alpha_n = \frac{\langle f, \cos nt \rangle}{\langle \cos nt, \cos nt \rangle} \quad n = 0, 1, \ldots, N \]
and
\[ \beta_n = \frac{\langle f, \sin nt \rangle}{\langle \sin nt, \sin nt \rangle} \quad n = 1, \ldots, N. \]
These coefficients are known as the Fourier coefficients of \( f \) in \( L_2(-\pi, \pi) \) and \( Pf \) is the \( N \)th partial sum of the Fourier series of \( f \). This partial sum is the best approximation, in the mean square sense, over the interval \((-\pi, \pi)\) in terms of the given sine and cosine functions (i.e., in terms of a so-called trigonometric polynomial).

If in an application the linearly independent vectors \( \{x_j\} \) spanning \( M \) are not orthogonal then it may be advantageous to compute an equivalent basis \( \{z_1, \ldots, z_N\} \) of \( M \) of mutually orthogonal vectors and to express the projection as a linear combination of these new basis vectors. The process of finding the orthogonal basis \( \{z_j\} \) equivalent to the basis \( \{x_j\} \) is known as the Gram-Schmidt orthogonalization process and proceeds recursively as follows.

We set
\[ z_1 = x_1. \]
Assume that for \( j = 1, \ldots, k-1 \) we have found orthogonal vectors \( \{z_j\} \subset \text{span}\{x_j\} \). Then we set
\[ z_k = x_k - \sum_{j=1}^{k-1} \alpha_j z_j \]
where the $\alpha_j$ are computed such that $\langle z_k, z_j \rangle = 0$ for $j = 1, \ldots, k - 1$. Since $\langle z_i, z_j \rangle = 0$ for $i \neq j$ and $i, j < k$ this requires that

$$
\alpha_j = \frac{\langle x_k, z_j \rangle}{\langle z_j, z_j \rangle}.
$$

When $k = N$ we have generated $N$ mutually orthogonal vectors, each of which is obtained as a combination of the basis vectors $\{x_j\}$. Hence $\{z_j\}$ forms an orthogonal basis of $M$.

**Examples:**

1) Let us find an orthogonal basis of $M = \text{span}\{(1, 2, 1, 2), (0, 1, 0, 1), (1, 0, 0, -1)\} \subset E_4$.

The notation $E_4$ implies that the inner product is the dot product. We set

$$z_1 = (1, 2, 1, 2)$$

and compute

$$z_2 = (0, 1, 0, 1) - \alpha_1 (1, 2, 1, 2)$$

where

$$\alpha_1 = \frac{(0, 1, 0, 1) \cdot (1, 2, 1, 2)}{(1, 2, 1, 2) \cdot (1, 2, 1, 2)} = \frac{4}{10}$$

so that

$$z_2 = (-\frac{2}{5}, \frac{1}{5}, -\frac{2}{5}, \frac{1}{5}).$$

Since the span of a set remains unchanged if the vectors are scaled we can simplify the notation in our long-hand calculation by setting

$$z_2 = (-2, 1, -2, 1).$$

Then

$$z_3 = (1, 0, 0, -1) - \alpha_1 (1, 2, 1, 2) - \alpha_2 (-2, 1, -2, 1)$$

where

$$\alpha_1 = \frac{(1, 0, 0, -1) \cdot (1, 2, 1, 2)}{10} = \frac{-1}{10},$$

$$\alpha_2 = \frac{(1, 0, 0, -1) \cdot (-2, 1, -2, 1)}{(-2, 1, -2, 1) \cdot (-2, 1, -2, 1)} = \frac{-3}{10}.$$
so that
\[ z_3 = (1/2, 1/2, -1/2, -1/2). \]

Hence as an orthogonal basis of \( M \) we may take
\[ \{(1, 2, 1, 2), (-2, 1, -2, 1), (1, 1, -1, -1)\}. \]

The orthogonal projection of an arbitrary vector \( y \in E_4 \) onto \( M \) is then given by
\[ Py = \sum_{i=1}^{3} \frac{\langle y, z_i \rangle}{\langle z_i, z_i \rangle} z_i. \]

On occasion the vectors \( z_j \) are scaled so that they are unit vectors in the norm \( \|z\| = \langle z, z \rangle^{1/2} \). The vectors \( \{z_j\} \) are said to be orthonormal in this case.

2) Find an orthogonal basis of \( \text{span}\{1, t, t^2\} \subset L_2(-1, 1) \). Scale the functions such that they assume a value of 1 at \( t = 1 \). The notation indicates that the inner product is
\[ \langle f, g \rangle = \int_{-1}^{1} f(t)g(t)dt. \]

We now apply the Gram-Schmidt process to generate orthogonal functions \( \{\phi_0, \phi_1, \phi_2\} \).

We set
\[ \phi_0(t) \equiv 1 \quad (\text{which already satisfies } \phi(1) = 1) \]
and compute
\[ \phi_1(t) = t - \alpha_0 1 \]
where
\[ \alpha_0 = \frac{\langle t, 1 \rangle}{\langle 1, 1 \rangle} = 0. \]

Hence
\[ \phi_1(t) = t \quad (\text{which also is already properly scaled}). \]

Then
\[ \phi_2(t) = t^2 - \alpha_0 \phi_0(t) - \alpha_1 \phi_1(t) \]
with
\[ \alpha_0 = \frac{\langle t^2, 1 \rangle}{\langle 1, 1 \rangle} = \frac{1}{3}. \]
and
\[ \alpha_1 = \frac{\langle t^2, t \rangle}{\langle t, t \rangle} = 0. \]

Hence a function orthogonal to \( \phi_0(t) \) and \( \phi_1(t) \) is \( c(t^2 - \frac{1}{3}) \) for any constant \( c \). To insure that \( \phi_2(1) = 1 \) we choose
\[ \phi_2(t) = \frac{1}{2} (3t^2 - 1). \]

These three orthogonal polynomials are known as the first three Legendre polynomials which arise, for example, in the solution of the heat equation in spherical coordinates. Legendre polynomials of order up to \( N \) can be found by applying the Gram-Schmidt process to the linearly independent functions \( \{1, t, \ldots, t^N\} \) with the \( L_2 \) inner product.

We shall conclude our discussion of projections and best approximations with a problem which, strictly speaking, leads neither to a projection nor a best approximation, but which has much the same flavor as the material presented above.

Suppose we are in a vector space \( V \) with inner product \( \langle \ , \ \rangle \) and its associated norm. Suppose further that we wish to find an element \( u \in V \) which satisfies the following \( N \) constraint equations
\[ \langle u, x_i \rangle = b_i, \quad i = 1, \ldots, N \]
where the vectors \( \{x_1, \ldots, x_N\} \) are assumed to be linearly independent in \( V \). In general, the solution to this problem will not be unique. The solution does exist and is unique if we restrict it to lie in
\[ M = \text{span}\{x_1, \ldots, x_n\}. \]

In this case \( u \) has to have the form
\[ u = \sum_{j=1}^{N} \alpha_j x_j. \]

Substitution into the \( N \) equations shows that \( \{\alpha_j\} \) is a solution of the system
\[ A\vec{\alpha} = b \]
where as before
\[ A_{ij} = \langle x_j, x_i \rangle. \]
Linear independence of the \( \{x_j\} \) guarantees that \( A \) is invertible and that therefore the \( \{\alpha_j\} \) are uniquely defined.

**Theorem:** Let \( y \) be any vector in \( V \) which satisfies the \( N \) constraint equations and let \( u \) be the specific solution which belongs to \( M \) then

\[
\|u\| \leq \|y\|.
\]

**Proof:** For any \( y \) we can write \( y = y - u + u \). Then

\[
\|y\|^2 = \langle y - u + u, y - u + u \rangle = \langle y - u, y - u \rangle + 2\text{Re}\langle y - u, u \rangle + \langle u, u \rangle.
\]

But

\[
\langle y - u, u \rangle = \sum_{j=1}^{N} \alpha_j \langle y - u, x_j \rangle = 0 \quad \text{because} \quad \langle y, x_j \rangle = b_j = \langle u, x_j \rangle.
\]

Hence

\[
\|y\|^2 = \|y - u\|^2 + \|u\|^2 > \|u\|^2 \quad \text{for} \quad y \neq u.
\]

Problems of this type arise in the theory of optimal controls for linear state equations and quadratic cost functionals. We shall not pursue this subject here but instead consider the following simpler geometric problem.

**Problem:** Let \( \vec{x} = (x, y) \) be a point in \( \mathbb{E}_3 \). Find the point in the intersection of the planes

\[
x + 2y + 3z = 1
\]

\[
3x + 2y + z = 5
\]

which is closest to the point \((0, 1, 0)\).

**Answer:** The geometry is clear. These are two planes which are not parallel to each other. Hence they intersect in a line. The problem then is to find the point on the line which is closest to \((0, 1, 0)\). Since the setting is \( \mathbb{E}_3 \), closest means closest in Euclidean distance.

We shall examine three different approaches to solving this problem.

1) If we define \( u = \vec{x} - (0, 1, 0) \) then we want the minimum norm \( u \) which satisfies

\[
\langle u, (1, 2, 3) \rangle = \langle \vec{x}, (1, 2, 3) \rangle - \langle (0, 1, 0), (1, 2, 3) \rangle = -1
\]

\[
\langle u, (3, 2, 1) \rangle = \langle \vec{x}, (3, 2, 1) \rangle - \langle (0, 1, 0), (3, 2, 1) \rangle = 3.
\]
According to the last theorem the minimum norm solution is the uniquely defined solution belonging to $M = \text{span}\{(1, 2, 3), (3, 2, 1)\}$. It is computed as

$$u = \alpha_1(1, 2, 3) + \alpha_2(3, 2, 1)$$

where $\{\alpha_1, \alpha_2\}$ are found from the linear system

$$\begin{pmatrix} \langle (1, 2, 3), (1, 2, 3) \rangle & \langle (3, 2, 1), (1, 2, 3) \rangle \\ \langle (1, 2, 3), (3, 2, 1) \rangle & \langle (3, 2, 1), (3, 2, 1) \rangle \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 14 & 10 \\ 10 & 14 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \end{pmatrix}.$$  

Doing the arithmetic we find that

$$(x, y, z) = (7/6, 7/6, -5/6).$$

ii) We shall find the equation of the line of intersection. Subtracting the equation of one plane from the other we find that coordinates of points in the intersection must satisfy

$$x - z = 2.$$  

We set $z = s$. Then $x = 2 + s$ and $y = -1/2 - 2s$. Hence the line is

$$(x, y, z) = (2, -1/2, 0) + s(1, -2, 1)$$

i.e., the line through $(2, -1/2, 0)$ with direction $(1, -2, 1)$. The square of the distance from $(0, 1, 0)$ to a point on the line is

$$g(s) = (2 + s)^2 + (-3/2 + 2s)^2 + s^2.$$  

This function is minimized for $s = -5/6$ yielding

$$(x, y, z) = (7/6, 7/6, -5/6).$$

iii) The problem can be solved with Lagrange multipliers. We want to minimize

$$g(\vec{x}) = g(x, y, z) = \langle \vec{x} - (0, 1, 0), \vec{x} - (0, 1, 0) \rangle.$$  

subject to the constraints

$$\langle \vec{x}, (1, 2, 3) \rangle - 1 = 0$$
The Lagrangian is

\[ \mathcal{L} = x^2 + (y - 1)^2 + z^2 + \lambda_1(x + 2y + 3z - 1) + \lambda_2(3x + 2y + z - 5). \]

The minimizer has to satisfy

\[ \frac{\partial \mathcal{L}}{\partial x} = \frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \mathcal{L}}{\partial z} = 0 \]

as well as the constraint equations. This leads to the linear system

\[
\begin{pmatrix}
2 & 0 & 0 & 1 & 3 \\
0 & 2 & 0 & 2 & 2 \\
0 & 0 & 2 & 3 & 1 \\
1 & 2 & 3 & 0 & 0 \\
3 & 2 & 1 & 0 & 0
\end{pmatrix}
\begin{pmatrix}
x \\
y \\
z \\
\lambda_1 \\
\lambda_2
\end{pmatrix}
=
\begin{pmatrix}
0 \\
2 \\
0 \\
1 \\
5
\end{pmatrix}
\]

which again has the solution

\[(x, y, z) = \left( \frac{7}{6}, \frac{7}{6}, -\frac{5}{6} \right). \]
Module 6 - Homework

1) Consider the vectors \{ (1, 1, 0), (1, 0, 1), (0, 1, 1), (1, 1, 1) \}.
   i) Why can these vectors not be linearly independent?
   ii) Carry out the Gram-Schmidt process. How does the dependence of the vectors
       \{ x_j \} affect the calculation of the \{ z_j \}?

2) Let \( y = (1, 2, 3) \in \mathbb{E}_3 \). Find three orthonormal vectors \{ u_1, u_2, u_3 \} such that \( u_1 \) is
   parallel to \( y \).

3) Let \( V \) be the set of all continuous real valued functions which are square integrable over
   \((0, \infty)\) with respect to the weight function
   \[
   w(t) = e^{-t},
   \]
   i.e.,
   \[
   V = \left\{ f : f \in C^0(0, \infty) \text{ and } \int_0^\infty f(t)^2 e^{-t} dt < \infty \right\}
   \]
   i) Show that \( V \) is a vector space over \( \mathbb{R} \).
   ii) Show that
       \[
       \langle f, g \rangle = \int_0^\infty f(t)g(t)e^{-t} dt
       \]
       defines an inner product on \( V \).
   iii) Show that \( M = \text{span}\{1, t, t^2\} \subset V \).
   iv) Find an orthogonal basis of \( M \). Scale the vectors such that they assume the value
       1 at \( t = 0 \) (if you solve this problem correctly you will find the first three so-called
       Laguerre polynomials).
We are going to discuss functions = mappings = transformations = operators from one vector space $V_1$ into another vector space $V_2$. However, we shall restrict our sights to the special class of linear operators which are defined as follows.

**Definition:** An operator $L$ from $V_1$ into $V_2$ is linear if

$$L(x + \alpha y) = Lx + \alpha Ly \quad \text{for all } x, y \in V_1 \quad \text{and } \alpha \in F.$$ 

In this case $V_1$ is the domain of $L$, and its range, denoted by $R(L)$, is contained in $V_2$.

**Examples:**
1) The most important example for us:

$$V_1 = \mathbb{R}_n \quad \text{(or } \mathbb{C}_n), \quad V_2 = \mathbb{R}_m \quad \text{(or } \mathbb{C}_m)$$

and

$$Lx \equiv Ax$$

where $A$ is an $m \times n$ real (or complex) matrix.

2) Let $K(t, s)$ be a function of two variables which is continuous on the square $[0, 1] \times [0, 1]$. Define $Lf$ by

$$(Lf)(t) \equiv \int_0^1 K(t, s)f(s)ds$$

then $L$ is a linear operator from $C^0[0, 1]$ into $C^0[0, 1]$. $L$ is called an integral operator.

3) Define $Lf$ by

$$(Lf)(t) \equiv \int_0^t f(s)ds$$

then $L$ is a linear operator from $C^0[0, 1]$ into the subspace $M$ of $C^1[0, 1]$ defined by

$$M = \{g : g \in C^1[0, 1], \ g(0) = 0\}.$$ 

4) Define the operator $Df$ by

$$(Df)(t) \equiv f'(t)$$

then $D$ is a linear operator from $C^1[0, 1]$ into $C^0[0, 1]$. 

39
5) Define the linear operator \( Lu \) by

\[
(Lu)(t) \equiv \sum_{n=0}^{N} a_n(t)u^{(n)}(t)
\]

then \( L \) is a linear operator from \( C^N[a,b] \) into \( C^0[a,b] \). \( L \) will be called an \( N \)th order linear differential operator with variable coefficients.

**Definition:** The inverse of a linear operator is the operator which maps the element \( Lx \) in the range of \( L \) to \( x \) in the domain of \( L \).

**Theorem.** A linear operator can have an inverse only if \( Lx = 0 \) implies that \( x = 0 \).

**Proof.** If \( Lx = y \) then the inverse of \( L \) is the mapping which takes \( y \) to \( x \). Suppose now that \( Lx_1 = y \) and \( Lx_2 = y \). Then by linearity \( L(x_1 - x_2) = 0 \). If \( x_1 - x_2 \neq 0 \) then there is no function which maps every \( y \) in the range of \( L \) uniquely into the domain of \( D \), i.e., the inverse function does not exist.

As an illustration we consider examples 3 and 4. We see that for any \( f \in C^0[0,1] \)

\[
(D(Lf))(t) \equiv \frac{d}{dt} \int_{0}^{t} f(s)ds = f(t).
\]

On the other hand,

\[
(L(Df))(t) \equiv \int_{0}^{t} f'(s)ds = f(t) - f(0).
\]

So \( D \) is the inverse of \( L \) on the range of \( L \) in the first case but \( L \) is not the inverse of \( D \) in the second case. Note that

\[
(Lf)(t) \equiv 0
\]

implies that \( f(t) \equiv 0 \) as seen by differentiating both sides, but

\[
(Df)(t) \equiv 0
\]

does not imply that \( f(t) \equiv 0 \) since any constant function would also serve. However, if we consider \( D \) as an operator defined on the space \( M \) defined in 3) above then \( f(0) = 0 \) and the integration denoted by \( L \) is indeed the inverse of the differentiation denoted by \( D \). These examples serve to illustrate that when we define an operator we also have to specify its domain.
Linear operators from $\mathbb{R}^n$ (or $\mathbb{C}^n$) into $\mathbb{R}^m$ (or $\mathbb{C}^m$)

We are now considering the case of

$$Lx \equiv Ax$$

where $A$ is an $m \times n$ matrix with entries $a_{ij}$. It is assumed throughout that you are familiar with the rules of matrix addition and multiplication. Thus we know that if $Ax = y$ then $y$ is a vector with $m$ components where

$$y_i = \sum_{j=1}^{n} a_{ij}x_j$$

i.e., we think of dotting the rows of $A$ into the column vector $x$ to obtain the column vector $y$. However, this not a helpful way of interpreting the action of $A$ as a linear operator. A MUCH MORE useful way of looking at $Ax$ is the following decomposition

$$(7.1) \quad Ax = \sum_{j=1}^{n} x_j A_j$$

where $x = (x_1, \ldots, x_n)$ and $A_j$ is the $j$th column of $A$ which is a column vector with $m$ components. That this relation is true follows by writing

$$x = \sum_{j=1}^{n} x_j \hat{e}_j$$

where $\hat{e}_j$ is the $j$th unit vector, and by observing that $A\hat{e}_j = A_j$. The immediate consequence of this interpretation of $Ax$ is the observation that

$$R(A) = \text{span}\{A_1, \ldots, A_n\}.$$ 

Many problems in linear algebra revolve around solving the linear system

$$Ax = b$$

where $A$ is an $m \times n$ matrix and $b$ is a given vector $b = (b_1, \ldots, b_m)$. It follows immediately from (7.1) that a solution can exist only if $b \in \text{span}\{A_1, \ldots, A_n\}$. Moreover, if the columns of $A$ are linearly independent then $Ax = 0$ has only the zero solution so that the solution of
Ax = b would have to be unique. In this case the inverse of A would have to exist on \( R(A) \) even if A is not square.

However, we usually cannot tell by inspection whether the columns of A are linearly independent or whether b belongs to the range of A. That question can only be answered after we have attempted to actually solve the linear system. But how do we find the solution \( x \) of \( Ax = b \) for an \( m \times n \) matrix A?

**Gaussian elimination**

It is assumed that you are familiar with Gaussian elimination so we shall only summarize the process. We subtract multiples of row 1 of the system from the remaining equations to eliminate \( x_1 \) from the remaining \( m - 1 \) equations. If \( a_{11} \) should happen to be zero then this process cannot get started. In this case we reorder the equations of \( Ax = b \) so that in the new coefficient matrix \( a_{11} \neq 0 \). The process then starts over again on the remaining \( m - 1 \) equations in the \( m - 1 \) unknowns \( \{x\}_{j=2}^n \). Eventually the system is so small that we can find its solution or observe that a solution cannot exist. Back-substitution yields the solution of the original system, if it exists.

**The LU decomposition of A**

For non-singular square matrices Gaussian elimination is equivalent to factoring A (or a modification \( PA \) of A obtained by interchanging certain rows of A). A consistent approach to Gaussian elimination for an \( n \times n \) matrix is as follows:

1) Let \( a_{ij}^{(1)} = a_{ij} \) (the original entries of A). Then for \( k = 1, \ldots, n - 1 \) compute the multipliers

\[
m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad i = k + 1, \ldots, n
\]

(where we have assumed that we do not divide by zero) and overwrite \( a_{ij}^{(k+1)} \) with

\[
a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)} \quad i = k + 1, \ldots, n \quad j = 1, \ldots, n.
\]

We denote by \( A^{(k)} \) the matrix with elements \( a_{ij}^{(k)}, \quad i = k, \ldots, n; \quad j = 1, \ldots, n. \)

2) Let L be the lower triangular matrix with entries

\[
L_{ii} = 1
\]
\[ L_{ij} = m_{ij} \quad j < i. \]

Let \( U \) be the upper triangular matrix with entries
\[ U_{ij} = a_{ij}^{(i)} \quad j \geq i. \]

\((U\) is actually the matrix \( A^{(m)} \), but in computer implementations of the \( LU \) factorization
the zeros below the diagonal are not computed. In fact, the elements of \( L \) below the
diagonal are usually stored there in \( A^{(m)}. \))

**Theorem:** \( A = LU. \)

**Proof.** \((LU)_{ij} = (m_{i1}, m_{i2}, \ldots, m_{i,i-1}, 1, 0, \ldots, 0) \cdot (u_{1j}, u_{2j}, \ldots, u_{jj}, 0, \ldots, 0) = \sum_{k=1}^{i-1} m_{ik}a_{kj}^{(k)} + a_{ij}^{(i)} = \sum_{k=1}^{i-1} \left[ d_{ij}^{(k)} - d_{ij}^{(k+1)} \right] + a_{ij}^{(i)} = a_{ij}^{(1)} = a_{ij}. \) We see that under the hypothesis that all elements of \( L \) can be found the original matrix has been factored into
the product of two triangular matrices. This product allows an easy solution of
\[ Ax = b. \]

Let \( y \) be the solution of
\[ Ly = b \]

then \( x \) is the solution of
\[ Ux = y \]

because \( Ax = LUx = Ly = b. \) An example may serve to clarify this algorithm. Consider
the system
\[ Ax = \begin{pmatrix} 4 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 4 \end{pmatrix} x = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}. \]

For \( k = 1 \) we obtain
\[ L_{11} = 1 \]
\[ L_{21} = m_{21} = \frac{2}{4} \]
\[ L_{31} = m_{31} = \frac{1}{4} \]

and
\[ A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{12}^{(1)} \\ a_{21}^{(2)} & a_{22}^{(2)} & a_{22}^{(2)} \\ a_{31}^{(2)} & a_{32}^{(2)} & a_{33}^{(2)} \end{pmatrix} = \begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & \frac{3}{2} & \frac{15}{4} \end{pmatrix}. \]
For $k = 3$ we obtain $L_{32} = m_{32} = \frac{1}{2}$ and

$$A^{(3)} = \begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & 0 & 3 \end{pmatrix}$$

Thus

$$LU = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{4} & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & 0 & 3 \end{pmatrix} = A.$$

In order to solve $Ax = b$ we now solve

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{4} & \frac{1}{2} & 1 \end{pmatrix} y = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix},$$

yielding $y = (1, \frac{3}{2}, 2)$, and

$$\begin{pmatrix} 4 & 2 & 1 \\ 0 & 3 & \frac{3}{2} \\ 0 & 0 & 3 \end{pmatrix} x = \begin{pmatrix} 1 \\ \frac{3}{2} \\ 2 \end{pmatrix}.$$

We find that the solution of $Ax = b$ is

$$x = (0, \frac{1}{6}, \frac{2}{3}).$$

Of course, it is possible that the diagonal element $a_{kk}^{(k)}$ is zero. In this case we interchange row $k$ of the matrix $A^{(k)}$ with row $i$ for some $i > k$ for which $a_{ik}^{(k)} \neq 0$. If $A$ is non-singular this can always be done. In fact, computer codes for the $LU$ decomposition of a matrix $A$ routinely exchange row $k$ with row $i$ for that $i$ for which $|a_{ik}^{(k)}| \geq |a_{jk}^{(k)}|$, $j = k, \ldots, n$. This process is called partial pivoting. It assures that $|m_{kj}| \leq 1$ and stabilizes the numerical computation. Any text on numerical linear algebra will discuss the $LU$ factorization and its variants in some detail. In this course we shall hand over the actual solution to the computer.

For subsequent modules we shall retain the following observations:

1) Gaussian elimination can also be applied to non-square system of the form

$$Ax = b.$$

If $m < n$ then the last equation obtained is a linear equation in $\{x_m, \ldots, x_n\}$. If $m > n$ then the the last $m - n$ equations all have zero coefficients. A solution of $Ax = b$ can
exist only if the last \( m - n \) terms of the source term \( b' \) generated during the elimination likewise vanish.

2) If \( A \) is a non-singular \( n \times n \) matrix then there always exists an \( LU \) decomposition of the form

\[
LU = PA
\]

where \( P \) is a non-singular matrix which permutes the rows of \( A \). If no partial pivoting has to be carried out then \( P = I \).
Module 7 - Homework

1) Define \( Lx \equiv Ax \) where \( A \) is an \( m \times n \) complex matrix. What are the domain and range of \( L \)?

2) Define
\[
(Lf)(t) \equiv f'(t) + \int_0^1 K(t, s)f(s)ds
\]
where \( K \) is continuous in \( s \) and \( t \) on the unit square.

i) What is a suitable domain for \( L \)? What is the corresponding range?

ii) Let \( f(t) \equiv \cos t \) and \( K(x, y) \equiv e^{x-y} \). Find \( Lf \).

3) Let \( L : V_1 \rightarrow V_2 \) denote the following operator and spaces:

i) \( V_1 = \{ f : f \in C^2[0, 1], f(0) = f(1) = 0 \} \)
   \( V_2 = C^0[0, 1] \)
   \( (Lf)(t) \equiv f''(t) + f(t). \)

ii) \( V_1 = \{ f : f \in C^2[0, 1], f(1) = 0 \} \)
   \( V = C^0[0, 1] \)
   \( (Lf)(t) \equiv f''(t) + f(t). \)

In both cases prove or disprove: \( Lf = 0 \) if and only if \( f \equiv 0 \).

4) Compute the \( LU \) decomposition of
\[
A = \begin{pmatrix}
3 & 2 & 1 \\
6 & 6 & 3 \\
0 & 2 & 2
\end{pmatrix}.
\]
Use the \( LU \) decomposition to solve
\[
Ax = \hat{e}_1.
\]

5) Let \( P_{ij} \) be the matrix obtained from the \( m \times m \) identity matrix \( I \) by interchanging rows \( i \) and \( j \). Let \( A \) be any \( m \times n \) matrix. What is the relation between \( P_{ij}A \) and \( A \)?

Let
\[
A = \begin{pmatrix}
1 & 2 & 3 \\
1 & 2 & 4 \\
2 & 6 & 1
\end{pmatrix}.
\]
Apply the \( LU \) factorization to \( A \) and to \( P_{23}A \).
**Definition:** Let \( L : V_1 \rightarrow V_2 \) be a linear operator. The null space \( \mathcal{N}(L) \) of \( L \) is the subspace of \( V_1 \) defined by

\[
\mathcal{N}(L) = \{ x \in V_1 : Lx = 0 \}
\]

*Note:* The null space of \( L \) is sometimes called the kernel of \( L \).

**Examples:**

i) \( Lx \equiv Ax \equiv \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} x = 0 \) then \( \mathcal{N}(A) = \text{span}\{(1, -1)\} \in \mathbb{R}_2 \).

ii) \( Lf \) defined by \( (Lf)(t) \equiv f''(t) \) for \( f \in C^2[a, b] \) then \( \mathcal{N}(L) = \text{span}\{1, t\} \).

iii) \( L : C^0[-1, 1] \rightarrow R \) defined by

\[
Lf \equiv \int_{-1}^{1} f(s)ds
\]

then \( \mathcal{N}(L) \) contains the subspace of all odd continuous functions on \([-1, 1]\) plus many other functions such as \( f(t) = t^2 - 1/3 \).

We shall now restrict ourselves to \( m \times n \) real matrices. We note that always \( 0 \in \mathcal{N}(A) \). If this is the only vector in \( \mathcal{N}(A) \), i.e., if \( \mathcal{N}(A) = \{0\} \) then the null space is the trivial null space with dimension 0.

We also know from

\[
Ax = \sum_{j=1}^{n} x_j A_j
\]

that \( R(A) = \text{span}\{A_1, \ldots, A_n\} \in \mathbb{R}_m \). The range of \( A \) is often called the column space of \( A \) and the dimension of this space is called the rank of \( A \), i.e.,

\[
r(A) = \text{rank}(A) = \dim R(A) = \dim \text{column space of } A.
\]

We note that \( r(A) < \min\{m, n\} \).

**Example:** Let \( x \) and \( y \) be two column vectors in \( \mathbb{R}_n \). Then the \( n \times n \) matrix

\[
x \cdot y^T = (y_1 \vec{x} \cdot y_2 \vec{x} \cdots y_n \vec{x})
\]

is a matrix with rank 1 since every column is a multiple of \( \vec{x} \).
**Theorem:** Let $A$ be an $m \times n$ matrix. Then

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

**Proof:** Let $\{y_1, \ldots, y_r\}$ be a basis of $R(A)$. Let $\{x_1, \ldots, x_r\}$ be the vectors which satisfy

$$Ax_j = y_j \quad \text{for } j = 1, \ldots, r.$$ 

Let $\{z_1, \ldots, z_p\}$ be a basis of $\mathcal{N}(A)$. Then the vectors $\{x_1, \ldots, x_j, z_1, \ldots, z_p\}$ are linearly independent because if

$$\sum_{j=1}^{r} \alpha_j x_j + \sum_{j=1}^{p} \beta_j z_j = 0$$

then

$$A \left( \sum_{j=1}^{r} \alpha_j x_j + \sum_{j=1}^{p} \beta_j z_j \right) = \sum_{j=1}^{r} \alpha_j y_j = 0$$

which implies that $\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0$. But then the linear independence of the $\{z_j\}$ implies that the $\{\beta_j\}$ also must vanish. Finally, let $x$ be arbitrary in $R_n$. Then

$$Ax = \sum_{j=1}^{r} \gamma_j y_j$$

for some $\{\gamma_j\}$. This implies that $A \left( x - \sum_{j=1}^{r} \gamma_j x_j \right) = 0$ so that $x - \sum_{j=1}^{r} \gamma_j x_j \in \mathcal{N}(A)$, i.e.,

$$x - \sum_{j=1}^{r} \gamma_j x_j = \sum_{j=1}^{p} \beta_j z_j.$$ 

Hence the linearly independent vectors $\{x_1, \ldots, x_r, z_1, \ldots, z_p\}$ span $R_n$ and

$$r + p \equiv \text{rank}(A) + \dim \mathcal{N}(A) = n.$$ 

It follows immediately that if $A$ is an $m \times n$ matrix and $m < n$ then $\dim \mathcal{N}(A) \geq 1$ because $\text{rank}(A) \leq \min\{m, n\}$. In particular, this implies that $Ax = 0$ has a non-zero solution so that such a matrix cannot have an inverse.

So far we have looked at the columns of $A$ as $n$ column vectors in $R_m$. Likewise, the $m$ rows of $A$ define a set of $m$ vectors in $R_n$. What can we say about the number of linearly independent rows of $A$?

We recall from the homework of Module 2 that if $\langle x, y \rangle$ denotes the dot product then

$$\langle Ax, y \rangle = \langle x, A^T y \rangle.$$ 

48
for \( x \in R_n \) and \( y \in R_m \). Next, let \( \{y_1, y_2, \ldots, y_r\} \) be a basis of \( R(A) \) and apply the Gram-Schmidt orthogonalization process to the vectors

\[
\{y_1, y_2, \ldots, y_r, \hat{e}_1, \hat{e}_2, \ldots, \hat{e}_m\}
\]

then the first \( r \) orthogonal vectors will be a basis of \( R(A) \) and the remaining \( m - r \) vectors \( \{Y_1, Y_2, \ldots, Y_{m-r}\} \) will be orthogonal to \( R(A) \). Since \( A^T Y_j \in R_n \) it it follows from

\[
\langle A^T Y_j, A^T Y_j \rangle = \langle Y_j, A(A^T Y_j) \rangle = 0
\]

that \( A^T Y_j = 0 \) so that \( \dim \mathcal{N}(A^T) \geq (m - r) \). Finally, we observe that if \( Ax \neq 0 \) then \( \langle A^T(Ax), x \rangle > 0 \) so that \( Ax \) cannot belong to \( \mathcal{N}(A^T) \). Hence \( \dim \mathcal{N}(A^T) = m - r \) so that \( \text{rank}(A^T) = \) number of linearly independent rows of \( A = m - (m - r) = r \). In other words, an \( m \times n \) matrix has as many independent rows as columns.

Finally, we observe that if we add to any row of \( A \) a linear combination of the remaining rows we do not change the number of independent rows. Hence we can apply Gaussian elimination to the rows of \( A \) and read off the number of independent rows of \( A \) from the final form of \( A \) where all elements below the diagonal are zero.

Implications for the solution of the linear system

\[
Ax = b
\]

where \( A \) is an \( m \times n \) matrix.

1) We shall assume that \( b \in R(A) \).

i) If the columns of \( A \) are linearly independent then \( Ax = b \) has a unique solution regardless of the size of the system. In this case the inverse mapping exists for every element \( y \in R(A) \).

ii) If the columns of \( A \) are linearly dependent then \( \dim \mathcal{N}(A) \geq 1 \) and there are infinitely many solutions. One can then constrain the solution by asking, for example, for the minimum norm solution.

iii) If \( m \geq n \) the columns of \( A \) may or may not be linearly dependent. If \( m < n \) then the columns of \( A \) must be linearly dependent.
iv) If rank($A$) = $m$ then $b \in R(A)$.  

2) Regardless of the size of the system, if $b \not\in R(A)$ there cannot be a solution. If $b \not\in R(A)$ then Gaussian elimination will lead to inconsistent equations.

Two points of view for finding an approximate solution of $Ax = b$ when $b \not\in R(A)$.

I. The “Least Squares Solution”:

When the system $Ax = b$ is inconsistent then for any $x \in R_n$ the residual, defined as

$$r(x) \equiv b - Ax,$$

cannot be zero. In this case it is common to try to minimize the residual (in some sense) over all $x \in R_n$ (or possibly over some specially chosen set of “admissible” $x \in R_n$). We shall consider here only the case of minimizing a norm of the residual which is obtained from an inner product. This means we need to find the minimum of the function $f$ defined by

$$f(x) \equiv \langle r(x), r(x) \rangle = \langle b - Ax, b - Ax \rangle.$$

Let us assume now that we are dealing with real valued vectors. Then $f$ is a function of the $n$ real variables $x_1, \ldots, x_n$, and calculus tells us that a necessary condition for the minimum is that

$$\nabla f(x) = 0.$$

We find that

$$\frac{\partial f}{\partial x_j} \equiv \langle -A_j, b - Ax \rangle + \langle b - Ax, A_j \rangle = 0.$$

Since in a real vector space the inner product is symmetric it follows that $x$ must be a solution of

$$\langle A_j, Ax \rangle = \langle A_j, b \rangle \quad \text{for } j = 1, \ldots, n.$$

If the inner product is the dot product on $R_n$ then these $n$ equations can be written in matrix form as

$$A^T Ax = A^T b.$$

If the $n \times n$ matrix $A^T A$ has rank $n$ then dim $\mathcal{N}(A^T A) = 0$ and $(A^T A)^{-1}$ exists so that

$$x = (A^T A)^{-1} Ab.$$
This is the least squares solution of \( Ax = b \) in Euclidean \( n \)-space. If \( A \) and hence \( A^T \) are square and have rank \( n \) then \( A^T \) is invertible and \( x \) solves \( Ax = b \).

II. We know that we can solve \( Ax = b' \) for any \( b' \in R(A) \) since Gaussian elimination will give the answer. One may now pose the problem:

Find the solution \( x \) of \( Ax = b' \) where \( b' \) is the vector in \( R(A) \) which is “closest” in norm to \( b \). As we saw in module 4 the vector \( b' \) is the orthogonal projection of \( b \) onto \( \text{span}\{A_1, \ldots, A_n\} \). Thus

\[
b' = \sum_{j=1}^{n} \alpha_j A_j = A\tilde{\alpha}
\]

where \( \tilde{\alpha} \) is computed from

\[
A\tilde{\alpha} = d
\]

with \( A_{ij} = \langle A_j, A_i \rangle \) and \( d_i = \langle b, A_i \rangle \). It follows that \( A \) and \( d \) can be written in matrix notation as

\[
A = A^T A, \quad d = A^T b
\]

so that by inspection the solution of

\[
Ax = b' = A\alpha = A(A^T A)^{-1}A^T b
\]

is

\[
x = (A^T A)^{-1}A^T b
\]

provided \( A \) has rank \( n \). Hence the least squares solution is the exact solution of the “closest” linear system for which there is an exact solution.
Module 8 - Homework

1) Let \( V_1 = \{ u : u \in C^0[-1, 1] \} \)
\[ V_2 = C^0[-1, 1] \]
Define
\[ (Lu)(t) = \int_{-1}^{t} su(s)ds. \]
Show that \( L \) is linear and find \( \mathcal{N}(L) \). Show that the range of \( L \) is not all of \( V_2 \).

2) Let
\[ A = \begin{pmatrix} 1 & 5 & 9 & 13 & 6 \\ 2 & 6 & 10 & 14 & 8 \\ 3 & 7 & 11 & 15 & 10 \\ 4 & 8 & 12 & 16 & 12 \end{pmatrix}. \]
What is the rank of \( A \)?
Find an orthogonal (with respect to the dot product) basis of the null space and range of \( A \).

3) Let \( A \) be an \( m \times n \) matrix. Assume that its columns are linearly independent.
   i) Show that in this case \( n \leq m \).
   ii) Show that one can find an \( n \times m \) matrix \( B \) such that
   \[ BA = I_n \quad \text{where } I_n \text{ is the } n \times n \text{ identity matrix}. \]

4) Suppose the cost \( C(t) \) of a process grows quadratically with time, i.e.,
\[ C(t) = a_0 + a_1 t + a_2 t^2 \]
Company records contain the following data:

<table>
<thead>
<tr>
<th>time taken</th>
<th>measured cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.911</td>
</tr>
<tr>
<td>.2</td>
<td>.84</td>
</tr>
<tr>
<td>.3</td>
<td>.788</td>
</tr>
<tr>
<td>.4</td>
<td>.76</td>
</tr>
<tr>
<td>.5</td>
<td>.747</td>
</tr>
<tr>
<td>.6</td>
<td>.77</td>
</tr>
</tbody>
</table>

What would be your estimate of the cost of the process if it takes one unit of time?

52
Throughout $A$ is an $n \times n$ real (or complex) matrix. Let $V$ be the vector space of all these matrices with respect to the usual rules of matrix addition and scalar multiplication.

**Definition:** The determinant of a matrix, denoted by $\det(A)$, is a function from $V$ to the real (or complex) numbers with the following properties: If $A_1, \ldots, A_n$ denote the columns of $A$ then

1) $\det(A_1, \ldots, A_i + \alpha B_i, \ldots, A_n) = \det(A_1, \ldots, A_n) + \alpha \det(A_1, \ldots, A_{i-1}, B_i, A_{i+1}, \ldots, A_1)$
   for each $i$ (i.e., the function is linear in each coordinate, i.e., it is said to be multilinear)
2) $\det(A) = (-1)^{i} \det(B)$ where $B$ is the matrix obtained from $A$ by interchanging two columns (the function is said to be anti-symmetric)
3) $\det(I) = 1$ where $I$ is the $n \times n$ identity matrix.

(Not so obvious consequences of the definition):
1) The determinant (function) exists and is uniquely defined.
2) $\det A = \det(A^T)$
3) if $A$ has a zero row (or column) then $\det(A) = 0$
4) $\det(A) = 0$ iff the columns of $A$ are linearly dependent
5) $\det(AB) = \det(A) \det(B)$
6) computation: $\det(a_{11}) = a_{11}$ if $A$ is an $n \times n$ matrix then

$$\det(A) = \sum_{j=1}^{n} (-1)^{i+j} a_{ij} \det(A^{ij})$$

where $i$ refers to any row of $A$ and $A^{ij}$ is the $(n-1) \times (n-1)$ matrix obtained from $A$ by deleting row $i$ and column $j$.

7) if $A$ is triangular (with zeroes either above or below the diagonal) then $\det(A) = \prod_{j=1}^{n} a_{jj}$.

53
Eigenvalues and eigenvectors of a square matrix

**Definition:** A (complex) number $\lambda$ for which there exists a nontrivial solution $u \in C_n$ of the equation

$$Au = \lambda u$$

is an eigenvalue of $A$. $u$ is an eigenvector associated with this $\lambda$.

We observe that if $u$ is an eigenvector then any non-zero scalar multiple of $u$ is also an eigenvector so that $A$ maps span{$u$} into span{$u$}.

**Theorem:** An $n \times n$ matrix has $n$ eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$, although they may not be all distinct.

**Proof:** If $Au = \lambda u$ for a non-trivial $u$ then the matrix $A - \lambda I$ has a nontrivial null space so that its columns are linearly dependent. This implies that $\det(A - \lambda I) = 0$. The rule for the computation of determinants shows that $\det(A - \lambda I)$ is an $n$th order polynomial in $\lambda$ whose coefficients are functions of the entries of $A$. We “know” that an $n$th order polynomial has $n$ roots $\{\lambda_1, \ldots, \lambda_n\}$ and for each such $\lambda_i$ the null space of $A - \lambda_i I$ is not trivial. Hence we have $n$ eigenvalues.

If $A$ is a real matrix then the $n$th order polynomial will have real coefficients. This does not guarantee real roots, but if there is a complex eigenvalue $\lambda$ with a corresponding complex eigenvector $u$ then the complex conjugate $\bar{\lambda}$ is also an eigenvalue with eigenvector $\bar{u}$.

**Examples:** i) The matrix $A = I$ has the $n$ eigenvalues $\lambda = 1$ and every non-zero vector $u \in C_n$ is an eigenvector.

ii) The matrix

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

has eigenvalues $\lambda_1 = \lambda_2 = 1$ and for both we have the same eigenvector $u = (1, 0)$.

iii) The matrix

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

has eigenvalues and eigenvectors

$$\lambda_1 = 0 \text{ with } u_1 = (1, -1) \quad \text{and} \quad \lambda_2 = 2 \text{ with } u_2 = (1, 1).$$
Note that $u_1$ and $u_2$ are linearly independent; in fact they are orthogonal in $\mathbb{E}_2$.

iv) The matrix

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

has the eigenvalues and eigenvectors

$$\lambda_1 = i \text{ with } u_1 = (1, i) \quad \text{and} \quad \lambda_2 = -i \text{ with } u_2 = (1, -i).$$

**Observation:** Just as it is helpful to interpret $Ax$ as

$$Ax = \sum_{j=1}^{n} x_j A_j,$$

it is useful to write the $n$ equations $Au_i = \lambda_i u_i$ for $i = 1, \ldots, n$ in matrix form as

$$AU = U \Lambda$$

where $U = (u_1 \ u_2 \ \cdots \ u_n)$ and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. For example, if the $n$ eigenvectors should happen to be linearly independent, i.e., if the eigenvectors form a basis of the space $\mathbb{C}_n$, then the matrix $U$ is invertible and

$$A = U \Lambda U^{-1}.$$ 

Such a transformation is advantageous if one needs to compute powers of $A$. For example, the matrix exponential $e^A$ is defined by its Taylor series as

$$e^A = I + A + \frac{1}{2} A^2 + \frac{1}{3!} A^3 + = \sum_{n=1}^{\infty} \frac{1}{n!} A^n.$$ 

Matrix products are quite laborious to evaluate. In addition, this exponential series converges very slowly so that many terms are needed. As a consequence a direct evaluation of $e^A$ is generally not feasible. But if

$$A = U \Lambda U^{-1} \text{ then } A^k = U \Lambda^k U^{-1} = U \text{ diag}\{\lambda_1^k, \ldots, \lambda_n^k\} U^{-1}$$

and hence

$$e^A = U \text{ diag}\{e^{\lambda_1}, \ldots, e^{\lambda_n}\} U^{-1}.$$
The existence of an eigenvector basis is not easy to resolve. However, we have the following partial result.

**Theorem:** Suppose that the eigenvalues \( \{\lambda_1, \ldots, \lambda_k\} \) are distinct. Then the corresponding eigenvectors \( \{u_1, \ldots, u_k\} \) are linearly independent.

**Proof:** Suppose that \( Au_i = \lambda_i u_i \) and that the first \( k \) eigenvalues are distinct. Suppose further that the eigenvectors \( \{u_1, \ldots, u_m\} \) are linearly independent for some \( m < k \) but that \( u_{m+1} \in \text{span}\{u_1, \ldots, u_m\} \). Then

\[
(9.1) \quad u_{m+1} = \sum_{j=1}^{m} \alpha_j u_j
\]

and

\[
(9.2) \quad Au_{m+1} = \lambda_{m+1} u_{m+1} = \sum_{j=1}^{m} \alpha_j \lambda_j u_j.
\]

If we multiply (9.1) by \( \lambda_{mn} \) and subtract from (9.2) we find that

\[
\sum_{j=1}^{m} (\lambda_{m+1} - \lambda_j) \alpha_j u_j = 0.
\]

Since \( \lambda_{mn} - \lambda_j \neq 0 \) and the \( \{u_j\} \) are linearly independent this would require that \( \alpha_1 = \cdots = \alpha_m = 0 \) which contradicts that \( u_{m+1} \in \text{span}\{u_1, \ldots, u_m\} \). Hence there cannot be such an \( m \) so that \( \{u_1, \ldots, u_k\} \) is a set of \( k \) linearly independent vectors.
Module 9 - Homework

1) Let $A$ be an $n \times n$ matrix, $U = (u_1 \ u_2 \ldots u_n)$ and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. Compute the $j$th column of $AU = UA$ and show that you obtain the equation defining eigenvalues and eigenvectors.

Compute the $j$th column of $AU = \Lambda U$ and show that this is not the equation defining eigenvalues and eigenvectors.

2) Let

$$A = \begin{pmatrix} 1 & 3 & 8 & 1 & -2 \\ 1 & 2 & 9 & 2 & -1 \\ 1 & 2 & 11 & 3 & 0 \\ 1 & 2 & 11 & 6 & 1 \\ 1 & 2 & 11 & 6 & 5 \end{pmatrix}.$$

Find the determinant of $A$ with pen and paper, i.e., without using a computer. (You may verify your answer with a machine, if you wish).

3) Find the eigenvalues and eigenvectors of

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

and show by direct calculation that they are linearly independent.

4) Let

$$A = \frac{1}{40} \begin{pmatrix} -129 & 21 \\ -21 & -71 \end{pmatrix}.$$

Find a matrix $U$ such that

$$UAU^{-1}$$

is a diagonal matrix.

Compute $A^{10}$. You may use the computer to find $U$, although it can be done longhand easily enough. Do not use a computer to find $A^{10}$.
1) In a small chemical plant three tanks are connected with each other with an inflow and an outflow pipe.

i) Characterize all admissible flow rates for which the volume in each tank will remain constant.

ii) Suppose in this closed system the volume of each tank changes at a prescribed rate. Characterize the admissible volume changes.

**Answer:** Let $c(i, j)$ be the flow rate from tank $i$ to tank $j$. Then a mass balance requires that

\[
c(1, 2) + c(1, 3) = c(2, 1) + c(3, 1) \]
\[
c(2, 1) + c(2, 3) = c(2, 1) + c(3, 2) \]
\[
c(3, 1) + c(3, 2) = c(1, 3) + c(2, 3). \]

Let $x_1 = c(1, 2)$, $x_2 = c(1, 3)$, $x_3 = c(2, 1)$, $x_4 = c(2, 3)$, $x_5 = c(3, 1)$ and $x_6 = c(3, 2)$ then the mass balance equations can be rewritten as

\[
Ax = 0
\]

where

\[
A = \begin{pmatrix}
1 & 1 & -1 & 0 & -1 & 0 \\
1 & 0 & -1 & -1 & 0 & 1 \\
0 & 1 & 0 & 1 & -1 & -1
\end{pmatrix}.
\]

If we carry out Gaussian elimination we find that

\[
U = \begin{pmatrix}
1 & 1 & -1 & 0 & -1 & 0 \\
0 & -1 & 0 & -1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.
\]

Hence $\text{rank}(A) = 2$, $R(A) = \text{span}\{(1,1,0),(1,0,1)\}$ and $\text{dim} N(A) = 4$. A basis of the null space is found from $Ux = 0$ as

\[
\begin{align*}
    u_1 &= (1,0,1,0,0,0) & \text{(tank 1 and 2 exchange fluid)} \\
    u_2 &= (0,1,0,0,1,0) & \text{(tank 1 and 3 exchange fluid)} \\
    u_3 &= (0,0,0,1,0,1) & \text{(tank 2 and 3 exchange fluid)} \\
    u_4 &= (1,0,0,1,1,0) & \text{(the three tanks are connected in series)}.
\end{align*}
\]
Any flow schedule in the span\(\{u_i\}\) is an admissible flow schedule.

ii) The mass balance equations become

\[Ax = b\]

where \(b\) is the prescribed change of fluid in each tank. In order to solve the system we need that

\[b \in \text{span}\{(1,1,0),(1,0,1)\}\]

Hence \(b = \alpha(-1,1,1)\) would not be allowed for any \(\alpha \neq 0\). Of course, the model breaks down when a tank becomes empty or overflows.

2) Eigenvalues are usually obtainable only through a numerical calculation, but on occasion it is possible to obtain some useful a-priori estimates of what they might be. Suppose that

\[Au = \lambda u.\]

Since \(u\) is not the zero vector we can normalize \(u\). We shall write

\[y = \frac{u}{\|u\|_\infty}\]

so that \(|y_k| = 1\) for some \(k\) and \(|y_j| \leq 1\) for all \(j\). If we now look at the \(k\)th equation of \(Ay = \lambda y\) we obtain

\[(a_{kk} - \lambda)y_k = \sum_{\substack{j=1\ j\neq k}}^n a_{kj}y_j\]

so that for each eigenvalue there is a \(k\) such that

\[|a_{kk} - \lambda| \leq \sum_{\substack{j=1\ j\neq k}}^n |a_{kj}|.\]

Hence the eigenvalues have to lie in a union of disks given by

\[\bigcup_{i=1}^n \left\{ z : |a_{ii} - z| \leq \sum_{\substack{j=1\ j\neq i}}^n |a_{ij}| \right\}. \]
For example, suppose that $A$ is strictly diagonally dominant so that

$$|a_{ii}| > \sum_{j=1}^{n} |a_{ij}| \quad \text{for each } i$$

Then none of these circles contains the origin. Hence $\lambda = 0$ cannot be an eigenvalue which implies that

$$Ax = 0$$

cannot have a non-zero solution (which otherwise would be an eigenvector corresponding to $\lambda = 0$). Hence if $A$ is strictly diagonally dominant then $A$ is invertible.

3) Let $L$ denote the linear transformation in $\mathbb{R}_2$ which describes a reflection in $\mathbb{R}_2$ about the line $x_2 = x_1$. Find the matrix of $A$ and its eigenvalues and eigenvectors.

**Answer:** We know that a linear transformation from $\mathbb{R}_2$ to $\mathbb{R}_2$ has a matrix representation

$$Lx \equiv Ax$$

where the $i$th column of $A$ is the image of the $i$th unit vector. It follows in this case that

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$ 

From the geometry we observe that the vector $u_1 = (1, 1)$ stays unchanged and the vector $u_2 = (1, -1)$ goes into $(-1, 1) = -u_2$. Hence without calculation

$$Au_1 = u_1 \quad \text{and} \quad Au_2 = -u_2$$

so that $\lambda = 1$ with eigenvector $u_1$ and $\lambda = -1$ with eigenvector $u_2$. It is straightforward to verify these results algebraically.

4) Find the matrix for the orthogonal projection in $\mathbb{E}_3$ onto the plane

$$x_1 + x_2 + x_3 = 0$$

and determine its eigenvalues and eigenvectors geometrically.
Answer: In order to write down the matrix we need to find the images of the three unit vectors \( \{\hat{e}_i\} \). We can find these images once we have a basis for the subspace onto which we project. Since the subspace is a plane in \( \mathbb{E}_3 \) any two linearly independent vectors in this plane will serve as a basis. By inspection we see that

\[
    u_1 = (1, -1, 0) \quad \text{and} \quad u_2 = (1, 1, -2)
\]

form an orthogonal basis for the plane which simplifies the calculation of \( P \). It follows that

\[
P\hat{e}_i = \frac{\langle \hat{e}_i, u_1 \rangle}{\langle u_1, u_1 \rangle} u_1 + \frac{\langle \hat{e}_i, u_2 \rangle}{\langle u_2, u_2 \rangle} u_2
\]

so that

\[
A = \begin{pmatrix}
    2/3 & -1/3 & -1/3 \\
   -1/3 & 2/3 & -1/3 \\
   -1/3 & -1/3 & 2/3
\end{pmatrix}.
\]

A quick check shows that at least all images \( P\hat{e}_i \) belong to the plane which is necessary but not sufficient for the correctness of the derivation of \( A \). Without calculation we recognize that \( Pu_1 = u_1 \) and \( Pu_2 = u_2 \) so that \( \lambda = 1 \) must be an eigenvalue which occurs twice with corresponding orthogonal eigenvectors. We also note that any vector orthogonal to the plane is mapped to the origin. Hence \( P(1,1,1) = 0 \) or, if you prefer,

\[
P(u_1 \times u_2) = A(u_1 \times u_2) = 0
\]

so that \( \lambda = 0 \) is also an eigenvalue with eigenvector \( (1,1,1) \). Again, these results can be verified algebraically.

5) Suppose we have a rotation in \( \mathbb{R}_3 \) around the \( x_1 \)-axis in the clockwise direction (looking along the positive \( x_1 \)-axis toward the origin) followed by a rotation through \( \pi/4 \) clockwise around the \( x_3 \)-axis. Find the matrix for the combined rotation and the axis of rotation.

Answer: Let \( A \) be the matrix for the first rotation. Then

\[
A\hat{e}_1 = \hat{e}_1
\]

\[
A\hat{e}_2 = -\hat{e}_3
\]

\[
A\hat{e}_3 = \hat{e}_2.
\]
Hence
\[ A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}. \]

Let \( B \) be the matrix for the second rotation, then
\[ B\hat{e}_1 = 1/\sqrt{2}\hat{e}_1 - 1/\sqrt{2}\hat{e}_2 \]
\[ B\hat{e}_2 = 1/\sqrt{2}\hat{e}_1 + 1/\sqrt{2}\hat{e}_2 \]
\[ B\hat{e}_3 = \hat{e}_3 \]
so that
\[ B = \begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ -1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 0 & 0 & 1 \end{pmatrix}. \]

The combined rotation is
\[ C = BA = \begin{pmatrix} 1/\sqrt{2} & 0 & 1/\sqrt{2} \\ -1/\sqrt{2} & 0 & 1/\sqrt{2} \\ 0 & -1 & 0 \end{pmatrix}. \]

The eigenvalues and eigenvectors of \( C \) are obtained from the computer.

\[ \text{In}[1] := m = \{\{1/Sqrt[2],0,1/Sqrt[2]\},\{-1/Sqrt[2],0,1/Sqrt[2]\},\{0,-1,0\}\} \]
\[ \text{Out}[1] = \{\{1 + Sqrt[2], -1, 1\}\}, \{\{-1/2 + Sqrt[2] + I Sqrt[10 + 4 Sqrt[2]], 0, 1\}\}, \{0,-1,0\}\} \]
\[ \text{In}[2] := \text{Eigenvalues}[m] \]
\[ \text{Out}[2] = \left\{ \frac{1}{4} \left[ -2 + Sqrt[2] + I Sqrt[10 + 4 Sqrt[2]] \right] , \right. \]
\[ \left. \frac{-2 + Sqrt[2] + I Sqrt[10 + 4 Sqrt[2]]}{4} \right\} \]
\[ \text{In}[3] := \text{Eigenvectors}[m] \]
\[ \text{Out}[3] = \left\{ \{1 + Sqrt[2], -1, 1\}, \left\{ \frac{-Sqrt[2] - I Sqrt[10 + 4 Sqrt[2]] + I Sqrt[2 \ (10 + 4 Sqrt[2])]}{4} , 1 \right\} \right. \]
\[ \left. \frac{2 - Sqrt[2] + I Sqrt[10 + 4 Sqrt[2]]}{4} , 1 \right\} \]
The relevant information is the axis of rotation which is the eigenvector
\[ u_1 = \left(1 + \sqrt{2}, -1, 1\right) \]
corresponding to the eigenvalue \( \lambda = 1 \).

6) Find the matrix for the rotation about about an axis of rotation parallel to the vector \( \vec{u}_1 \) through an angle \( \theta \) counterclockwise when looking along \( \vec{u}_1 \) toward \( \vec{0} \).

**Answer:** The rotation is easy to describe in a right handed orthogonal coordinate system \( \{\vec{u}_1, \vec{u}_2, \vec{u}_3\} \) of \( \mathbb{E}_3 \) where \( \vec{u}_2 \) and \( \vec{u}_3 \) are orthogonal vectors in the plane perpendicular to \( \vec{u}_1 \). Let the given vector \( \vec{u}_1 \) be \( \vec{u}_1 = (u_1, u_2, u_3) \). Then a vector perpendicular to \( \vec{u}_1 \) is the vector \( \vec{u}_2 = (u_2, -u_1, 0) \). A right handed coordinate system is obtained if we set \( \vec{u}_3 = \vec{u}_1 \times \vec{u}_2 = (u_1 u_3, +u_2 u_3, -u_1^2 - u_2^2) \). Let us normalize the vectors and choose
\[ v_i = \frac{\vec{u}_i}{\|\vec{u}_i\|_2} \quad \text{for } i = 1, 2, 3. \]
The set \( \{v_i\} \) will play much the same role as the set of unit vectors \( \{\hat{e}_i\} \). (We are dropping the arrows indicating vectors because the components will no longer appear explicitly.)

Let \( R \) denote the rotation operator. Then
\[ Rv_1 = v_1 \]
because \( v_1 \) is the axis of rotation and does not change.
\[ Rv_2 = \alpha_2 v_2 + \beta_2 v_3 \]
because the image of \( v_2 \) remains in the plane spanned by \( \{v_2, v_3\} \). Moreover
\[ \langle Rv_2, v_2 \rangle = \alpha_2 \langle v_2, v_2 \rangle + \beta_2 \langle v_3, v_2 \rangle = \|Rv_2\|\|v_2\| \cos \theta = \cos \theta \]
which together with
\[ \langle v_2, v_2 \rangle = \langle Rv_2, Rv_2 \rangle = \alpha_2^2 + \beta_2^2 = 1 \]
determines $\alpha_2$ and $\beta_2$. Similarly we find

$$Rv_3 = \alpha_3v_2 + \beta_3v_3.$$  

Since $\{v_1, v_2, v_3\}$ forms an orthonormal basis of $\mathbb{E}_3$ there are constants $\{\gamma_{1i}, \gamma_{2i}, \gamma_{3i}\}$ such that

$$\hat{e}_i = \gamma_{1i}v_1 + \gamma_{2i}v_2 + \gamma_{3i}v_3 = (v_1 \ v_2 \ v_3) \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \gamma_{3i} \end{pmatrix}$$

It follows that

$$I = (e_1 \ e_2 \ e_3) = (v_1 \ v_2 \ v_3) \begin{pmatrix} \gamma_{11} & \gamma_{12} & \gamma_{13} \\ \gamma_{21} & \gamma_{22} & \gamma_{23} \\ \gamma_{31} & \gamma_{32} & \gamma_{33} \end{pmatrix}$$

so that the last matrix is $(v_1 \ v_2 \ v_3)^{-1}$. Now consider the image of the unit vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$. We have

$$Re_i = \gamma_{1i}v_1 + \gamma_{2i}v_2 + \gamma_{3i}v_3$$

so that

$$Re_i = \gamma_{1i}v_1 + \gamma_{2i}(\alpha_2v_2 + \beta_2v_3) + \gamma_{3i}(\alpha_3v_2 + \beta_3v_3)$$

$$= V \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} + \gamma_{3i} \alpha_3 \\ \gamma_{2i} + \gamma_{3i} \beta_3 \end{pmatrix} = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha_2 & \alpha_3 \\ 0 & \beta_2 & \beta_3 \end{pmatrix} \begin{pmatrix} \gamma_{1i} \\ \gamma_{2i} \\ \gamma_{3i} \end{pmatrix}$$

where $V = (v_1 \ v_2 \ v_3)$. Hence the matrix describing the rotation can be written as

$$(Re_1 \ Re_2 \ Re_3) = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha_2 & \alpha_3 \\ 0 & \beta_2 & \beta_3 \end{pmatrix} V^{-1}$$

Finally we observe from

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

that $(v_1 \ v_2 \ v_3)^T(v_1 \ v_2 \ v_3) = I$ so that $V^{-1} = V^T$. Hence the rotation matrix can be simplified to

$$R = V \begin{pmatrix} 1 & 0 & 0 \\ 0 & \alpha_2 & \alpha_3 \\ 0 & \beta_2 & \beta_3 \end{pmatrix} V^T.$$  

The matrix $V^T$ maps the units vectors $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ onto the basis $\{v_1, v_2, v_3\}$, the next matrix tells us how these basis vectors transform, and the matrix $V$ maps the $\{v_1, v_2, v_3\}$ basis back to the $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$ basis.
Module 10 - Homework

1) Let

\[ A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}. \]

Prove or disprove: \( A \) describes a rotation in \( \mathbb{R}^3 \).

2) Find the matrix \( P \) for the projection in \( \mathbb{E}^3 \) onto the subspace \( M = \text{span}\{(1, 1, 1), (1, 2, 1)\} \).

Find the eigenvalues \( \{\lambda_i\} \) of \( A \). Determine the dimension and a basis of the null space of \( A - \lambda_i I \) for each \( i \). If your basis is not orthogonal find an orthogonal basis of the null spaces.

3) Find the matrix for a rotation about the axis span\{(1, 1, 1)\} through \( \pi/2 \) radians in the counterclockwise direction when looking from \((1, 1, 1)\) toward \((0, 0, 0)\).
MODULE 11

Topics: Hermitian and symmetric matrices

Setting: $A$ is an $n \times n$ real or complex matrix defined on $\mathbb{C}^n$ with the

complex dot product $\langle x, y \rangle = \sum_{j=1}^{n} x_j \overline{y_j}$.

Notation: $A^* = A^T$, i.e., $a_{ij} = \overline{a_{ji}}$.

We know from Module 4 that

$\langle Ax, y \rangle = \langle x, A^* y \rangle$

for all $x, y \in \mathbb{C}_n$.

Definition: If $A = A^T$ then $A$ is symmetric.

If $A = A^*$ then $A$ is Hermitian.

Examples: \[
\begin{pmatrix}
1 & i \\
i & 1
\end{pmatrix}
\]
is symmetric but not Hermitian

\[
\begin{pmatrix}
1 & i \\
-i & 1
\end{pmatrix}
\]
is Hermitian but not symmetric.

Note that a real symmetric matrix is Hermitian.

Theorem: If $A$ is Hermitian then

$\langle Ax, x \rangle$ is real for all $x \in \mathbb{C}_n$.

Proof: $\langle Ax, x \rangle = \overline{\langle x, Ax \rangle} = \overline{\langle Ax, x \rangle}$, so the number $\langle Ax, x \rangle$ is real.

Theorem: If $A$ is Hermitian then its eigenvalues are real.

Proof: Suppose that $Au = \lambda u$, then

$\lambda \langle u, u \rangle = \langle \lambda u, u \rangle = \langle Au, u \rangle = \langle u, A^* u \rangle = \langle u, Au \rangle = \langle u, \lambda u \rangle = \lambda \langle u, u \rangle$.

It follows that $\lambda = \overline{\lambda}$ so that $\lambda$ is real.

Theorem: If $A$ is Hermitian then eigenvectors corresponding to distinct eigenvalues are orthogonal.

Proof: Suppose that $Au = \lambda u$ and $Av = \mu v$ for $\lambda \neq \mu$. Then

$\lambda \langle u, v \rangle = \langle Au, v \rangle = \langle u, Av \rangle = \langle u, \mu v \rangle = \mu \langle u, v \rangle$. 

66
Since $\mu \neq \lambda$ it follows that $\langle u, v \rangle = 0$.

Suppose that $A$ has $n$ distinct eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ with corresponding orthogonal eigenvectors $\{u_1, \ldots, u_n\}$. Let us also agree to scale the eigenvectors so that

$$\langle u_i, u_j \rangle = \delta_{ij}$$

where $\delta_{ij}$ is the so-called Kronecker delta

$$\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases}$$

We recall that the eigenvalue equation can be written in matrix form

$$AU = U\Lambda$$

where $U = (u_1 \ u_2 \ \ldots \ u_n)$ and $\Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}$. We now observe from the orthonormality of the eigenvectors that

$$U^*U = I.$$ 

Hence $U^{-1} = U^*$ and consequently

$$A = U\Lambda U^*.$$ 

In other words, $A$ can be diagonalized in a particularly simple way.

**Example:** Suppose

$$A = \begin{pmatrix} 1 & i \\ -i & 1 \end{pmatrix}$$

We saw that $A$ is Hermitian; its eigenvalues are the roots of $(1 - \lambda)^2 - 1 = 0$ so that

$$\lambda_1 = 0, \quad u_1 = (1, i)/\sqrt{2}$$

$$\lambda_2 = 2, \quad u_2 = (1, -i)/\sqrt{2}$$

which shows that $\langle u_1, u_2 \rangle = 0$. Thus

$$U^*U = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -i \\ 1 & i \end{pmatrix} \cdot \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ i & -i \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

and

$$A = U \begin{pmatrix} 0 & 0 \\ 0 & 2 \end{pmatrix} U^*.$$
We saw from the example

\[ A = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \]

that the eigenvector system for \( A \) can only be written in the form

\[ A \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix}. \]

This matrix \( A \) cannot be diagonalized because we do not have \( n \) linearly independent eigenvectors. However, a Hermitian matrix can always be diagonalized because we can find an orthonormal eigenvector basis of \( \mathbb{C}_n \) regardless of whether the eigenvalues are distinct or not.

**Theorem:** If \( A \) is Hermitian then \( A \) has \( n \) orthonormal eigenvectors \( \{u_1, \ldots, u_n\} \) and

\[ A = U \Lambda U^* \]

where

\[ \Lambda = \text{diag}\{\lambda_1, \ldots, \lambda_n\}. \]

**Proof:** If all the eigenvalues were distinct then the results follow from the forgoing discussion. Here we shall outline a proof this result for the case where \( A \) is real so that the setting is \( \mathbb{R}_n \) rather than \( \mathbb{C}_n \). Define the function

\[ f(y) = \langle Ay, y \rangle \]

Analysis tells us that this function has a minimum on the set \( \langle y, y \rangle = 1 \). The minimizer can be found with Lagrange multipliers. We have the Lagrangian

\[ \mathcal{L}(y) = \langle Ay, y \rangle - \lambda (\langle y, y \rangle - 1) \]

and the minimizer \( x_1 \) is found from the necessary condition

\[ \nabla \mathcal{L}(y) = 0. \]

We compute

\[ \frac{\partial \mathcal{L}}{\partial y_k} = \langle A_k, y \rangle + \langle Ay, e_k \rangle - 2\lambda y_k = 2(\langle A_k, y \rangle - \lambda y_k) \]
so that we have to solve
\[ \nabla \mathcal{L} \equiv Ay - \lambda y = 0 \]
\[ \langle y, y \rangle = 1. \]

The solution to this problem is the eigenvector \( x_1 \) with eigenvalue \( \lambda_1 \).

If we now apply the Gram-Schmidt method to the set of \( n + 1 \) dependent vectors \( \{x_1, e_1, \ldots, e_n\} \) we end up with \( n \) orthogonal vectors \( \{x_1, y_2, \ldots, y_n\} \). We now minimize

\[ f(y) = \langle Ay, y \rangle \text{ subject to } \|y\| = 1 \]

over all \( y \in \text{span}\{y_2, \ldots, y_n\} \) and get the next eigenvalue \( \lambda_2 \) and eigenvector \( x_2 \) which then necessarily satisfies \( \langle x_1, x_2 \rangle = 0 \). Next we find an orthogonal basis of \( \mathbb{R}_n \) of the form \( \{x_1, x_2, y_3, \ldots, y_n\} \) and minimize \( f(y) \) over \( \text{span} \{y_3, \ldots, y_n\} \) subject to the constraint that \( \|y\| = 1 \). We keep going and find eventually \( n \) eigenvectors, all of which are mutually orthogonal. Afterwards all eigenvectors can be normalized so that we have an orthonormal eigenvector basis of \( \mathbb{R}_n \) which we denote by \( \{u_1, \ldots, u_n\} \).

If \( A \) is real and symmetric then this choice of eigenvectors satisfies

\[ U^T U = I. \]

If \( A \) is complex then the orthonormal eigenvectors satisfy

\[ U^* U = I. \]

In either case the eigenvalue equation

\[ AU = U \Lambda \]

can be rewritten as

\[ A = U \Lambda U^* \text{ or equivalently, } U^* AU = \Lambda. \]

We say that \( A \) can be diagonalized.

We shall have occasion to use this result when we talk about solving first order systems of linear ordinary differential equations.
The diagonalization theorem guarantees that for each eigenvalue one can find an eigenvector which is orthogonal to all the other eigenvectors regardless of whether the eigenvalue is distinct or not. This means that if an eigenvalue $\mu$ is a repeated root of multiplicity $k$ (meaning that $\det(A - \lambda I) = (\lambda - \mu)^k g(\lambda)$, where $g(\mu) \neq 0$), then $\dim N(A - \mu I) = k$. We simply find an orthogonal basis of this null space. The process is, as usual, the application of Gaussian elimination to

$$(A - \mu I)x = 0$$

and finding $k$ linearly independent vectors in the null space which can then be made orthogonal with the Gram-Schmidt process.

**Some consequences of this theorem:** Throughout we assume that $A$ is Hermitian and that $x$ is not the zero vector.

**Theorem:** $\langle Ax, x \rangle > 0$ if and only if all eigenvalues of $A$ are strictly positive.

**Proof:** Suppose that $\langle Ax, x \rangle > 0$. If $\lambda$ is a non-positive eigenvalue with eigenvector $u$ then $\langle Au, u \rangle = \lambda \langle u, u \rangle \leq 0$ contrary to assumption. Hence the eigenvalues must be positive. Conversely, assume that all eigenvalues are positive. Let $x$ be arbitrary, then the existence of an orthonormal eigenvector basis $\{u_j\}$ assures that

$$x = \alpha_1 u_1 + \cdots + \alpha_n u_n.$$ 

If we substitute this expansion into $\langle Ax, x \rangle$ we obtain

$$\langle Ax, x \rangle \sum_{j=1}^{n} \lambda_j \alpha_j \overline{\alpha_j} > 0$$

for any non-zero $x$.

**Definition:** $A$ is positive definite if $\langle Ax, x \rangle > 0$ for $x \neq 0$.

$A$ is positive semi-definite if $\langle Ax, x \rangle \geq 0$ for all $x$.

$A$ is negative definite if $\langle Ax, x \rangle < 0$ for $x \neq 0$.

$A$ is negative semi-definite if $\langle Ax, x \rangle \leq 0$ for all $x$.

Note that for a positive or negative definite matrix $Ax = 0$ if and only if $x = 0$ while a semi-definite matrix may have a zero eigenvalue.
An application: Consider the real quadratic polynomial

\[ P(x, y, z) = a_{11}x^2 + a_{12}xy + a_{13}xz + a_{22}y^2 + a_{23}yz + a_{33}z^2 + b_1x + b_2y + b_3z + c \]

We “know” from analytic geometry that \( P(x, y, z) = 0 \) describes an ellipsoid, paraboloid or hyperboloid in \( \mathbb{R}^3 \). But how do we find out? We observe that \( P(x, y, z) = 0 \) can be rewritten as

\[ P(x, y, z) = \left\langle A \begin{pmatrix} x \\ y \\ z \end{pmatrix}, \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\rangle + \left\langle b, \begin{pmatrix} x \\ y \\ z \end{pmatrix} \right\rangle c = 0 \]

where

\[ A = \begin{pmatrix} a_{11} & a_{12}/2 & a_{13}/2 \\ a_{12}/2 & a_{22} & a_{23}/2 \\ a_{13}/2 & a_{23}/2 & a_{33} \end{pmatrix} \]

is a symmetric real matrix. Hence \( A \) can be diagonalized

\[ A = U\Lambda U^T \]

Define the new vector (new coordinates)

\[ \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} = U^T \begin{pmatrix} x \\ y \\ z \end{pmatrix} \]

Then

\[ P(X, Y, X) = \left\langle \Lambda \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}, \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \right\rangle + \left\langle b, U \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \right\rangle c = \lambda_1 X^2 + \lambda_2 Y^2 + \lambda_3 Z^2 + \text{lower order terms} \]

where \( \{\lambda_j\} \) are the eigenvalues of \( A \). If all are positive or negative we have an ellipsoid, if all are non-zero but not of the same sign we have a hyperboloid, and if at least one eigenvalue is zero we have a paraboloid. The lower order terms determine where these shapes are centered but not their type.
Module 11 Homework

1) Suppose that $A$ has the property $A = -A^*$. In this case $A$ is said to be skew-Hermitian.
   i) Show that all eigenvalues of $A$ have to be purely imaginary.
   ii) Prove or disprove: The eigenvectors corresponding to distinct eigenvalues of a skew-Hermitian matrix are orthogonal with respect to the complex dot product.
   iii) If $A$ is real and skew-Hermitian what does this imply about $A^T$?

2) Let $A$ be a skew-Hermitian matrix.
   i) Show that for each $x \in \mathbb{C}_n$ we have $\text{Re}\langle Ax, x \rangle = 0$
   ii) Show that for any matrix $A$ we have

   $$\langle Ax, x \rangle = \langle Bx, x \rangle + \langle Cx, x \rangle$$

   where $B$ is Hermitian and $C$ is skew-Hermitian.

3) Suppose that $U^*U = I$. What can you say about the eigenvalues of $U$?

4) Find a new coordinate system such that the conic section

   $$x^2 - 2xy + 4y^2 = 6$$

   is in standard form so that you can read off whether it is an ellipse, parabola or hyperbola.
Let \( \| \cdot \| \) denote a norm on \( \mathbb{R}^m \) and \( \mathbb{R}^n \). Typically, we think of \( \| x \|_\infty = \max_i |x_i| \), but it can be any norm.

We define the vector norm of a matrix \( A \) by

\[
\| A \| = \max_{\| x \|=1} \| Ax \|.
\]

We say that the vector norm \( \| A \| \) is “induced” by the norm \( \| \cdot \| \). It is a measure of the “size” of the operator.

It is straightforward to show that this definition yields a norm on the vector space of all \( m \times n \) matrices. Moreover, for any vector \( x \neq 0 \) we have that

\[
\| Ax \| = \left\| A \frac{x}{\| x \|} \right\| \| x \| \leq \| A \| \| x \|
\]

and

\[
\| ABx \| \leq \| A \| \| B \| \leq \| A \| \| B \| \| x \|
\]

so that

\[
\| AB \| \leq \| A \| \| B \|.
\]

Finding the actual number for the norm of a matrix may be complicated for some norms on \( \mathbb{R}^n \). However, for the infinity norm it is easy.

**Terminology:** Given an \( m \times n \) matrix \( A \) its maximum row sum is the number

\[
R = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|.
\]

**Theorem:** The vector norm of a matrix \( A \) induced by the infinity norm is equal to its maximum row sum.

**Proof:** Let \( \| x \|_\infty = 1 \) then by definition \( |x_i| \leq 1 \) and \( |x_k| = 1 \) for some \( k \). Then

\[
\| Ax \|_\infty = \max_i \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \max_i \sum_{j=1}^{n} |a_{ij}| = R.
\]
Hence the maximum row sum is always greater than or equal to the infinity vector norm of $A$.

Conversely, suppose the maximum row sum is obtained from row $k$ of the matrix $A$. Then choose the vector $x$ defined by

$$x_j = 1 \quad \text{if } a_{kj} \geq 0$$

$$x_j = -1 \quad \text{if } a_{kj} < 0.$$  

Then $\|x\|_{\infty} = 1$ and

$$\|A\|_{\infty} \geq \|Ax\|_{\infty} \geq \left| \sum_{j=1}^{n} a_{kj}x_j \right| = \sum_{j=1}^{n} |a_{kj}| = R.$$  

Hence here we have a specific vector of length 1 for which the vector norm $A$ dominates the maximum row sum. Therefore,

$$\|A\|_{\infty} = R.$$  

**Application:** Suppose that the $n \times n$ matrix $A$ is strictly diagonally dominant, i.e.

$$|a_{ii}| > \sum_{\substack{j=1\atop j \neq i}}^{n} |a_{ij}| \quad \text{for } i = i, \ldots, n.$$  

Suppose we want to solve $Ax = b$.

First we observe that if $\lambda$ is an eigenvalue of $A$ with eigenvector $x$ then we can scale the eigenvector so that its maximum component is equal to +1 for some component $k$. Then it follows from the eigenvalue equation

$$(a_{kk} - \lambda)x_k + \sum_{\substack{j=1\atop j \neq k}}^{n} a_{kj}x_j = 0$$  

and strict diagonal dominance that

$$|a_{kk} - \lambda| \leq \left| \sum_{\substack{j=1\atop j \neq k}}^{n} a_{kj}x_j \right| \leq \sum_{\substack{j=1\atop j \neq k}}^{n} |a_{kj}| < |a_{kk}|.$$  

This strict inequality implies that $\lambda \neq 0$. Hence the null space of $A$ contains only the zero vector so that

$$Ax = b$$
has a unique solution.

The solution can be found iteratively. We write

\[ A = D - B \]

where \( D \) is the diagonal of \( A \) and \( B = A - D \). The solution \( x^* \) of \( Ax = b \) satisfies the equation

\[ x = D^{-1}Bx + D^{-1}b \]

where \( D^{-1} = \text{diag}\{1/a_{11}, 1/a_{22}, \ldots, 1/a_{nn}\} \).

We shall find it from the so-called Jacobi iteration

\[ x^{k+1} = D^{-1}Bx^k + D^{-1}b \]

where \( x^0 \) is an initial guess. The advantage of such iterative solution is its applicability to huge linear systems where the entries of \( A \) are mostly zero and do not enter into the actual computation. However, the iteration will not always converge. The next theorem gives insight when the iteration will work.

**Theorem:** For a strictly diagonally dominant matrix \( A \) the Jacobi iteration converges to the unique solution \( x^* \) of \( Ax = b \).

**Proof:** Let \( e^k \) denote the error in iteration \( k \)

\[ e^k = x^k - x^*. \]

Then

\[ e^{k+1} = D^{-1}Be^k = (D^{-1}B)^{k+1}e^0 \]

so that

\[ \|e^k\| = \|(D^{-1}B)^ke^0\| \leq \|(D^{-1}B)^k\|\|e^0\|. \]

If \( (D^{-1}B)^k \to 0 \) as \( k \to \infty \) then \( e^k \to 0 \) so that

\[ \lim_{k \to \infty} x^k = x^*. \]
It follows from the vector norm properties of matrices that
\[ \| (D^{-1}B)^k \| \leq \| D^{-1}B \|^k. \]

From the above theorem we know that
\[ D^{-1}B = \max_i \left| \sum_{j=1, j \neq i}^n a_{ij} \right|. \]

Strict diagonal dominance implies that
\[ \| D^{-1}B \|_\infty \leq R < 1. \]

Thus
\[ \| e^k \| < R^k \| e^0 \| \rightarrow 0 \quad \text{as} \quad k \rightarrow \infty. \]

The condition \( \| A \| < 1 \) for a square matrix \( A \) in some vector norm insures that
\[ \lim_{k \to \infty} A^k = 0 \quad (\text{the zero matrix}) \]
because
\[ 0 \leq \| A^k \| < \| A \|^k. \]

However, it may well be that in a particular norm
\[ \| A \| > 1 \quad \text{and yet that} \quad \lim_{k \to \infty} A^k \rightarrow 0. \]

For example,
\[ A = \begin{pmatrix} 0 & 10^{10} \\ 0 & 0 \end{pmatrix} \quad \text{satisfies} \quad \| A \|_\infty = 10^{10} \]
but \( A^k = 0 \) for \( k \geq 2 \).

The relevant question for applications is: Given a square matrix \( A \) what is the choice of vector norm for which \( \| A \| \) is smallest?

We observe first that for any eigenvalue and eigenvector of \( A \) we have
\[ \| Ax \| = |\lambda| \| x \|. \]
regardless of the norm chosen. Moreover, if in this norm the eigenvector is scaled to have length 1 then it follows that

$$|\lambda| = \|Ax\| \leq \|A\|.$$ 

Hence in any vector norm chosen for $R^*_n$ the associated norm of $A$ must satisfy

$$\rho(A) = \max_i |\lambda_i| \leq \|A\|.$$ 

$\rho(A)$ is known as the spectral radius of the matrix $A$ and represents a lower bound on any vector norm of $A$.

There are infinitely many norms which can be imposed on $R^*_n$ and each induces a vector norm on the $n \times n$ matrices. For example, if we define $\|x\| = \|Cx\|_p$ for $p \in [1, \infty]$ for a given non-singular matrix $C$ then $\| \|$ is a norm on $R^*_n$. Now, if $A$ is an $n \times n$ matrix and we consider the vector norm induced by $\| \|$; then we have by definition

$$\|A\| = \max_{\|x\|=1} \|Ax\| = \max_{\|x\|=1} \|CAx\|_p = \max_{\|Cx\|_p=1} \|CAx\|_p.$$ 

But for each such $x$ there is a unit vector $y$ such that $Cx = y$. Then

$$\|A\| = \max_{\|y\|_p=1} \|CAC^{-1}y\|_p$$

so that

$$\|A\| = \|CAC^{-1}\|_p.$$ 

The transformation of $A$ into $CAC^{-1}$ is called a similarity transformation. We have seen above if the $n \times n$ matrix $A$ has $n$ linearly independent eigenvectors then it follows from the eigenvector equations

$$AX = X\Lambda$$

where the $j$th column of $X$ is the eigenvector corresponding to the eigenvalue $\lambda_j$ that in the norm

$$\|x\| = \|X^{-1}x\|_p$$

the induced matrix norm $\|A\|$ is

$$\|A\| = \|X^{-1}AX\|_p = \|\Lambda\|_p.$$
In particular, if \( p = \infty \) then \( \|A\| \) is equal to the maximum row sum of \( \Lambda \) so that

\[
\|A\| = \rho(A).
\]

In this case the size of \( A \) is equal to its spectral radius and \( A^k \to 0 \) as \( k \to \infty \) whenever \( \rho(A) < 1 \). This analysis applies to matrices with \( n \) distinct eigenvalues and to Hermitian matrices.

Not every square matrix is similar to a diagonal matrix. However, it is possible, but not easy, to prove via the Jordan canonical form that for any \( \epsilon > 0 \) the matrix \( A \) can be transformed with a similarity transformation into a matrix whose diagonal entries are the \( n \) eigenvalues of \( A \), and whose entries \( a_{i-1,i} \) are either 0 or \( \epsilon \). Thus, even when \( A \) is not diagonalizeable there is a norm on \( \mathbb{R}^n \) such that

\[
\|A\| = \rho(A) + \epsilon.
\]

This implies that if \( \rho(A) < 1 \) then for sufficiently small \( \epsilon > 0 \)

\[
\|A\| < 1
\]

so that again \( A^k \to 0 \) as \( k \to \infty \). The matrix \( C \) in this similarity transformation is generally not available but also not needed for the conclusion that \( A^k \to 0 \).

Hence the smallest possible vector norm of a matrix \( A \) is roughly equal to its spectral radius.
We shall be concerned with mappings = functions = operators which map a function space into a function space. We shall write, sort of generically and in analogy to $Lx \equiv Ax$,

$$Lu \equiv L(u, u', u'', \ldots, u^{(n)}, t)$$

where $Lu(t)$ is either one function or a vector of functions. Here $t$ is the independent variable. Often it denotes time, but sometimes the independent variable may denote a spatial variable, the cost of a process, or whatever. $n$ denotes the highest derivative which has to be evaluated and is the “order” of the differential operator. If $Lu$ is vector valued we speak of a system of differential operators. In this course we shall concentrate on two types of differential operators:

i) $Lu = a(t)u'' + b(t)u' + c(t)u$ whose domain is $C^2[a, b]$ or some subspace thereof. $Lu$ is a second order (scalar) operator, because $Lu(t)$ is a single function of $t$ and not a vector valued function. For example, if

$$Lu \equiv u'' + u$$

is defined on $C^2[0, 1]$ then

$$L((2 - t)^{-1}) \equiv (2 - t)^{-3} + (2 - t)^{-1}$$

ii) $Lu \equiv u'(t) - F(u, t)$ where $u(t) = (u_1(t), \ldots, u_n(t))$, and

$$F(u, t) = \begin{pmatrix} f_1(u_1, \ldots, u_n, t) \\ \vdots \\ f_n(u_1, \ldots, u_n, t) \end{pmatrix}$$

In this setting $L$ is known as a system of first order operators which would be defined on a space of vectors $\vec{u}(t) = (u_1(t), \ldots, u_n(t))$ where each $u_i(t)$ is a continuously differentiable function. For example, if

$$Lu(t) \equiv \begin{pmatrix} u_1'(t) \\ u_2'(t) \end{pmatrix} + \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}$$

then

$$L \begin{pmatrix} \sin t \\ \cos t \end{pmatrix} \equiv \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$
The objective of our study will be to solve the differential equation

\[(Lu)(t) = f(t)\]

associated with the operator \(L\), where \(f(t)\) is a given source term. In analogy to the matrix equation \(Ax = b\) we should expect a solution only when \(f\) belongs to the range of \(L\). However, we need to clarify what we mean by a “solution.”

**Definition:** A (classical) solution of the \(n\)th order ordinary differential equation

\[Lu = f(t)\]

is an \(n\)-times continuously differentiable function of \(t\) which satisfies the differential equation in some interval \((a, b)\).

The fact that we specifically define a classical solution suggests that there are other solutions. There are indeed strong solutions, weak solutions and distributional solutions which in general satisfy the differential equation only in some “average” (integrated) sense. We shall be concerned only with a classical solution which we shall call henceforth simply a “solution.”

Much of the literature on ordinary differential equations deals with first order systems. In particular, numerical methods are usually given only for such systems. The reason is that any \(n\)th order equation can be reduced to a first order system. For example, consider the general \(n\)th order scalar equation

\[L(u, u’, \ldots, u^{(n)}, t) = f(t)\]

If we introduce new functions \(\{v_1, v_2, \ldots, v_n\}\) defined by

\[
\begin{align*}
v_1(t) &= u(t) \\
v_2(t) &= u'(t) \\
\vdots \\
v_n(t) &= u^{(n-1)}(t)
\end{align*}
\]

then the \(n\)th order equation is equivalent to the \(n\)-dimensional first order system

\[
\begin{align*}
L_1 v &= v_1' - v_2(t) = 0 \\
\vdots \\
L_{n-1} v &= v_{n-1}' - v_n(t) = 0 \\
L_n(v_1, v_2, \ldots, v_n, v'_n) &= f(t)
\end{align*}
\]
where \( L = (L_1, \ldots, L_n) \) is vector valued. Hence any \( n \)th order equation can be reduced to a first order system. The reverse is generally not true. There is no general method for transforming a first order system into a single \( n \)th order equation.

Loosely speaking, the solution to a differential equation contains as many degrees of freedom as the number of derivatives. We can think of these parameters as the constants of integration which arise when we solve the differential equation. In order to pin down these parameters additional conditions are imposed on the solution \( u(t) \) of \( Lu = f \). The most common form is to prescribe \( u \) and some of its derivatives at a single point \( t_0 \). In this case we speak of an initial value problem. If \( u \) and some of its derivatives are given at more than one point then we have a boundary value problem. In general, boundary value problems are more difficult to solve than initial value problems.

Existence and uniqueness of the solution for an initial value problem: We shall consider the general case of a first order system, written in the form

\[
\frac{du}{dt}(t) = F(u, t),
\]

and subject to the initial value

\[
u(t_0) = u_0
\]

where \( t_0 \) and the vector \( u_0 \) are given. In the shorthand notation introduced above this means that \( L \) can be solved for the highest derivative, which is often the case in applications. When can we guarantee that there is a function \( u \) which satisfies the equation and takes on the given value at \( t_0 \)?)

**Theorem:** Suppose that in a neighborhood of \((u_0, t_0)\) the function \( F \) is continuous in \( t \) and continuously differentiable with respect to \( u \). Then there exists a unique solution of the problem over an interval \((t_0 - \epsilon, t_0 + \epsilon)\) for some \( \epsilon > 0 \).

We shall not prove this theorem, but some additional comments are in order. First of all, since \( F \) is a vector valued function with \( n \) components \((f_1, \ldots, f_n)\) in the vector variable \((u_1, \ldots, u_n)\) the continuous differentiability of \( F \) with respect to \( u \) means that all the partial derivatives

\[
\frac{\partial f_i}{\partial u_j} \quad i, j = 1, \ldots, n
\]
are continuous functions in a neighborhood of \((u_0, t_0)\). Secondly, the existence is guaranteed only for \(t\) in some interval around \(t_0\). As one moves away from \(t_0\) the solution may well blow up or enter a region when \(F'(u, t)\) is no longer defined. For example, consider the following cases:

i) \(u' = F(u, t) = 1 - u^2, \ u(0) = 0\)

We note that for this (one dimensional) system the function \(F\) is constant with respect to \(t\) and \(\partial F/\partial u = -2u\) which is continuous in \(u\) everywhere. It is straightforward to verify that

\[
 u(t) = \tanh t
\]

satisfies the differential equation and the initial condition. Obviously, this solution exists for all \(t \in (-\infty, \infty)\).

ii) \(u' = F(u, t) = 1 + u^2, \ u(0) = 0\).

This function \(F\) has the same continuity properties as in case i), but the solution of the problem is now seen to be

\[
 u(t) = \tan t
\]

which exists only on \((-\pi/2, \pi/2)\). Hence local existence is the best one can hope for in general.

iii) \(u'(t) = \sqrt{|u|}, \ u(0) = 0\).

The right hand side of the equation is continuous in \(u\) but not differentiable with respect to \(u\) in a neighborhood of \(u = 0\). By inspection we see that

\[
 u(t) = 0
\]

is a solution. However, we can verify that

\[
 u(t) = (t/2)^2
\]

also solves the equation and satisfies the initial condition. Hence the lack of differentiability with respect to \(u\) leads to multiple solutions.

We note in this context that our existence and uniqueness theorem is not stated in its strongest form. It suffices for the existence of a solution to have continuity in \(t\) and \(u\), and
for the uniqueness to have a so-called Lipschitz continuity in $u$. However, this latter type of continuity is implied by differentiability which is much easier to verify.

In this course the continuity and differentiability properties of $F$ are always assumed so that we can expect a unique local solution. Rather than specifying exactly how many derivatives $F$ has we shall simply say that $F$ is “smooth.” This is a convenient (if imprecise) way to express that $F$ is expected to have as many derivatives as the application of a theorem or a calculation demands.

For example, if $F$ is smooth then the solutions of

$$u' = F(u, t), \quad u(t_0) = u_0$$
$$v' = F(v, t), \quad v(t_0) = v_0$$

for $u_0 \neq v_0$ can never satisfy

$$u(t) = v(t) \quad \text{for any value of } t$$

because if they were to cross at a given $t_1$ then this would imply that $u(t)$ and $v(t)$ are both solutions of the initial value problem

$$w' = F(w, t), \quad w(t_1) = u(t_1) = v(t_1)$$

which has a unique solution so that $w(t) = u(t) = v(t)$ for all $t$ in its interval of existence.

We shall cite one additional theorem for the system

$$u' = F(u, t, \alpha)$$
$$u(t_0) = u_0$$

where $\alpha$ is a parameter.

**Theorem:** If $F$ is smooth in $u$ and $t$ and continuous in $\alpha$ then the solution $u(t)$ depends continuously on $\alpha$, $u_0$ and $t_0$.

In other words, as we change parameters of the problem smoothly the solution also changes smoothly. To indicate the dependence of $u$ on these parameters the solution is sometimes denoted by $u(t, t_0, u_0, \alpha)$. For example, suppose we would like to solve the boundary value problem

$$u'_1 = f_1(u_1, u_2, t), \quad u_1(0) = a$$
\[ u_2' = f_2(u_1, u_2, t), \quad u_2(1) = b \]

then we can solve this system subject to the initial value

\[ u_1(0) = a \]
\[ u_2(0) = s \]

where \( s \) is a parameter. The task then is to search for that value of \( s \) for which the solution \( u_2(t, s) \) satisfies \( u_2(1, s) = b \). A numerical method based on this idea is the so-called “shooting method,” which is quite effective and versatile for many different boundary value problems.

**Analytic solution of first order equations:**

Very few differential equations can be solved analytically. If one needs an analytic solution of an equation which cannot be solved with the standard methods discussed here then one may try to discover the solution in the various tables of solutions for differential equations which have been assembled over the years (see, e.g., Kamke, E. GT Library No. QA371 .K2814). Here we can discuss only a single first order equation and later, in some depth, linear systems.

The analytic solution of a “separable” first order equation: Suppose we wish to solve an equation of the form

\[ u'(t) = \frac{f(t)}{g(u)} \quad u(t_0) = u_0, \]

i.e., the right hand side can be factored into the product of a function of \( t \) and a function of \( u \). Such an equation is called separable. Suppose further that for \( g \) and \( f \) we can find antiderivatives \( F \) and \( G \) such that

\[ \int f(t)dt = F(t) \]

and

\[ \int g(u)du = G(u). \]

Then the equation

\[ G(u) = F(t) \]
implicitly defines \( u \) as a function of \( t \), i.e.,

\[ G(u(t)) = F(t). \]

Implicit differentiation shows that

\[ \frac{d}{dt} G(u(t)) = g(u(t)) \frac{du}{dt} = f(t) \]

so that \( u(t) \) is a solution of the differential equation. Note that the antiderivatives are determined only up to an arbitrary constant which can be determined such that

\[ G(u_0) = F(t_0). \]

If one can solve

\[ G(u(t)) = F(t) \]

explicitly for \( u(t) \) we have an explicit solution of the differential equation. Otherwise all one has is a relation linking \( u \) and \( t \) which is consistent with the differential equation. Operationally, the separable equation is written as

\[ \int g(u) du = \int f(t) dt, \]

both sides are integrated and a constant of integration is added to the right side. One obtains \( G(u) = F(t) \) and then solves for \( u \) as a function of \( t \), if possible.

Unfortunately, it many instances one can find either no antiderivatives or no explicit function \( u(t) \) so that one is forced into numerical methods. Nonetheless, separable equations arise quite commonly and afford a chance for finding an analytic solution, which is always preferable to an approximate solution.

**Examples:**

1) \( u' = 1 + u^2 \), \( u(1) = 0 \) is rewritten as

\[ \int \frac{du}{1 + u^2} = \int dt. \]

Integration yields

\[ \tan^{-1} u = t + c \]
The initial condition requires

\[ \tan 0 = 1 + c \]

or \( c = -1 \). Solving for \( u \) we find

\[ u(t) = \tan(t - 1). \]

2) \( u' = \sqrt{u}, \ u(0) = 0 \)

\[
\int \frac{du}{u^{1/2}} = \int dt \\
2u^{1/2} = t + c \\
0 = 0 + c \\
u(t) = (t/2)^2
\]

3) \( u' = a(t)u, \ u(t_0) = u_0 \) describes exponential growth. We write

\[
\int \frac{du}{u} = \int a(t)dt \\
\ln |u| = \int_{t_0}^{t} a(s)ds + c
\]

(Differentiation shows that the right hand side is an antiderivative of \( a(t) \).) Solving for \( u \) we obtain

\[ |u(t)| = e^{\int_{t_0}^{t} a(s)ds + c} \]

or

\[ u(t) = ke^{\int_{t_0}^{t} a(s)ds} = u_0e^{\int_{t_0}^{t} a(s)ds} \]

4)

\[ u'(t) = \frac{t}{\ln u}, \quad u(1) = 3 \]

\[
\int \ln u \, du = \int t \, dt \\
u \ln u - u = t^2/2 + c
\]

where

\[ 3 \ln 3 - 3 = 1/2 + c \]
determines \( c \). We cannot solve for \( u \) as a function of \( t \). However, here we can solve for \( t \) as a function of \( u \) which is helpful for plotting the curve \((t,u)\) in the \( t-u \) plane.

5) The equation
\[
u' = au - bu^2 = u(a - bu), \quad u(t_0) = u_0, \quad a, b > 0,
\]
is the so-called logistic equation and describes growth or decay in a finite environment. We see that if \( 0 < u_0 < a/b \) then \( u' > 0 \) so that \( u \) can grow until \( u_\infty = a/b \). If \( u_0 > a/b \) then \( u \) will decay to \( u_\infty = a/b \). This is a separable equation. We write the partial fraction expansion
\[
\int \frac{du}{u(a - bu)} = \int \frac{du}{au} + \int \frac{b du}{a(a - bu)} = \int dt
\]
which integrates into
\[
\ln \left| \frac{u}{a - bu} \right| = at + c.
\]
Solving for \( u \) we obtain
\[
u(t) = \frac{a}{b + ke^{-at}}
\]
where \( k \) is determined from the initial condition.

NOTE: IN ALL EXAMPLES THE CONSTANT OF INTEGRATION APPEARS AT THE TIME THE INTEGRATION IS CARRIED OUT. DO NOT STICK IT IN AS AN AFTERTHOUGHT IN SOME RANDOM PLACE AT SOME RANDOM TIME IN YOUR CALCULATION.
Module 12 - Homework

1) Solve analytically
   i) \( u' = tu\sqrt{1+u}, \ u(1) = 1 \)
   ii) \( u' = \frac{u}{1+t^2}, \ u(0) = 1 \)
   iii) \( u' = te^u, \ u(0) = 1 \)
   iv) \( u' = 2t^2 + u - t^2u + tu - 2t - 2, \ u(0) = 0 \)

   In each case state for what \( t \) the solution exists and find \( \lim_{t \to \pm\infty} u(t) \), if possible.

2) Think of the solution \( u(t) \) of the \( n \)-dimensional system
   \[ u' = F(u, t) \]
   \[ u(t_0) = u_0 \]
   as a position vector in \( \mathbb{E}_n \) which traces out a trajectory. Find the equation of the straight line tangent to the trajectory at a given time \( t_1 \).

3) Using the existence and uniqueness theorem for a system of first order equations prove an existence and uniqueness theorem for
   \[ Lu = u'' + a_1(t)u' + a_0(t)u = f(t). \]
   (You decide what the correct corresponding initial conditions are.)

4) Let \( u \) and \( v \) be solutions of \( Lu = f(t) \) where \( L \) is the linear operator of problem 3. Suppose
   \[ u(t_0) \neq v(t_0) \]
   for some \( t_0 \). Prove or disprove:
   i) \( u(t) \neq v(t) \) for all \( t \)
   ii) \( u(t) = v(t) \)
      \[ u'(t) = v'(t) \] for some \( t \).

5) Suppose you invest \( M \) in a savings account which pays an interest rate of \( r \) percent per year, compounded continuously. Suppose you withdraw money at a constant rate of \( C \) dollars/year. What should \( C \) be so that after 5 years you have exactly \( M/2 \) in your account?

6) Suppose you buy a bond today which will pay exactly $100 in two years. The cost of the bond today is $87.00. What is the effective interest rate for your investment?
We shall consider linear operators and the associated linear differential equations. Specifically we shall have operators of the form

i) \( Lu \equiv u' - A(t)u \)

where \( A(t) \) is an \( n \times n \) matrix with continuous functions of \( t \) in its entries and \( u = (u_1, \ldots, u_n) \).

ii) \( Lu \equiv \sum_{j=0}^{n} a_j(t)u^{(j)}(t) \).

In most applications this \( L \) will operate on a single function \( u(t) \). We assume that the coefficient of the \( j \)th derivative is continuous in \( t \), and that \( a_n(t) \neq 0 \) for any \( t \) in the interval of interest. Hence without loss of generality we may assume that \( a_n(t) \equiv 1 \).

Both operators have in common that they are linear, i.e.,

\[
L(u + \alpha v) = Lu + \alpha Lv.
\]

In either case we shall consider the differential equation

\[
Lu = 0
\]

In other words, we are going to examine the null space of \( L \) in the domain of \( L \) consisting of all sufficiently smooth functions.

As we noted before, the \( n \)th order equation can be transformed into a first order system if we set

\[
\begin{align*}
v_1 &= u \\
v_2 &= u' \\
\cdots \\
v_n &= u^{(n-1)}
\end{align*}
\]

then the first order system equivalent to the single \( n \)th order scalar equation is

\[
Lv \equiv v' - A(t)v = 0
\]

where

\[
A(t) = \begin{pmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
-a_0 & -a_1 & -a_2 & \cdots & -a_{n-1}
\end{pmatrix}
\]
The equivalence between the \( n \)th order scalar equation and the first order system will be exploited time and again. We also note that the equation \( Lu = 0 \) for a first order system will usually be written as

\[
  u' = A(t)u
\]

so that we can apply the above existence and uniqueness theorem.

**Theorem:** \( \dim \mathcal{N}(L) = n \).

**Proof.** Consider the \( n \) initial value problems

\[
  u'_i = A(t)u_i \\
  u_i(0) = \hat{e}_i, \quad i = 1, \ldots, n.
\]

It follows from our existence and uniqueness theorem that the functions \( \{u_i(t)\} \) exist, at least in a neighborhood of \( t = 0 \). Without proof we shall accept that if the coefficients are bounded for all \( t \) then these solutions in fact exist for all \( t \). These solutions are linearly independent, for if

\[
  \sum_{j=1}^{n} \alpha_j u_j(t) = U(t)\alpha = 0
\]

where \( U \) is the matrix whose \( i \)th column is \( u_i(t) \), then also

\[
  U(0)\alpha = 0.
\]

But \( U(0) = I \), hence \( \alpha = (\alpha_1, \ldots, \alpha_n) = 0 \). Now suppose that \( w(t) \) is any vector \( \in \mathcal{N}(L) \). Then by linearity the function

\[
  v(t) = U(t)w(0)
\]

satisfies

\[
  v'(t) = A(t)v(t) \\
  v(0) = w(0).
\]

Uniqueness guarantees that \( v(t) \equiv w(t) \). Hence any function in the null space of \( L \) is representable as a linear combination of the solutions \( \{u_i\} \). Hence these \( n \) functions form a basis of \( \mathcal{N}(L) \).
Example: If
\[ u' = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} u \]
then
\[ u_1(t) = \begin{pmatrix} \cosh t \\ \sinh t \end{pmatrix}, \quad u_2(t) = \begin{pmatrix} \sinh t \\ \cosh t \end{pmatrix}. \]
By inspection,
\[ w(t) = \begin{pmatrix} e^{-t} \\ -e^{-t} \end{pmatrix} \in \mathcal{N}(L) \]
and \( 2w(t) = u_1(t) - u_2(t). \)

What does this theorem tell us about
\[ Lu \equiv \sum_{j=0}^{n} a_j(t) u^{(j)}(t) = 0 \]
The corresponding vector system \( v' = A(t)v \) has \( n \) linearly independent solutions \( v_i(t) \), \( i = 1, \ldots, n \). The first component of each vector \( v_i(t) \) is a solution of
\[ Lu = 0. \]
The remaining components are the derivatives of \( u \). If we set
\[ u_i(t) = \text{first component of } v_i(t) \]
then
\[ Lu_i(t) = 0 \]
and
\[ u_i^{(k)}(0) = \delta_{i,k+1}. \]
In other words, \( u_1(0) = 1 \) and all derivatives up to order \( (n-1) \) are zero at \( t = 0 \), \( u_2(0) = 0 \), \( u_2'(0) = 1 \), and all other derivatives of \( u_2 \) are zero at \( t = 0 \). If we look at the Wronskian of these \( n \) functions we see from
\[ W(t) = \det \begin{pmatrix} u_1(t) & u_2(t) & \cdots & u_n(t) \\ \cdots & \cdots & \cdots & \cdots \\ u_1^{(n-1)}(t) & u_2^{(n-1)}(t) & \cdots & u_n^{(n-1)}(t) \end{pmatrix} = \det(v_1 \ v_2 \ \ldots \ v_n) \]
that $W(0) = 1$ so that the functions $\{u_i(t)\}$ are linearly independent. The uniqueness of solutions for an initial value problem can then be used to demonstrate that these functions are a basis of $\mathcal{N}(L)$. We summarize:

i) The equation $u' = A(t)u$ has $n$ linearly independent solutions $\{u_i\}$ and any solution of this system can be written as a linear combination of the $\{u_i\}$.

ii) The equation

$$Lu = \sum_{j=0}^{n} a_j(t)u^{(j)}(t) = 0$$

has $n$ linearly independent solutions $\{u_i\}$ and any solution of $Lu = 0$ can be written as a linear combination of the $\{u_i\}$.

**Example:** Consider the boundary value problem

$$Lu \equiv u'' + u = 0$$

$$u(0) + u(1) = 1$$

$$u'(2) = 2.$$ 

We verify that two linearly independent solutions of $Lu = 0$ are

$$u_1(t) = \cos t$$

and

$$u_2(t) = \sin t.$$ 

A solution $u(t)$ of our boundary value problem, if it exists, must belong to the span$\{u_1(t), u_2(t)\}$, i.e.,

$$u(t) = \alpha_1 u_1(t) + \alpha_2 u_2(t).$$

We satisfy the boundary conditions if

$$\begin{pmatrix} 1 + \cos 1 & \sin 1 \\ -\sin 2 & \cos 2 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}. $$

The system is not singular. $\alpha_1$ and $\alpha_2$ can be found, hence a solution $u(t)$ exists. Moreover, this solution is unique because if there were two solutions $v(t)$ and $w(t)$ then the function

$$u(t) = v(t) - w(t)$$
would have to satisfy

\[ Lu = 0 \]

\[ u(0) + u(1) = 0 \]

\[ u'(2) = 0. \]

But the only function in the span of \( u_1(t) \) and \( u_2(t) \) which satisfies these boundary conditions is the zero function.

The dominant practical difficulty in solving linear equations is the calculation of the basis functions. Only for constant coefficient systems is there a consistent way of generating the basis. This topic will be studied at length a little later. However, given one element in the null space of an \( n \)th order linear operator one can often find a second linearly independent element from a related simpler problem. We shall discuss only the case of a second order equation of the form

\[ Lu \equiv a_2(t)u'' + a_1(t)u + a_0(t)u = 0 \]

where the coefficients are continuous in \( t \) on some interval. Here the computation proceeds as follows:

Let \( u_1(t) \) be an element in \( \mathcal{N}(L) \), i.e.,

\[ Lu_1(t) = 0 \]

then one can find a second element \( \in \mathcal{N}(L) \) of the form

\[ u_2(t) = \phi(t)u_1(t) \]

where \( \phi \) is found from (essentially) a first order separable equation. Since we want \( Lu_2 = 0 \) we find that \( \phi \) must be chosen so that

\[ a_2(t)[\phi''u_1 + 2\phi'u_1' + \phi u_1'''] + a_1(t)[\phi'u_1 + \phi u_1'] + a_0(t)\phi u_1 = 0. \]

Since \( Lu_1 = 0 \) this equation simplifies to

\[ a_2(t)u_1\phi'' + [2a_2(t)u_1' + a_1(t)u_1]\phi' = 0. \]
We now have a first order separable equation in $\psi \equiv \phi'$ which we can write as

$$\frac{\psi'}{\psi} = \frac{2a_2(t)u_1'(t) + a_1(t)u_1(t)}{a_2(t)u_1(t)}$$

and which has a non-zero exponential solution. The function $\phi$ is then obtained by integrating $\psi$.

**Example:** Consider

$$Lu \equiv t^2u'' + 3tu' + u = 0.$$ 

It is known that one can find a solution of the form

$$u_1(t) = t^\alpha.$$ 

If we substitute this function into $Lu = 0$ we find that $\alpha$ must be chosen such that

$$[\alpha(\alpha - 1) + 3\alpha + 1]t^\alpha = 0,$$

i.e., $\alpha^2 + 2\alpha + 1 = 0$,

which yields the solution $\alpha = -1$ and

$$u_1(t) = \frac{1}{t}$$

in any interval not including the origin. Since the above quadratic in $\alpha$ has only one repeated root we do not obtain two different elements in $\mathcal{N}(L)$. Let us then find a second element of the form

$$u_2(t) = \phi(t) - \left(\frac{1}{t}\right)$$

If we substitute $u_2$ into the differential equation we obtain

$$t^2 \left[\phi'' \frac{1}{t} - 2\phi' \frac{1}{2} + \phi \frac{2}{3}\right] + 3t \left[\phi' \frac{1}{t} - \phi \frac{1}{2}\right] + \phi \frac{1}{t} = 0$$

so that

$$t\phi'' + \phi' = 0.$$ 

It follows that

$$\phi'(t) = \frac{K}{t}$$
and hence that

$$\phi(t) = K \ln t + c.$$ 

Hence a second element in $\mathcal{N}(L)$ is

$$u_2(t) = \frac{1}{t} \ln t.$$ 

By inspection $u_2$ is not a scalar multiple of $u_1$ so we have a basis for $\mathcal{N}(L)$. Finally we observe that the element $\phi$ generated in this way is always linearly independent from $u_1$. We simply look at the Wronskian:

$$\det \begin{pmatrix} u_1(t) & \phi(t)u_1(t) \\ u_1'(t) & \phi'(t)u_1(t) + \phi(t)u_1'(t) \end{pmatrix} = u_1^2(t)\phi'(t)$$

which is non-zero at some point because $\phi'$ is an exponential function.
Homework Module 13

1) Consider
\[ Lu \equiv u'' + 3tu' - 4\sin u = f(t). \]

i) Prove or disprove: \( L \) is linear on \( C^2[0,1] \).
ii) Determine \( f(t) \) such that \( u(t) = 5^t \) is a solution of \( Lu = f(t) \).

2) Consider
\[ Lu = u'' - 2u' + u = te^t. \]

i) Show that \( \{e^t, te^t\} \) is a basis of \( \mathcal{N}(L) \).
ii) Determine a particular integral of the form
\[ u_p(t) = (a_0 + a_1t + a_2t^2 + a_3t^3)e^t. \]

iii) Find a solution of \( Lu = te^t, \ u(1) = 1, \ u'(2) = 2 \).
iv) Write a first order system equivalent to \( Lu = te^t \),
\( \ u(1) = 1, \ u'(2) = 2 \) and give its solution.

3) Find a basis of \( \mathcal{N}(L) \) when
\[ Lu = u'' + 4u. \]

Solve
\[ Lu = 0 \]
subject to
\[ \int_0^\pi u(t)dt = 1 \]
or explain why no such solution exists.

4) Find a solution of the form \( u(t) = t^\alpha \) for
\[ Lu \equiv t^2u'' + 5tu' + 4u = 0 \]
then find a basis for \( \mathcal{N}(L) \).
We recall that if $A(t)$ is an $n \times n$ matrix with differentiable entries $a_{ij}(t)$ then

$$A'(t) = \lim_{\Delta t \to 0} \frac{A(t + \Delta t) - A(t)}{\Delta t} = (a'_{ij}(t))$$

and that if $B(t)$ is an $m \times n$ matrix with differentiable entries then

$$(B(t)A(t))' = B'(t)A(t) + B(t)A'(t).$$

Moreover, if $A(t)$ is differentiable and invertible in a neighborhood of a point $t_0$ then it follows from

$$(A(t)A^{-1}(t))' = A'(t)A^{-1}(t) + A(t)(A^{-1}(t))' = I' = 0$$

that

$$A^{-1}(t)' = -A^{-1}(t)A'(t)A^{-1}(t)$$

which is the matrix analog of $(1/f(t))' = f'(t)/f^2(t)$. We shall use these relationships in discussing further the solution of

$$Lu \equiv u' - A(t)u = 0, \quad \text{or equivalently, of } u'(t) = A(t)u.$$ 

As we have seen, the null space of $L$ has dimension $n$, and any $n$ functions which solve the equation and which are linearly independent at one point span the null space of $L$. For example, a convenient set of $n$ such functions may be found by integrating

$$u'_i = A(t)u, \quad u_i(0) = \hat{e}_i, \quad i = 1, \ldots, n.$$ 

It is clear that the matrix $U(t) = (u_1(t) \cdots u_n(t))$ satisfies

$$U'(t) = A(t)U(t),$$

and since $U(0) = I$ it is invertible in a neighborhood of $t = 0$. The next theorem shows that it is invertible for all $t$.

**Theorem:** Let $\{u_i(t)\}$ be a basis of $\mathcal{N}(L)$. The $U(t) = (u_1(t) \cdots u_n(t))$ is invertible everywhere.
**Proof.** We know that \( U(t)\vec{a} \equiv 0 \) for all \( t \) if and only if \( \vec{a} = 0 \) because of the linear independence of the functions \( \{u_i(t)\} \), but this does not rule out that \( U(t_0)\vec{\beta} = 0 \) with \( \vec{\beta} \neq 0 \) at some particular point \( t_0 \). So let us suppose that for some \( t_0 \) we have indeed \( U(t_0)\vec{\beta} = 0 \) for some \( \vec{\beta} \neq 0 \). But then the function \( v(t) = U(t)\vec{\beta} \) satisfies

\[
v' = A(t)v, \quad v(t_0) = 0.
\]

By uniqueness this implies that \( v(t) \equiv 0 \) for all \( t \) which contradicts the linear independence of the \( \{u_i(t)\} \). Hence there cannot be a point \( t_0 \) where \( U(t_0) \) is singular.

**Definition:** A non-singular matrix \( U(t) \) which satisfies

\[
U' = A(t)U
\]

is called a fundamental matrix for the differential equation

\[
Lu \equiv u' - A(t)u = 0.
\]

We observe that \( U \) is a fundamental matrix if and only if its columns span \( \mathcal{N}(L) \). Let us define the \( n \times n \) matrix

\[
\phi(t, s) = U(t)U^{-1}(s)
\]

then it follows that

\[
\frac{\partial}{\partial t} \phi(t, s) = U'(t)U^{-1}(s) = A(t)\phi(t, s)
\]

\[
\phi(s, s) = I,
\]

in other words, \( \phi(t, s) \) is the fundamental matrix which is the identity for \( t = s \). If we have computed a \( U(t) \), say by imposing initial conditions \( u_i(0) = \hat{e}_i \) on the columns of \( U \), then the solution to the initial value problem

\[
u' = A(t)u, \quad u(t_0) = u_0
\]

is simply

\[
u(t) = \phi(t, t_0)u_0.
\]
For completeness we observe that
\[
\frac{\partial}{\partial s} \phi(t, s) = U(t)U^{-1}(s)' = -U(t)[U^{-1}(s)U''(s)U^{-1}(s)]
\]
\[= -U(t)U^{-1}(s)A(s)U(s)U^{-1}(s) = -\phi(t, s)A(s).
\]

We shall now discuss the solution of
\[
Lu \equiv u' - A(t)u = F(t)
\]
which is usually written in the form

\[
u' = A(t)u + F(t)
\]
where \(F(t) = (f_1, \ldots, f_n)\) is a given source term. Such an equation is called a linear inhomogeneous equation.

We already learned that for the matrix problem \(Ax = b\) with \(b \in R(A)\) the solution is not unique when \(\dim\mathcal{N}(A) \geq 1\). For the differential equation we know that \(\dim\mathcal{N}(L) = n \neq 0\) and we shall see that the solution of the equation cannot be unique unless additional conditions, such as initial or boundary conditions, are imposed which do make the solution unique.

Let \(u_p(t)\) be ANY function which satisfies the equation

\[
u' = A(t)u + F(t)
\]
then the most general solution of this equation is

\[
u(t) = u_c(t) + u_p(t)
\]
where for any basis \(\{u_i(t)\}\) of \(\mathcal{N}(L)\) and arbitrary \(\alpha \in \mathbb{R}_n\)

\[u_c(t) = U(t)\alpha = (u_1 \ u_2 \ \cdots u_n)\alpha
\]
is known as the complementary solution. \(u_p(t)\) is called a particular integral of the equation. We note that if \(u_p\) and \(w_p\) are two particular integrals then \(u_p - w_p \in \mathcal{N}(L)\) so that two particular integrals differ by an element of \(\mathcal{N}(L)\).
The method of variation of parameters for the calculation of $u_p(t)$: Given the differential equation $u' = A(t)u + F(t)$, i.e., $Lu = u' - A(t)u = F(t)$, and a basis $\{u_i(t)\}$ of $\mathcal{N}(L)$ we can find a particular integral of the form

$$u_p(t) = U(t)v(t).$$

We compute that

$$u'_p = U'(t)v(t) + U(t)v'(t) = A(t)U(t)v + U(t)v' = A(t)u_p + U(t)v'.$$

Substitution into the differential equation yields

$$A(t)u_p + U(t)v' = A(t)u_p + F(t)$$

so that $v(t)$ is to be found by integrating

$$v' = U^{-1}(t)F(t).$$

Hence

$$v(t) = \int_{t_0}^{t} U^{-1}(s)F(s)ds$$

where $t_0$ is a convenient lower limit. Note that changing the lower limit changes $v(t)$ by a constant and hence $u_p(t)$ by an element in $\mathcal{N}(L)$ as expected. Thus we have found a particular integral of the form

$$u_p(t) = U(t)v(t) = \int_{t_0}^{t} U(t)U^{-1}(s)F(s)ds = \int_{t_0}^{t} \phi(t,s)F(s)ds.$$  

To illustrate this process consider the scalar equation

$$u' = a(t)u + F(t)$$

$$u(t_0) = u_0.$$  

The null space is one-dimensional and is spanned by the solution of

$$U' = a(t)U$$
\[ U(t_0) = 1 \]

which by separation of variables we find to be

\[ U(t) = e^{\int_{t_0}^{t} a(r) \, dr}. \]

(In all these calculations be super-careful not to confuse the independent variable \( t \) with any dummy variable of integration!) Because we have a scalar \( U(t) \) we find by inspection that

\[ U^{-1}(s) = e^{-\int_{t_0}^{s} a(r) \, dr} \quad \text{and} \quad \phi(t, s) = e^{\int_{s}^{t} a(r) \, dr}. \]

The general solution is

\[ u(t) = U(t) \alpha + \int_{t_0}^{t} \phi(t, s) F(s) \, ds. \]

The initial condition requires that

\[ \alpha = u_0. \]

Note that had we not chosen a particular integral which vanishes at \( t_0 \) then the computation of \( \alpha \) would have been more involved.

**A word of caution.** This scalar calculation has no matrix equivalent.

Given a matrix \( A(t) \) we know that

\[ A'(t) = (a'_{ij}(t)) \quad \text{and} \quad \int A(s) \, ds = \left( \int a_{ij}(s) \, ds \right) \]

so it is tempting to write a solution of

\[ U' = A(t)U, \quad U(t_0) = I \]

for an \( n \times n \) non-constant matrix \( A(t) \) as

\[ U(t) = \exp \left( \int_{t_0}^{t} A(s) \, ds \right) \]

where as before the matrix exponential is defined by the Taylor series

\[ e^{\int_{t_0}^{t} A(s) \, ds} = \sum_{n=0}^{\infty} \frac{\left( \int_{t_0}^{t} A(s) \, ds \right)^n}{n!}. \]
The series can be differentiated term by term, but unfortunately
\[
\frac{d}{dt} \left( \int_{t_0}^{t} A(s) ds \right) \left( \int_{t_0}^{t} A(s) ds \right) = A(t) \int_{t_0}^{t} A(s) ds + \int_{t_0}^{t} A(s) ds A(t)
\]
\[
\ne 2A(t) \int_{t_0}^{t} A(s) ds
\]
because in general \( A(t) \) and its integral do not commute. This implies that usually
\[
\frac{d}{dt} e^{\int_{t_0}^{t} A(s) ds} \ne A(t) e^{\int_{t_0}^{t} A(s) ds}
\]
i.e.,
\[
U'(t) \ne A(t)U(t)
\]
so that \( \exp \left( \int_{t_0}^{t} A(s) ds \right) \) is NOT a fundamental matrix. This raises the question of just when can we actually find a basis of \( \mathcal{N}(L) \) analytically which will be discussed in the next two modules.

Let us conclude this discussion by relating the computation of \( u_p \) for a first order system to the calculation of a solution \( u_p \) of the second order equation
\[
Lu \equiv u'' + a_1(t)u' + a_0(t)u = f(t).
\]
As we saw, if we set
\[
v_1 = u \\
v_2 = u'
\]
then the equivalent first order system is
\[
v' = \begin{pmatrix} 0 & 1 \\ -a_0(t) & -a_1(t) \end{pmatrix} v + \begin{pmatrix} 0 \\ f(t) \end{pmatrix}.
\]
Let \( \left\{ \begin{pmatrix} v_{11} \\ v_{12} \end{pmatrix}, \begin{pmatrix} v_{21} \\ v_{22} \end{pmatrix} \right\} \) be two linearly independent solutions of \( v' = A(t)v \) then
\[
u_1 = v_{11} \quad u_2 = v_{12} \\
u'_1 = v_{21} \quad u'_2 = v_{22}
\]
defines two linearly independent solutions \( \{u_1, u_2\} \) in \( \mathcal{N}(L) \). The particular integral \( v_p \) of the first order system is

\[
v_p(t) = V(t)z(t)
\]

where \( v(t) = (v_1(t) \ v_2(t)) \) and

\[
\begin{pmatrix}
v_1' \\
v_2'
\end{pmatrix} = V^{-1}(t) \begin{pmatrix}
0 \\
f(t)
\end{pmatrix} = \frac{1}{u_1' u_2' - u_1 u_2'} \begin{pmatrix}
-u_2 f(t) \\
u_1 f(t)
\end{pmatrix}.
\]

The first component of \( v_p(t) \) is a solution of \( Lu = f(t) \). Hence we have the particular integral

\[
u_p(t) = u_1(t)z_1(t) + u_2(t)z_2(t)
\]

were \( z_1 \) and \( z_2 \) are found by integration. We point out that this form of a particular integral for the second order equation \( Lu = f \) is usually derived by assuming a solution of the form

\[
u_p(t) = u_1(t)z_1(t) + u_2(t)z_2(t)
\]

where \( z_1 \) and \( z_2 \) are determined such that

\[
u_1 z_1' + u_2 z_2' = 0
\]

and

\[
Lu_p = f(t).
\]

When these two equations are solved for \( z_1' \) and \( z_2' \) exactly the same equations as those derived above via the first order system are found. Hence the method of variation of parameters exists for a first order system or a higher order scalar equation. Its dominant advantage is its applicability for all continuous functions \( F(t) \) whatever their form and structure. Its major drawback is the necessity to carry out integrations which may not be possible in closed form.

**Example:** Find a particular integral for

\[
Lu = u'' - u = \sin t
\]

**Answer:** It is easy to verify that two linearly independent solutions of \( Lu = 0 \) are

\[
u_1(t) = \cosh t, \quad u_2(t) = \sinh t
\]
It then follows that $u_1 u_2' - u_1' u_2 = 1$ and

$$z_1' = - \sinh t \sin t$$

$$z_2' = \cosh t \sin t.$$ Integrating from 0 to $t$ we obtain

$$z_1(t) = -(\cosh t \sin t - \cos t \sinh t)/2$$

$$z_2(t) = (1 - \cosh t \cos t + \sinh t \sin t)/2$$

so that

$$u_p(t) = -(\cosh^2 t \sin t - \cosh t \sinh t \cos t - \sinh t \cosh t \cos t - \sinh^2 t \sin t)/2$$

$$= (- \sin t + \sinh t)/2.$$ Since $\sinh t \in \mathcal{N}(L)$ we may take

$$u_p(t) = -(\sin t)/2.$$
Module 14 Homework

1) Let 
\[ A(t) = \begin{pmatrix} 1 & t \\ -t & 1 \end{pmatrix}. \]

i) Prove or disprove: 
\[ 2A'(t)A(t) = \frac{d}{dt} A^2(t) \]
for all \( t \).

ii) Prove or disprove 
\[ \frac{d}{dt} e^{A(t)} = A(t)e^{A(t)} \]
for all \( t \).

2) Solve 
\[ u' = (\cos t)u + \sin t \]
\[ u(3) = 4 \]

3) Let 
\[ B(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \]
and 
\[ C(t) = (3I + B)t. \]

i) Show that 
\[ e^{C(t)} = e^{3It}e^{Bt}. \]

ii) Compute explicitly 
\[ e^{3It} \]
and 
\[ e^{Bt} \]
by summing their Taylor series.

iii) Solve 
\[ u' = \begin{pmatrix} 3 & 1 \\ 0 & 3 \end{pmatrix} u + \begin{pmatrix} 1 \\ t \end{pmatrix} \]
\[ u(0) = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \]

iv) Find a scalar second order equation equivalent to the system of iii) and give its solution.

4) Consider 
\[ t^2u'' - tu' + u = e^{-t^2} \]
\[ u(1) = 1, \quad u'(1) = 0. \]

Find \( u_c(t) \) and \( u_p(t) \) and solve the problem. It is acceptable to have an integral for \( u_p \).
In general we can assert only that there exist $n$ linearly independent solutions of

$$Lu \equiv u' - A(t)u = 0,$$

where $A(t)$ is an $n \times n$ matrix, or of the scalar equation

$$Lu = \sum_{j=0}^{n} a_j(t)u^{(j)}(t) = 0$$

provided $a_n(t) \neq 0$ for any $t$. In practice these functions generally have to be found numerically. However, for equations with constant coefficients we may be able to do better.

The solution of a first order system with constant coefficients: Let us consider the first order system when the matrix $A(t)$ is independent of $t$, i.e., constant. We recall that for any scalar $t$

$$e^{At} = \sum_{n=0}^{\infty} \frac{(At)^n}{n!}.$$

We see that $Ae^{At} = e^{At}A$ and compute that

$$\frac{d}{dt} e^{At} = \sum_{n=0}^{\infty} \frac{nA^n t^{n-1}}{n!} = Ae^{At}.$$

Hence the matrix $U(t) = e^{At}$ solves

$$U'(t) = AU(t)$$

$$U(0) = I$$

which means that its columns form a basis of $\mathcal{N}(L)$ and that $U(t)$ is a fundamental matrix. Moreover, it follows from

$$\frac{d}{dt} [e^{At} e^{-At}] = Ae^{At}e^{-At} - e^{At} Ae^{-At} = 0$$

that $e^{At} e^{-At}$ is a constant matrix so that $e^{At} e^{-At} = e^{0}e^{0} = I$ or

$$U^{-1}(t) = e^{-At}.$$
Similarly one can show that \( e^{A(t-s)} \) and \( e^{At}e^{-As} \) both satisfy

\[
\begin{align*}
    u' &= Au \\
    u(s) &= I
\end{align*}
\]

so that \( e^{A(t-s)} \equiv e^{At}e^{-As} \) for all \( t \) and for any \( s \). Hence the initial value problem

\[
\begin{align*}
    u' &= Au \\
    u(t_0) &= u_0
\end{align*}
\]

has the explicit solution

\[
    u(t) = \phi(t, t_0)u_0,
\]

where

\[
    \phi(t, s) = e^{A(t-s)}.
\]

Similarly, the inhomogeneous initial value problem

\[
\begin{align*}
    u' &= Au + F(t) \\
    u(t_0) &= u_0
\end{align*}
\]

has the explicit solution

\[
    u(t) = \phi(t, t_0)u_0 + \int_{t_0}^{t} \phi(t, s)F(s)ds.
\]

Unfortunately, the definition of the matrix exponential in terms of its Taylor series does not usually lend itself to an actual computation of \( \phi(t, s) \). A sense of the difficulty can be obtained already from the scalar case

\[
    e^{-t} = \sum_{n=0}^{\infty} \frac{(-t)^n}{n!}.
\]

We know that this series representation is valid for all \( t \); however, in order to actually evaluate we are forced to truncate the series after \( N \) terms. For large positive \( t \) the exponential is essentially zero, but the truncated series is an \( N \)th order polynomial. Many terms of the
series are required for $n!$ to drag down the growth of $t^n$. It may be taken as an axiom of numerical mathematics that what is bad in one dimension is infinitely worse in higher dimensions. Since matrix multiplication is expensive and many powers of $A$ would have to be evaluated it generally is impossible to find a reasonable numerical expression for $\phi(t, s)$ for large $A$. However, suppose that $A$ has $n$ linearly independent eigenvectors, then it follows from

$$AX = X\Lambda$$

with

$$\Lambda = \begin{pmatrix} \lambda_1 & \lambda_2 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}$$

that

$$A = X\Lambda X^{-1}, \quad A^n = X\Lambda^n X^{-1}$$

and that

$$\phi(t, s) = e^{A(t-s)} = X \sum_{n=0}^{\infty} \frac{(t-s)^n \Lambda^n}{n!} X^{-1} = X \begin{pmatrix} e^{\lambda_1(t-s)} \\ \vdots \\ e^{\lambda_n(t-s)} \end{pmatrix} X^{-1}$$

which removes the infinite series and makes the solution computable.

**Example:** Solve

$$u' = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} u + \begin{pmatrix} t \\ \ln t \end{pmatrix}$$

$$u(1) = \begin{pmatrix} 3 \\ 4 \end{pmatrix}.$$

**Answer:** First of all we observe that the matrix $A$ is symmetric. Hence we know we can find two orthonormal eigenvectors $x_1$ and $x_2$. A simple calculation shows that the eigenvalues of $A$ are

$$\lambda_1 = 1$$

$$\lambda_2 = 3.$$

Since the eigenvalues are distinct the corresponding eigenvectors are automatically orthogonal. Orthonormal eigenvectors are

$$x_1 = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

108
If \( X = (x_1 \ x_2) \) then \( X^{-1} = X^T \) and hence

\[
\phi(t - s) = \frac{1}{2} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} e^{t-s} & 0 \\ 0 & e^{3(t-s)} \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}.
\]

The solution of the problem then is given by

\[
u(t) = \phi(t, 1) \begin{pmatrix} 3 \\ 4 \end{pmatrix} + \int_{1}^{t} \phi(t, s) \begin{pmatrix} s \\ \ln s \end{pmatrix} ds.
\]

A closer look at this formula suggests a simpler method to find a representation for \( u \) than the detour via the matrix exponential and its diagonalization. Given

\[u' = Au\]

we are looking for a solution of the form

\[u = xe^{\lambda t}\]

where \( x \) is a vector in \( \mathbb{R}_n \) and \( \lambda \) is a scalar. Substitution into the differential equation yields

\[(A - \lambda I)xe^{\lambda t} = 0.\]

Hence a nontrivial solution is found if \( \lambda \) is an eigenvalue of \( A \) and \( x \) is the corresponding eigenvector. Suppose we have found the \( n \) eigenvalues \( \{\lambda_i\} \) and corresponding eigenvectors \( \{x_i\} \). If \( U(t) = (x_1e^{\lambda_1} \ldots x_1e^{\lambda_n}) \) and if the eigenvectors \( \{x_i\} \) are linearly independent then \( U(0) \) is non-singular and hence \( U(t) \) is a fundamental matrix for the differential equation. A particular integral is then given by

\[u_p(t) = U(t)v(t)\]

where

\[v'(t) = U^{-1}(t)F(t),\]

and the general solution is

\[u(t) = U(t)\alpha + u_p(t)\]
where $\alpha$ is determined from initial and boundary conditions imposed on $u$.

Yet another view of exploiting the existence of $n$ linearly independent eigenvectors of $A$ is the following approach which is perhaps the simplest of all. We consider again

$$u' = Au + F(t).$$

If $A$ has $n$ linearly independent eigenvectors \{${x_i}$\} corresponding to the eigenvalues \{${\lambda_i}$\} then

$$X^{-1}AX = \Lambda$$

where $\Lambda$ is the diagonal matrix $\text{diag}\{\lambda_1, \ldots, \lambda_n\}$. From the differential equation we obtain

$$X^{-1}u'(t) = X^{-1}AXX^{-1}u + X^{-1}F(t).$$

If we set

$$y(t) = X^{-1}u(t) \quad \text{and} \quad G(t) = X^{-1}F(t)$$

then this equation is the same as

$$y'(t) = \Lambda y(t) + G(t).$$

Since $\Lambda$ is diagonal we have obtained $n$ uncoupled first order equations which generally are solvable. Once we have found $y(t)$ we know

$$u(t) = Xy(t).$$

If we cannot find $n$ linearly independent eigenvectors of $A$ then this simple approach breaks down, and neither can we diagonalize $e^{tA}$. However, $U(t) = e^{tA}$ remains a fundamental matrix of the equation. There are other means for avoiding a calculation of the matrix exponential from its Taylor series but we shall not pursue this topic in this course.
Module 15 Homework

1) Consider

\[ u' = \begin{pmatrix} 2 & 1 & -1 \\ 3 & 2 & -3 \\ 3 & 1 & -2 \end{pmatrix} u \]

\[ u(1) = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \]

Find a fundamental matrix and solve the problem.

2) Let

\[ A = \begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 2 \\ 0 & 2 & 1 \end{pmatrix} \]

Solve

\[ u' = Au + \begin{pmatrix} t \\ 1 \\ -t \end{pmatrix} \]

\[ u(1) = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \]

What is a fundamental matrix for this problem?

3) Suppose that two states \( A \) and \( B \) exchange populations. Suppose \( A \) loses people at a rate of 15% of its population to \( B \) and gains people at a rate of 10% of the population of \( B \). Suppose that initially \( A \) has 70% of the total population. Assume that the total population remains constant throughout:

i) Find the population of each state as a function of time.

ii) Find the distribution of the total population as \( t \to \infty \).

Assume that the total population grows while the exchange between the states continues at the rates given above:

iii) Find the population in each state if that of \( A \) grows at a rate of 6% of its population while that of \( B \) grows at a rate of 8%. When will the total population have doubled and what is its distribution between \( A \) and \( B \) at that time?

iv) Assume that both populations grow at the same rates as in iii) but with respect to the populations of the states 9 months earlier. Write the equations which model the population growth. What are appropriate initial conditions?
4) Suppose liquid is pumped at a constant rate of a \( \ell/\text{min} \) from tank 1 to tank 2, from tank 2 to tank 3 and from tank 3 to tank 1. Suppose the liquid volume in tank 1 is \( V \) liters. Suppose further that the volume of tank 2 is twice that of tank 1 and that of tank 3 exactly three times that of tank 1. Assume that initially tank 1 contains brine with a salt concentration of 3.5 \( g/\ell \) while the other two tanks contain fresh water. Determine the flow rate such that the salt concentration in tank 3 is exactly 20\% of that of tank 1 after 2 hours. (The problem is probably easiest to formulate for the amount \( u_i(t) \) of salt in tank \( i \).)
Topics: Scalar differential equations with constant coefficients

We shall consider equations of the form

\[ Lu \equiv a_n u^{(n)} + a_{n-1} u^{(n-1)} + \cdots + a_0 u = f(t) \]

where the coefficients are real constants with \( a_n \neq 0 \). In principle we already know everything there is to know about the solution of this equation because it is convertible to an \( n \)-dimensional first order system with a constant coefficient matrix. However, in general this conversion is too cumbersome and it is easier to tackle the equation directly.

We have shown that \( L \) has an \( n \) dimensional null space and that the general solution of the differential equation can be written in the form

\[ u(t) = \sum_{j=1}^{n} \alpha_j u_j(t) + u_p(t) \]

where \( \{u_i(t)\} \) is a basis of \( \mathcal{N}(L) \) and the particular integral \( u_p(t) \) is any solution of

\[ Lu = f(t). \]

Moreover, the results for first order systems suggest that we should look for basis elements for our \( n \)th order equation of the form

\[ u(t) = e^{\lambda t}. \]

If we substitute \( u(t) \) into \( Lu = 0 \) we find that \( \lambda \) must be chosen such that

\[ \sum_{j=0}^{n} a_j \lambda^j = 0. \]

This is an \( n \)th order polynomial with real coefficients and has \( n \) roots which may be real or complex, but if complex then they appear as conjugate pairs. We recall at this point that if \( \lambda \) is complex then for \( \lambda = a + ib \) we have by definition

\[ e^{\lambda} = e^{a+ib} = e^a (\cos b + i\sin b). \]
We now observe that if $\lambda$ and $\bar{\lambda}$ are complex conjugate roots then the corresponding functions \(\{e^{\lambda t}, e^{\bar{\lambda} t}\}\) are linearly independent because the Wronskian at $t = 0$ is
\[
\det \begin{pmatrix} 1 & 1 \\ \lambda & \bar{\lambda} \end{pmatrix} = -2i \Im \lambda \neq 0.
\]

We also observe that for complex scalars
\[
\text{span}\{e^{\lambda t}, e^{\bar{\lambda} t}\} = \text{span}\{\Re e^{\lambda t}, \Im e^{\lambda t}\}.
\]

It is generally better for applications in the real world to choose $\Re e^{\lambda t}$ and $\Im e^{\lambda t}$ as basis elements instead of $e^{\lambda t}$ and $e^{\bar{\lambda} t}$. For example,
\[
Lu \equiv u'' + u = 0
\]

has the two complex solutions $u_1 = e^{it}$ and $u_2 = e^{-it}$ but $\{\cos t, \sin t\}$ is a much more useful basis of $\mathcal{N}(L)$.

Suppose that the $n$ roots $\{\lambda_i\}$ of the polynomial are distinct, then the corresponding functions $\{e^{\lambda_i t}\}$ are linearly independent as may be seen from the Wronskian at time $t = 0$
\[
\det W(0) = \det \begin{pmatrix} 1 & 1 & 1 \\ \lambda_1 & \lambda_2 & \lambda_n \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \lambda_n^{n-1} \end{pmatrix}.
\]

$W(0)$ is the transpose of the famous Vandermonde matrix which for distinct $\lambda$ is known to be non-singular. Hence the Wronskian is non-zero which guarantees the linear independence of the $\{u_i\}$. If, however, the polynomial has repeated roots then we obtain only as many linearly independent functions as we have distinct roots. Additional functions then may be found with the reduction in order method outlined for a second order scalar equation in Module 13. But in principle the method works for any order scalar equation. To illustrate, consider
\[
Lu \equiv u''' - 3u' + 2u = 0.
\]

Assuming a solution of the form
\[
u = e^{\lambda t}
\]
we find by substituting into $Lu = 0$ that $\lambda$ must be root of

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

The three roots of this cubic are

$$\lambda_1 = -2 \quad \text{and} \quad \lambda_2 = \lambda_3 = 1.$$  

Hence we only have two linearly independent elements $u_1(t) = e^{-2t}$ and $u_2(t) = e^t$ of $\mathcal{N}(L)$. If we set

$$u_3(t) = \phi(t)e^t$$

and compute $Lu_3(t) = 0$ we find that

$$\phi''' + 3\phi'' = 0.$$  

This is a separable first order equation for $\phi''$. We find that

$$\phi''(t) = e^{-3t}$$

If we integrate twice we obtain

$$\phi(t) = \frac{1}{9} e^{-3t} + c_1 t + c_2$$

so that

$$u_3(t) = \frac{1}{9} e^{-2t} + c_1 te^t + c_2 e^t.$$  

It follows by inspection that

$$\text{span}\{u_1, u_2, u_3\} = \text{span}\{u_1, u_2, te^t\}$$

because the other two terms of $u_3$ belong already to $\text{span}\{u_1, u_2\}$. It is straightforward to verify that these functions are linearly independent. Most of our applications will involve second order equations so that the generation of a second linearly independent solution in the case of a double root will be somewhat simpler than the above example (see also the homework in this connection).
Thus we have the means, at least for constant coefficients, to find a basis for $\mathcal{N}(L)$, and we can turn our attention to

$$Lu = f(t).$$

We know that the general solution of this equation is of the form

$$u(t) = u_c(t) + u_p(t)$$

where the complementary solution $u_c$ is of the form

$$u_c(t) = \sum_{j=1}^{n} \alpha_j u_j(t), \quad u_i \in \mathcal{N}(L)$$

and where the particular integral $u_p$ is any solution of $Lu = f(t)$.

In principle, we already know how to find a $u_p(t)$ with the method of variation of parameters applied to the first order system. If we translate the earlier result for systems to a scalar equation we find that $u_p(t)$ should have the form

$$u_p(t) = \sum_{j=1}^{n} u_j(t)v_j(t)$$

where $\{u_i\}$ is a basis of $\mathcal{N}(L)$ and the $n$ scalar functions $\{v_i\}$ are to be computed such that $Lu_p(t) = f(t)$. To see this we note that the fundamental matrix $U(t)$ for the first order system associated with the elements $\{u_i\}$ is

$$U(t) = \begin{pmatrix}
    u_1 & u_2 & \cdots & u_n \\
    u'_1 & u'_2 & \cdots & u'_n \\
    \vdots & \vdots & \ddots & \vdots \\
    u^{(n-1)}_1 & u^{(n-1)}_2 & \cdots & u^{(n-1)}_n \\
\end{pmatrix}$$

and that the particular integral for the system is given by

$$\bar{u}_p(t) = U(t)v(t)$$

where

$$U(t)v' = F(t) = \begin{pmatrix} 0 \\ 0 \\ \cdots \\ 0 \\ f(t)/a_n \end{pmatrix}.$$
The first component of $U(t)v(t)$ is the particular integral for the scalar equation. It is precisely

$$u_p(t) = \sum_{j=1}^{n} u_j(t)v_j(t).$$

We mention that the usual derivation of the method of variation of parameters for scalar equation does not involve the corresponding first order system. Instead, one starts with the above representation, imposes the constraint that the first $n-1$ derivatives of $u_p$ are independent of $v'_i$ so that only $u_p^{(n)}(t)$ depends on $v'_i$. The same equations for $\{v'_i\}$ result which were found from the equivalent first order system. The advantage of the method of variation of parameters is its independence of the coefficients of $L$, although we do need a basis of $\mathcal{N}(L)$ which generally can be found only for a constant-coefficient operator. A second advantage is its independence of the form of $f$. For example,

$$Lu \equiv u'' + u = e^{t^2}$$

has the following particular integral

$$u_p(t) = \cos tv_1(t) + \sin tv_2(t)$$

where

$$\begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} v'_1 \\ v'_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \exp(t^2) \end{pmatrix}$$

so that

$$v'_1 = -\sin t\exp(t^2)$$

$$v'_2 = \cos t\exp(t^2).$$

However, the $\{v_i\}$ may not be computable. For our example we cannot go beyond

$$v_1(t) = -\int_{t_1}^{t} \sin s\exp(s^2) ds$$

and

$$v_2(t) = \int_{t_2}^{t} \cos s\exp(s^2) ds$$
because the integrations cannot be carried out. (We remark the different lower limits in these integrals were chosen simply to illustrate that any particular integral will work as long as it satisfies

\[ Lu_p = f(t). \]

For a special class of problems a somewhat simpler method exists for finding a particular integral. In essence, one guesses its form. This method is known as the method of undetermined coefficients and, where it applies, tends to be quicker than the method of variation of parameters.

The method of undetermined coefficients: The method applies to

1) \( Lu \equiv \sum_{j=1}^{n} a_j u^{(j)} = f(t) \)

where the coefficients are constants and where

2) \( f(t) = P_k(t)e^{\omega t} \) for real or complex \( \omega \).

Here \( P_k(t) \) is a polynomial of degree \( k \), i.e.

\[ P_k(t) = \sum_{j=0}^{k} b_j t^j. \]

In this case we assume a particular integral of the form

\[ u_p(t) = Q_{n+k}(t)e^{\omega t} \]

where

\[ Q_{n+k}(t) = \sum_{j=0}^{n+k} c_j t^j. \]

It is easy to see that \( LQ_{n+k}(t)e^{\omega t} \) is \( e^{\omega t} \) times a polynomial of degree \( \leq n + k \). The task then simply is to find \( \{c_i\} \) so that \( LQ_{n+k}(t)e^{\omega t} = f(t) \). Let us give a couple of examples.

**Example 1:**

\[ Lu \equiv u''' = 1. \]

Since \( P_k(t) \) is a polynomial of degree 0 then \( u_p \) is a polynomial of degree \( (4 + 0) \), i.e.

\[ u_p(t) = c_0 + c_1 t + c_2 t^2 + c_3 t^3 + c_4 t^4. \]
Since \( \{1, t, t^2, t^3\} \) are a basis of \( \mathcal{N}(L) \) we find that \( Lu_p(t) = 24c_4 \) so that we may simply set
\[
  u_p(t) = \frac{1}{24} t^4.
\]

**Example 2:**

\[
Lu \equiv u'' + u = \cos t.
\]

We know that
\[
\cos t = \frac{e^{it} + e^{-it}}{2} = \text{Re } e^{it}
\]
and one could solve
\[
Lu_{p1} = \frac{e^{it}}{2}, \quad Lu_{p2} = \frac{e^{-it}}{2},
\]
and then take \( u_p = u_{p1} + u_{p2} \). However, if we solve
\[
Lw_p = e^{it}
\]
then necessarily
\[
L(\text{Re } w_p(t)) = \cos t.
\]

According to the method of undetermined coefficients we set
\[
w_p(t) = (c_0 + c_1 t + c_2 t^2)e^{it}.
\]

We compute
\[
Lw_p(t) = (2c_2 + 2i(c_1 + 2c_2 t) - c_0 - c_1 t - c_2 t^2)e^{it} + (c_0 + c_1 t + c_2 t^2)e^{it}
= (2c_2 + 2ic_1 + 4ic_2 t)e^{it}.
\]

Since \( f(t) = e^{it} \) we find that \( c_2 = 0 \) and \( c_1 = \frac{1}{2i} \). Hence
\[
  u_p(t) = \text{Re } w_p(t) = \frac{t}{2} \sin t.
\]

Finally we observe that if one can see by inspection that \( f(t) = P_k(t)e^{\omega t} \not\in \mathcal{N}(L) \) then one may choose the simpler form
\[
  u_p(t) = Q_k(t)e^{\omega t}
\]
because the coefficients of powers \( > k \) will come out to be zero. For example

\[
Lu \equiv u'''' + u = e^{2t}
\]

has the particular integral

\[
u_p(t) = c_0 e^{2t}
\]

where

\[
c_0 = \frac{1}{9}.
\]

**An application:** Newton’s second law for a spring-dashpot system (like a shock absorber) is

\[
Lu = mu'' + cu' + ku = F(t)
\]

where \( u(t) \) is the displacement of the mass from its equilibrium position. Here \( m \) is the mass of the body on a restoring spring with spring constant \( k \). The motion is retarded in proportion to its velocity. \( m, c, \) and \( k \) are non-negative known constants and \( F(t) \) describes an external force applied to the mass.

If \( c = 0 \) and \( F(t) = 0 \) the equation reduces to

\[
Lu = mu'' + ku = 0.
\]

It describes simple harmonic motion which is usually pinned down by imposing an initial displacement and velocity

\[
u(0) = u_0
\]

\[
u'(t) = u_1.
\]

Without actually writing down the solution we can obtain some information on the motion from the equation itself. We observe that since \( Lu = 0 \) it also is true the \( u'(t)Lu = 0 \), which may be written as

\[
mu'(t)u''(t) + ku'(t)u(t) = \frac{d}{dt} \left[ \frac{1}{2} mu'^2 + \frac{1}{2} ku^2 \right] = 0.
\]

The first term in the brackets is the kinetic energy of the system, the second is the potential energy. The equation states that the energy remains constant so that

\[
\frac{1}{2} mu'^2 + \frac{1}{2} ku^2 = \frac{1}{2} mu_1^2 + \frac{1}{2} ku_0^2 = C > 0.
\]
This implies that the vector \( R(t) = (u(t), u'(t)) \) traces out an ellipse in the \( u - u' \) plane (known as the phase plane) which is traversed in the direction of \( R'(t) = (u', u'') = (u', -\frac{k}{m}u(t)) \), i.e., in the clockwise direction.

It is often possible to gain a great deal of information from a phase plane analysis even in the case where the equations of motion have no explicit solution (see the nonlinear pendulum equation in the problem section of this module).

If \( c \neq 0 \) then it follows from

\[
\frac{d}{dt} \left( \frac{1}{2} mu'^2 + \frac{1}{2} ku^2 \right) = -cu'^2 < 0
\]

that energy is lost at a rate proportional to the kinetic energy. Since we cannot integrate this equation we do not get any phase plane information. We therefore shall write out the solution of \( Lu = 0 \) explicitly for further analysis. Let us compute first a basis of \( \mathcal{N}(L) \). A trial solution of the form \( e^{\lambda t} \) leads to

\[
m\lambda^2 + c\lambda + k = 0
\]

so that

\[
\lambda_{1,2} = \frac{-c \pm \sqrt{c^2 - 4km}}{2m}.
\]

We now consider three cases:

**Case 1:** \( c^2 > 4km \) - We now have two linearly independent solutions

\[
u_1(t) = e^{\lambda_1 t} \quad \text{and} \quad u_2(t) = e^{\lambda_2 t} \quad \text{with} \quad \lambda_2 < \lambda_1 < 0.
\]

The general solution is

\[
u(t) = \alpha_1 u_1(t) + \alpha_2 u_2(t).
\]

If the initial conditions are

\[
u(0) = u_0 \neq 0
\]

\[
u'(0) = 0
\]

then

\[
u'(0) = \alpha_1 \lambda_1 + \alpha_2 \lambda_2 = 0.
\]
Moreover, there cannot be another time $T$ where $u'(T) = 0$ because in that case also

$$\alpha_1 \lambda_1 e^{\lambda_1 T} + \alpha_2 \lambda_2 e^{\lambda_2 T} = 0.$$  

These two equations have the only solution $\alpha_1 \lambda_1 = \alpha_2 \lambda_2 = 0$ which is inconsistent with $u_0 \neq 0$. Hence the motion has at most one time where $u' = 0$ and it changes direction. The system is said to be overdamped.

**Case 2:** $c^2 = 4km$. Now we obtain only one basis function

$$u_1(t) = e^{-\frac{c}{2m} t}$$

and are forced to generate a second solution of the form

$$u_2(t) = \phi(t) u_1(t).$$

Substitution of $u_2$ into $Lu = 0$ shows that

$$\phi''(t) = 0$$

and that \{ $e^{\lambda t}$, $te^{\lambda t}$ \} is a basis of $\mathcal{N}(L)$. It is straightforward to show that in this case the general solution also can have only one time where $u' = 0$. The system is now said to be critically damped.

**Case 3:** $c < 4km$. Now we have two complex conjugate roots and two real basis functions by taking $\text{Re} e^{\lambda_1 t}$ and $\text{Im} e^{\lambda_1 t}$. We find

$$u_1(t) = e^{-pt} \cos \omega t, \quad u_2(t) = e^{-pt} \sin \omega t$$

where $p = \frac{c}{2m}$ and $\omega = \frac{\sqrt{4km-c^2}}{2m}$. The general solution can be written as

$$u(t) = e^{-pt} [\alpha_1 \cos \omega t + \alpha_2 \sin \omega t]$$

or alternatively

$$u(t) = e^{-pt} [A \cos(\omega t - \phi)]$$

where $\phi$ denotes a phase shift. We now have an oscillatory solution with decaying amplitude. The system is said to underdamped.
We shall conclude this discussion by considering forced motion. We shall assume that there is no dissipation (i.e., \( c = 0 \)) and that the forcing function is of the form
\[
F(t) = F_0 \cos \delta t.
\]
The complete model is now
\[
Lu \equiv mu'' + ku = F_0 \cos \delta t
\]
\[
u(0) = u_0
\]
\[
u'(0) = u_1.
\]
It has the solution
\[
u(t) = \alpha_1 \cos \omega t + \alpha_2 \sin \omega t + \nu_p(t)
\]
where \( \omega^2 = \frac{k}{m} \). Let us assume that \( \delta \neq \omega \). Then it is seen by inspection that we can choose a particular integral of the form
\[
u_p(t) = A \cos \delta t.
\]
Substitution into the differential equation shows that \( A \) is given by
\[
A = \frac{F_0}{m(\omega^2 - \delta^2)}.
\]
We note that this construction breaks down when \( \delta = \omega \), i.e., when the source term belongs to \( \mathcal{N}(L) \). The solution satisfying the initial conditions is found to be
\[
u(t) = u_0 \cos \omega t + \frac{u_1}{\omega} \sin \omega t + \frac{F_0}{m(\omega^2 - \delta^2)} \left[ \cos \delta t - \cos \omega t \right].
\]
The physics of the problem suggests that absolutely nothing disastrous should happen to the solution of this problem, at least over a short time span, if we drive the system at the natural frequency \( \omega \), i.e., as \( \delta \to \omega \). We observe that in the limit the last term of \( \nu(t) \) leads to the indeterminate form 0/0 as \( \delta \to \omega \). It can be evaluated with l'Hospital's rule by differentiating with respect to \( \delta \) and then taking the limit. We find
\[
\lim_{\delta \to \omega} \nu(t) = u_0 \cos \omega t + \frac{u_1}{\omega} \sin \omega t + \frac{F_0 t \sin \omega t}{2m\omega}.
\]
The first two terms describe a constant amplitude oscillation, but the last term is oscillatory with a linearly increasing amplitude. This phenomenon is called resonance and occurs because we drive the system at the natural frequency \( \omega \).
Module 16 - Homework

1) i) Find the solution \( u(t, \epsilon) \) of the boundary value problem

\[
\begin{align*}
  u'' - 6u' + (9 - \epsilon^2)u &= te^{3t} \\
  u(0) &= 0, \quad u(1) = 1.
\end{align*}
\]

ii) Compute

\[
\lim_{\epsilon \to 0} u(t, \epsilon).
\]

iii) Compute a basis of \( \mathcal{N}(L) \) for \( Lu \equiv u'' - 6u' + 9u \).

iv) Compute a particular integral of

\[
Lu = te^{3t}
\]

with a) the method of variation of parameters
and b) the method of undetermined coefficients.

v) Solve \( Lu = te^{3t}, \ u(0) = 0, \ u(1) = 1 \) using the results of iii) and iv) and compare with the answer obtained in ii).

2) Find a polynomial which belongs to \( \mathcal{N}(L) \) where \( L \) is given by

\[
Lu \equiv t^2u'' - 6u.
\]

Then find a basis of \( \mathcal{N}(L) \) and solve

\[
Lu = \frac{1}{t} \quad u(1) = 1, \quad u(2) = 0.
\]

3) Solve

\[
Lu = u''' - 3u'' + 3u' - u = \cos t
\]

\[
\begin{align*}
  u(0) &= u'(0) = u''(0) = 0.
\end{align*}
\]

4) Consider the pendulum problem

\[
Lu \equiv \ell u''(t) + g \sin u(t) = 0
\]
\[ u(0) = u_0, \quad u'(0) = u_1 \]

where \( u(t) \) is the angle in radians between the pendulum and the vertical. Show with energy arguments similar to those used in the module that if

\[ \frac{1}{2} \ell u_1^2 - g \cos u_0 > g \]

then the pendulum can never reverse direction and will swing round and round.

5) Consider

\[ u'' + \frac{1}{1 + t^2} u' - u = e^{t \sin t} \]

\[ u(0) = u(1) = 0. \]

Show that the solution cannot be positive at any point on \((0,1)\).

(Hint: Show that the equation and the second derivative test for a relative maximum are inconsistent.)
When we discussed the matrix problem

\[ Lx = Ax = b \]

we learned that the existence of a solution depends on whether \( b \in R(A) \) and its uniqueness on whether \( \dim \mathcal{N}(A) = 0 \). If the null space was not trivial then the solution was of the form

\[ x = x_c + x_p \]

where \( x_c = X\alpha = (x_1 \cdots x_k)\alpha \) with \( \{x_i\} \) a basis of \( \mathcal{N}(A) \) and where \( x_p \) was any solution of \( Lx = b \). Both \( X \) and \( x_p \) were generally found with Gaussian elimination. Special constraints were then imposed to pin down \( x_c \); for example, one could find a minimum norm solution of \( Lx = b \). Consider now

\[ Lu = F(t) \]

where either

\[ Lu \equiv u' - A(t)u \]

or

\[ Lu = \sum_{j=0}^{n} a_j(t)u^{(j)}. \]

We have learned that the equation has a solution for all continuous functions \( F(t) \) and that it is of the form

\[ u(t) = U(t)\alpha + u_p(t) \]

where \( U(t) \) is either an \( n \times n \) or a \( 1 \times n \) matrix whose columns are a basis of \( \mathcal{N}(L) \). Moreover, for constant coefficient equations we have a mechanism for actually computing \( u(t) \) which may be thought of as the counterpart to Gaussian elimination for algebraic equations. Boundary, initial conditions or other constraints can now be imposed to determine \( \alpha \). As an illustration consider the following problem: Suppose the motion of a mechanical system is described by the following mathematical model

\[ u'' + u = 2e^{-t} \quad \text{for } t \in [0, 2\pi]. \]
What initial conditions \( u(0), \ u'(0) \) should we choose so that the solution \( u(t) \) is closest in the mean square sense to the motion \( w(t) = t(2\pi - t) \) for \( t \in [0, 2\pi] \)?

**Answer:** This constant coefficient equation has the general solution

\[
 u(t) = \alpha_1 \cos t + \alpha_2 \sin t + e^{-t}.
\]

The problem now is to determine \( \{\alpha_1, \alpha_2\} \) such that

\[
 E(\alpha_1, \alpha_2) = \|u(t) - w(t)\|_2^2 = \int_0^{2\pi} (u(t) - w(t)^2)dt
\]

is minimized. We can either use calculus and solve \( \partial E/\partial \alpha_1 = \partial E/\partial \alpha_2 = 0 \) or recall that the closest element in \( \text{span}\{\cos t, \sin t\} \) to \( g(t) = w(t) - e^{-t} \) is the orthogonal projection of \( g(t) \) with respect to the usual \( L_2[0, 2\pi] \) inner product. Since \( \{\cos t, \sin t\} \) are orthogonal in \( L_2[0, 2\pi] \) we find immediately

\[
 Pg(t) = \frac{\langle g(t), \cos t \rangle}{\langle \cos t, \cos t \rangle} \cos t + \frac{\langle g(t), \sin t \rangle}{\langle \sin t, \sin t \rangle} \sin t.
\]

We find (from the computer) that

\[
 Pg(t) = -4.159 \cos t - .1589 \sin t
\]

so that the best trajectory is

\[
 u(t) = Pg(t) + e^{-t}
\]

which yields initial conditions

\[
 u(0) = -3.159 \\
 u'(0) = -1.1589.
\]

Note this problem is equivalent to finding the solution of \( Ax = b \) which is closest to a given vector \( y \) when \( \dim \mathcal{N}(A) \geq 1 \).

Let us now broaden our discussion of linear differential operators by introducing the concept of the inverse for a differential operator \( L \), i.e., the analog to \( A^{-1} \) for a matrix whose null space is \( \{0\} \).
In order to talk about $L^{-1}$ we need to be a bit more precise about the domain of $L$. We observe that if

\[ Lu \equiv u' - A(t)u \]

is to be considered for $t \in [a,b]$ then $L$ is defined on the vector space

\[ \mathcal{V} = \left\{ \bar{u}(t) = \begin{pmatrix} u_1(t) \\ \vdots \\ u_n(t) \end{pmatrix} : u_i(t) \in C^1[a,b] \right\}. \]

(Note again that here we distinguish between the vector valued function $\bar{u}(t)$ and its components $\{u_i(t)\}$, each of which is a scalar function.) If for $t \in [a,b]$

\[ Lu \equiv \sum_{j=0}^{n} a_j(t)u^{(j)} \]

the $L$ is defined on $\mathcal{V} = C^n[a,b]$. In either case $L$ cannot have an inverse because we already know that $\dim \mathcal{N}(L) = n$ while $\dim \mathcal{N}(L) = 0$ is necessary for the inverse to exist. We will have to restrict $L$ to some subspace $\mathcal{M}$ such that for $u \in \mathcal{M}$

\[ Lu = 0 \Rightarrow u = 0. \]

This is not hard to do. For the first order system $Lu \equiv u' - A(t)u$ let

\[ \mathcal{M} = \{ \bar{u}(t) = (u_1(t), \ldots, u_n(t) : u_i(t_0) = 0, \ i = 1, \ldots, n \} \]

for some $t_0 \in [a,b]$. $\mathcal{M}$ is clearly closed under vector addition because $u(t_0) + v(t_0) = 0$ and $\alpha u(t_0) = 0$ for $u, v \in \mathcal{M}$ and $\alpha$ a scalar. Then

\[ Lu = 0 \quad \text{and} \quad u \in \mathcal{M} \]

implies that $u \equiv 0$ because the initial value problem

\[ u' = A(t)u, \quad u(t_0) = 0 \]

has only the zero solution. We can actually write the inverse of $L$ in this case. Let $U(t)$ be any fundamental matrix of $u' = A(t)u$ then

\[ Lu = F(t) \]
has the inverse
\[ u(t) = L^{-1}F = \int_{t_0}^{t} \phi(t, s)F(s)ds \]
where as before \( \phi(t, s) = U(t)U^{-1}(s) \).

The corresponding result for the \( n \)th order scalar equation is obtained if we choose
\[ M = \left\{ u : u(t_0) = u'(t_0) = \cdots = u^{(n-1)}(t_0) = 0 \right\} \]
because the initial value problem
\[ Lu = 0 \]
\[ u(t_0) = u'(t_0) = \cdots = u^{(n-1)}(t_0) = 0 \]
likewise only has the zero solution. The inverse of \( L \) can be expressed as
\[ u(t) = L^{-1}f(t) = U(t)v(t) \]
where \( U(t) = (u_1(t) \cdots u_n(t)) \) is a \( 1 \times n \) matrix whose columns are the basis functions of \( \mathcal{N}(L) \) in \( C^n[a,b] \), and where \( v \) is the \( n \)-dimensional vector valued function
\[ v(t) = \int_{t_0}^{t} W^{-1}(s) \begin{pmatrix} 0 \\ \vdots \\ f(s) \\ a_n(s) \end{pmatrix} ds \]
and where \( W \) is the matrix whose determinant is the Wronskian.

In general, the inverse of \( L \) on a subspace associated with initial value problems is not particularly useful. More interesting are subspaces associated with boundary value problems. Here we shall consider only the following special case;
\[ Lu \equiv a_2(t)u'' + a_1(t)u' + a_0(t)u = f(t) \]
\[ u(a) = 0, \quad u(b) = 0. \]
The associated subspace of \( C^2[a,b] \) is
\[ M = \{ u \in C^2[a,b] : u(a) = u(b) = 0 \}. \]
It is clear that $\mathcal{M}$ is a subspace since sums and multiples of functions vanishing at $a$ and $b$ likewise vanish at $a$ and $b$. In order to have an inverse for $L$ we need that $Lu = 0$ has only the zero solution in $\mathcal{M}$. Sometimes this has to be decided by calculation, sometimes it is guaranteed by the structure of the problem. For example, consider

$$Lu = u'' + u = 0$$
$$u(0) = 0, \quad u(b) = 0.$$ 

Any solution of this problem will be an element of $\mathcal{N}(L) \subset \mathcal{M}$. This problem we can solve in closed form. We know that $u(t)$ also is an element of $\mathcal{N}(L)$ in the bigger space $C^2[0, b]$ which has dimension 2 and a basis $\{\cos t, \sin t\}$. Hence

$$u(t) = c_1 \cos t + c_2 \sin t.$$ 

The boundary conditions require that $c_1 = 0$ and $c_2 \sin b = 0$. If $b \neq n\pi$ for an integer $n$ then $c_2 = 0$ and $\mathcal{N}(L) = \{0\}$. Hence $L$ has an inverse and

$$Lu = f(t)$$
has the solution $u(t) = (L^{-1}f)(t)$.

By the end of this module we shall also be able to compute this inverse. As an example of a class of problems where the structure of $L$ guarantees an inverse we consider

$$Lu = (a(t)u')' - c(t)u = f(t)$$
$$u(a) = u(b) = 0.$$ 

The problem may be thought of as $Lu = f(t)$ defined on $\mathcal{M} = \{u \in C[a, b] : u(a) = u(b) = 0\}$.

**Theorem:** If $a(t) > 0$ and $c(t) \geq 0$ on $[a, b]$ then $Lu = 0$ has only the trivial solution $u(t) = 0$ in $\mathcal{M}$ (i.e., the boundary value problem has only the zero solution).

**Proof:** Assume that $u$ is a solution of $Lu = 0$, $u(a) = u(b) = 0$, then

$$\int_a^b u(t)Lu(t)dt = \int_a^b (a(t)u')'u - c(t)u^2dt = 0.$$

130
Integration by parts shows that
\[ a(t)u'(t)u(t) \big|_a^b - \int_a^b (a(t)u'^2 + c(t)u^2) dt = 0. \]
Since \( a(t) > 0 \) and \( c(t) \geq 0 \) this implies that \( u' \equiv 0; \ u(a) = 0 \) then assures that \( u \equiv 0 \) for all \( t \in [a, b] \).

We note that this theorem does not apply to our first example \( Lu = u'' + u \) where the trivial null space was found by actually solving \( Lu = 0 \) subject to the boundary data. On the other hand, the theorem assures that
\[ Lu \equiv u'' - tu = 0 \quad \text{on} \ 0 < a < t < b \]
\[ u(a) = u(b) = 0 \]
only has the zero solution. A closed form solution in terms of elementary functions does not exist for this \( L \).

The computation of the inverse:
We consider
\[ Lu \equiv a_2(t)u'' + a_1(t)u' + a_0(t)u = f(t) \]
\[ u(a) = u(b) = 0. \]
As before we assume that \( a_2(t) \neq 0 \) on \([a, b]\).

Our goal is to find a mapping which takes a given \( f \) to the solution \( u \), i.e.,
\[ u(t) = (L^{-1})f(t). \]
The process is mechanical although we may not be able to carry out the requisite calculations in analytic form.

Let \( s \) be an arbitrary but fixed point in the interval \((a, b)\). Then we compute two functions \( G_2(t, s) \) and \( G_1(t, s) \) which, as functions of \( t \), satisfy:
\[ LG_2(t, s) = 0, \quad G_2(a, s) = 0, \quad a < t < s \]
\[ LG_1(t, s) = 0, \quad G_2(b, s) = 0, \quad s < t < b. \]
At \( t = s \) the two functions are patched together such that

\[
G_2(s, s) = G_1(s, s)
\]

and

\[
\frac{\partial}{\partial t} G_1(s, s) - \frac{\partial}{\partial t} G_2(s, s) = \frac{1}{a_2(s)}.
\]

**Definition:** The Green’s function for

\[
Lu = f
\]

\[u(a) = u(b) = 0\]

is the function

\[
G(t, s) = \begin{cases} 
G_1(t, s) & a < s < t \\
G_2(t, s) & t < s < b.
\end{cases}
\]

Thus, \( G(t, s) \) is defined on the square \([a, b] \times [a, b]\). It is continuous in \( t \) on the whole square, and for a given \( s \) it is a solution of \( Lu(t) = 0 \) on \((a, s)\) and \((s, b)\). However, at \( t = s \) the first derivative with respect to \( t \) will show a jump.

The significance of the Green’s function is that it essentially defines the inverse of \( L \) because the boundary value problem has the solution

\[
u(t) = \int_a^b G(t, s)f(s)ds.
\]

This is of course not obvious but can be verified as follows. We write

\[
u(t) = \int_a^b G(t, s)f(s)ds = \int_a^t G_1(t, s)f(s)ds + \int_t^b G_2(t, s)f(s)ds
\]

and differentiate.

\[
u'(t) = G_1(t, t)f(t) + \int_a^t \frac{\partial}{\partial t} G_1(t, s)f(s)ds - G_2(t, t)f(t) + \int_t^b \frac{\partial}{\partial t} G_2(t, s)f(s)ds.
\]

Since \( G_1(s, s) = G_2(s, s) \) for all \( s \) two terms on the right cancel and we have

\[
u'(t) = \int_a^t \frac{\partial}{\partial t} G_1(t, s)f(s)ds + \int_t^b \frac{\partial}{\partial t} G_2(t, s)f(s)ds.
\]
Differentiating again we obtain

\[
u''(t) = \frac{\partial}{\partial t} G_1(t, t)f(t) + \int_a^t \frac{\partial^2}{\partial t^2} G_1(t, s)f(s)ds - \frac{\partial}{\partial t} G_2(t, t)f(t) + \int_t^b \frac{\partial^2}{\partial t^2} G_2(t, s)f(s)ds.
\]

When we substitute into the differential equation we find

\[
Lu(t) = \int_a^t L G_1(t, s)f(s)ds + \int_t^b L G_2(t, s)f(s)ds
\]

\[
+ a^2(t) \left[ \frac{\partial}{\partial t} G_1(t, t) - \frac{\partial}{\partial t} G_2(t, t) \right] f(t) = f(t).
\]

Furthermore

\[
u(a) = \int_a^b G_2(a, s)f(s)ds = 0 \quad \text{and} \quad u(b) = \int_a^b G_1(b, s)f(s)ds = 0
\]

by construction. Thus \(u(t)\) satisfies the differential equation and the boundary conditions regardless of what \(f\) is, and we can write

\[u = L^{-1} f\]

where \(L^{-1}\) is a linear “integral” operator defined by

\[(L^{-1} f)(t) = \int_a^b G(t, s)f(s)ds\]

for every function \(f \in C^0[a, b]\). Following are some examples for solving boundary value problems with a Green’s function.

**Problem:** Find the Green’s function for the problem

\[Lu \equiv u'' = f(t)\]

\[u(0) = u(1) = 0.\]

According to our recipe we need to solve

\[LG_2(t, s) = 0 \quad \text{on } (0, s)\]

\[G_2(0, s) = 0\]
and

\[ LG_1(t, s) = 0 \quad \text{on } (s, 1) \]

\[ G_1(1, s) = 0. \]

We find

\[ G_2(t, s) = at \]

\[ G_1(t, s) = b(1 - t) \]

where \( a \) and \( b \) will be functions of \( s \) yet to be determined. The two so-called interface conditions

\[ G_1(s, s) = G_2(s, s) \]

\[ \frac{\partial}{\partial t} G_1(s, s) - \frac{\partial}{\partial t} G_2(s, s) = 1 \]

lead to

\[ b(1 - s) = as \]

\[ -b - a = 1. \]

We solve for \( a \) and \( b \) and find

\[ G(t, s) = \begin{cases} 
  s(t - 1) & 0 < s < t \\
  t(s - 1) & t < s < 1 .
\end{cases} \]

so that

\[ u(t) = \int_0^1 G(t, s)f(s)ds. \]

Suppose we wanted to solve this simple problem directly, without use of a Green’s function. We then proceed the usual way. The solution is

\[ u(t) = c_1 1 + c_2 t + u_p(t) \]

where \( \{1, t\} \) is a basis of \( \mathcal{N}(L) \subset C^2[0, 1] \) and where \( u_p(t) \) is any solution of \( u'' = f(t) \). We see that

\[ u_p(t) = \int_0^t \int_0^r f(s)ds \, dr \]
will do because \( u'(t) = \int_0^t f(s)ds \) and \( u''(t) = f(t) \). It remains to find \( c_1 \) and \( c_2 \) from the boundary conditions. Since \( u_p(0) = 0 \) we find from \( u(0) = 0 \) that \( c_1 = 0 \). From \( u(1) = 0 \) we find that

\[
c_2 = - \int_0^1 \int_0^t f(s)ds dr.
\]

Thus

\[
 u(t) = - \left[ \int_0^1 \int_0^t f(s)ds dr - \int_0^t \int_0^r f(s)ds dr \right].
\]

If we reverse the order of integration we find

\[
\int_0^t \int_0^r f(s)ds dr = \int_0^t \int_s^t f(s)ds dr = \int_0^t (t-s)f(s)ds.
\]

Hence

\[
 u(t) = -t \int_0^1 (1-s)f(s)ds - \int_0^t (t-s)f(s)ds
\]

\[
= \int_0^1 t(s-1)f(s)ds + \int_0^t s(t-1)f(s)ds
\]

\[
= \int_0^1 G(t,s)f(s)ds
\]

as we would expect.

**Problem:** Solve

\[
 Lu \equiv u'' + u = f(t)
\]

\[
 u(0) = A, \quad u(T) = B
\]

with a Green’s function. We begin by noting that the boundary data are not satisfied by the zero function so \( L \) is not defined on a vector subspace. Hence we need to modify the problem. Let \( v(t) = A \frac{T-t}{T} + B \frac{t}{T} \) and define \( w(t) = u(t) - v(t) \). Then \( w \) satisfies

\[
 Lw = w'' + u = Lu - Lv = f(t) - v(t) \equiv g(t)
\]

\[
 w(0) = w(b) = 0.
\]

Now we can find \( w \) with the Green’s function. We verify that

\[
 G_2(t, s) = a \sin t, \quad G_1(t, s) = b \sin(t-T)
\]
are solutions of $Lu = 0$ which satisfy the conditions $G_1(0, s) = G_2(T, s) = 0$. The interface conditions are
\[
\begin{pmatrix}
\sin s & -\sin(s - T) \\
-\cos s & \cos(s - T)
\end{pmatrix}
\begin{pmatrix}
a \\
b
\end{pmatrix}
= \begin{pmatrix}
0 \\
1
\end{pmatrix}.
\]
We solve for $a$ and $b$ and find
\[
w(t) = \int_0^T G(t, s)g(s)ds
\]
where
\[
G(t, s) = \begin{cases}
\frac{\sin s \sin(t - T)}{\sin T} & 0 < s < t \\
\frac{\sin t \sin(s - T)}{\sin T} & t < s < T.
\end{cases}
\]
We observe that the Green’s function does not exist for $T = n\pi$. But as we showed earlier, the problem
\[
Lu = u'' + u = 0
\]
has the nontrivial solution $u(t) = \sin t$ in $\mathcal{M} = \{u \in C^2[0, n\pi], u(0) = u(n\pi) = 0\}$. Hence we cannot expect the inverse to exist in this case.

What is not clear at this point is why one would bother to compute the Green’s function since if we can find $G_1$ and $G_2$ then we very likely can also write down the general solution of the problem and simply compute its constants from the boundary conditions. However, consider the following nonlinear boundary value problem:
\[
u'' = F(t, u)
\]
\[
u(0) = 0, \quad u(T) = 0.
\]
It is clear from our discussion that any solution of this problem is also a solution of of the integral equation
\[
u(t) = \int_0^T G(t, s)F(s, u(s))ds
\]
where $G$ is the known Green’s function for $Lu = u''$, $u(0) = u(T) = 0$. This function was computed above and allows us to replace the nonlinear differential equation by a nonlinear integral equation. Both for the analysis of the problem, i.e., answering questions on existence and uniqueness of a solution, and for the numerical solution of the problem an integral equation is often preferable to the corresponding differential equation.
Module 17 - Homework

1) i) Integrate explicitly
\[ \int_0^t \int_r^t e^{-s^2} \, ds \, dr. \]

ii) Let \( u(t) = \int_1^t \cos(t - s^2) \, ds \). Compute \( u'(t) \) and \( u''(t) \).

2) Compute the Green’s function for
\[ Lu = t^2 u'' - tu' + u = f(t) \]
\[ u(1) = 0, \quad u(2) = 0, \]

or show that it cannot be done.

(Observe that \( u(t) \equiv t \in \mathcal{N}(L) \).)

3) Solve with a Green’s function only the problem
\[ u''(t) = 0 \]
\[ u(1) = 5, \quad u(2) = 2. \]

4) Show that the null space of \( Lu \equiv u'' \) defined on
\[ \mathcal{M} = \{ u \in C^2[0, 1] : u(0) = u'(0), \ u(1) = 0 \} \]
consists of the zero function only. Then mimic the derivation of the Green’s function in the module but impose the boundary condition \( u(0) = u'(0) \) instead of \( u(0) = 0 \). What is the final Green’s function?
In analogy to the matrix eigenvalue problem $Ax = \lambda x$ we shall consider the eigenvalue problem

$$Lu = \mu u$$

where $\mu$ is a real or complex number and $u$ is a non-zero function (i.e., $u$ is not identically equal to zero). $\mu$ will be called an eigenvalue of $L$ and $u(t)$ is the corresponding eigenvector = eigenfunction. As in the case of inverses we shall find that the existence of eigenvalues depends crucially on the domain of $L$. For example, suppose

$$Lu \equiv u'$$

is defined on $V = \{u(t) = (u_1(t), \ldots, u_n(t)) : u_i(t) \in C^1[a, b]\}$ then for any constant $\mu \in (-\infty, \infty)$ the equation

$$Lu = \mu u$$

has the non-zero solution

$$u(t) = xe^{\mu t}$$

where $x$ is any non-zero vector in $\mathbb{R}_n$. Hence every number $\mu$ is an eigenvalue and $x$ is a corresponding eigenfunction. On the other hand, if the same problem is considered in the subspace

$$M = \{u \in V : u(t_0) = 0\}$$

then it follows from the uniqueness of the solution of

$$u' = \mu u$$

$$u(t_0) = 0$$

that only the zero solution is possible. Hence there is no eigenvalue $\mu \in (-\infty, \infty)$. A more complicated example is provided by the following

**Problem:** Find all nontrivial solutions of

$$Lu = u' - Au = \mu u$$
$u(0) = u(1)$

where $A$ is an $n \times n$ matrix. Note that the problem can be restated as finding all eigenvalues and eigenvectors of the linear operator

$$Lu \equiv u' - Au$$

on the subspace of vector valued functions which satisfy $u(0) = u(1)$.

**Answer:** Any such solution solves the linear equation

$$u' - (A + \mu I)u = 0.$$

Our discussion of constant coefficient systems says that $u(t)$ should be of the form

$$u(t) = xe^{\lambda t}$$

where $x$ is a constant vector. Substitution into the equation yields

$$[\lambda x - (A + \mu)x]e^{\lambda t} = 0.$$

This implies that

$$[A - (\lambda - \mu)I]x = 0.$$

A non-trivial solution can exist only if $(\lambda - \mu)$ is an eigenvalue of $A$ and $x$ is a corresponding eigenvector. Let $\{\rho_1, \ldots, \rho_n\}$ be the eigenvalues of $A$ with corresponding eigenvectors $\{x_1, \ldots, x_n\}$. Then for each $\rho_j$ we obtain a vector valued function $x_je^{(\rho_j + \mu)t}$ which solves

$$u' - Au = \mu u.$$

The additional constraint on the solution is that

$$u(0) = u(1).$$

This requires

$$e^{(\rho_j + \mu)} = 1$$
and hence that
\[(\rho_j + \mu) = 2m\pi in\]
where \(i^2 = -1\) and \(m\) is any integer. Thus the eigenvalues are
\[\mu_{j,m} = 2m\pi i - \rho_j\]
with corresponding eigenfunctions
\[u_{j,m} = x_j e^{2m\pi it}.\]
Incidentally, the condition \(u(0) = u(1)\) is not so strange. If the first order system is equivalent to an \(n\)th order scalar equation then this condition characterizes smooth periodic functions with period 1.

All further discussion of eigenvalue problems for differential operators will be restricted to second order equations of the form
\[Lu \equiv (a(t)u')' + q(t)u = \mu p(t)u\]
defined on the interval \([0, T]\) with various boundary conditions at \(t = 0\) and \(t = T\). We shall assume that all coefficients are real and continuous in \(t\). In addition we shall require that
\[a(t) > 0 \quad \text{on } [0, T]\]
\[p(t) > 0 \quad \text{on } [0, T]\] except possibly at isolated points
where \(p\) may be zero.

The function \(p(t)\) multiplying the eigenvalue may look unusual but it does arise naturally in many applications. Its presence can greatly complicate the calculation of eigenfunctions and eigenvalues but it has little influence on the general eigenvalue theory. As we have seen time and again, a differential operator is completely specified only when we are given the domain on which it is to be defined. The domain of \(L\) given above will be a subspace \(\mathcal{M}\) of \(C^2[0, T]\). In applications a number of different subspaces are common, but here we shall restrict ourselves to
\[\mathcal{M} = \{u \in C^2[0, T]: u(0) = u(T) = 0\}.\]
We now can obtain a number of results which follow from the specific form of the operator.

**Theorem:** The eigenvalues of
\[ Lu = (au')' + q(t)u = \mu p(t)u \]
\[ u(0) = u(T) = 0 \]
are real and the corresponding eigenfunctions may be chosen to be real.

**Proof.** As in the case of a matrix eigenvalue problem we don’t know a priori whether an eigenvalue will turn out to be complex and require a complex eigenvector. Suppose that \( \mu \) is an eigenvalue. Let \( u \) be the corresponding eigenvector. Then
\[ \int_0^T \bar{u}Lu \, dt = \mu \int_0^T p(t)u \bar{u} \, dt. \]
The integral on the right is real and positive. Integration by parts shows that
\[ \int_0^T \bar{u}Lu \, dt = \int_0^T [(au')'\bar{u} + q(t)u\bar{u}] \, dt = \int_0^T [(au'\bar{u})' - au'\bar{u}' + q(t)u\bar{u}] \, dt \]
\[ = (au'\bar{u}) \bigg|_0^T - \int_0^T [au'\bar{u}' + q(t)u\bar{u}] \, dt = \int_0^T [-au'\bar{u}' - q(t)u\bar{u}] \, dt \]
is real. Hence \( \mu \) is real and \( u \) may be taken to be real since for a complex function both the real and the imaginary part would have to satisfy the eigenvalue equation.

**Theorem:** Let \( \mu_m \) and \( \mu_n \) be distinct eigenvalues and \( \phi_m \) and \( \phi_n \) the corresponding eigenfunctions. Then
\[ \langle \phi_m, \phi_n \rangle = \int_0^T \phi_m(t) \phi_n(t) p(t) \, dt = 0 \]
i.e., the functions \( \phi_m \) and \( \phi_n \) are orthogonal with respect to the inner product
\[ \langle f, g \rangle = \int_0^T f(t) g(t) p(t) \, dt. \]

**Proof:** It follows from
\[ L\phi_m = \mu_m \phi_m p(t) \]
\[ L\phi_n = \mu_n \phi_n p(t) \]
that
\[ \int_0^T [\phi_n L\phi_m - \phi_m L\phi_n] \, dt = (\mu_m - \mu_n) \int_0^T \phi_m(t) \phi_n(t) p(t) \, dt. \]
But
\[ \int_0^T [\phi_n L\phi_m - \phi_m L\phi_n] dt = \int_0^T [(a\phi'_m\phi_n)' - (a\phi'_n\phi_m)'] dt = 0 \]
because of the boundary conditions. Hence \( \langle \phi_m, \phi_n \rangle = 0 \) for \( m \neq n \).

**Theorem:** If \( \mu \) is an eigenvalue with eigenfunction \( \phi \) then
\[ \mu < \frac{\int_0^T q(t)\phi^2(t) dt}{\int_0^T p(t)\phi^2(t) dt} \]

**Proof:** We observe that \( \phi' \neq 0 \) because if \( \phi' \equiv 0 \) then \( \phi \) would be constant and \( \phi(0) = 0 \) would imply that \( \phi \equiv 0 \) which is not possible for an eigenfunction. It then follows from
\[ \int_0^T (\mu p(t) - q(t))\phi^2(t) dt = \int_0^T (a\phi')'\phi dt = -\int_0^T a\phi'\phi' dt \]
that necessarily
\[ \mu < \frac{\int_0^T q(t)\phi^2(t) dt}{\int_0^T p(t)\phi^2(t) dt} \]
In particular, this implies that the eigenvalue is strictly negative whenever \( q \leq 0 \).

**An example and an application:** Consider the eigenvalue problem
\[ Lu(x) \equiv u''(x) = \mu u(x) \]
\[ u(0) = u(b) = 0 \]
where in view of the subsequent application we have chosen \( x \) as our independent variable. In this case \( a(x) = p(x) = 1 \) and \( q(x) = 0 \). The above theorems assure that eigenfunctions corresponding to distinct eigenvalues are orthogonal with respect to the usual \( L_2 \) inner product on \([0, b]\), and that the eigenvalues are strictly negative. It is convenient to set
\[ -\mu = \lambda^2, \quad \lambda \neq 0 \]
so that the problem to be solved is
\[ u''(x) + \lambda^2 u(x) = 0 \]
\[ u(0) = u(b) = 0. \]
This is a constant coefficient second order equation with solution

\[ u(x) = c_1 \cos \lambda x + c_2 \sin \lambda x. \]

The boundary conditions can be expressed as

\[
\begin{pmatrix}
1 & 0 \\
\cos b & \sin b
\end{pmatrix}
\begin{pmatrix}
c_1 \\
c_2
\end{pmatrix} =
\begin{pmatrix}
0 \\
0
\end{pmatrix}.
\]

A non-trivial solution is possible only if the coefficient matrix is singular. It is singular whenever its determinant is zero. Hence \( \lambda \) has to be found such that

\[ \sin \lambda b = 0. \]

This implies that

\[ \lambda_n = \frac{n\pi}{b}, \quad n = \pm 1, \pm 2, \ldots. \]

A vector \((c_1, c_2)\) which satisfies the singular system is seen to be \((0, 1)\). Hence the eigenvalues and eigenfunctions for this problem are

\[ \mu_n = -\lambda_n^2, \quad \lambda = \left( \frac{n\pi}{b} \right), \quad \phi_n(x) = \sin \lambda_n x, \quad n = 1, 2, \ldots. \]

A direct computation verifies that

\[ \langle \phi_n, \phi_m \rangle = 0 \quad \text{for} \ m \neq n. \]

**An application:** The so-called wave equation and boundary and initial conditions

\[ u_{xx}(x,t) - \frac{1}{c^2} u_{tt} = F(x,t) \]

\[ u(0,t) = u(b,t) = 0 \]

\[ u(x,0) = u_0(x) \]

\[ u_t(x,0) = u_1(x) \]

describe the displacement \( u(x,t) \) of a string at point \( x \) and time \( t \) from its equilibrium position. The string is held fixed at \( x = 0, x = b \) and has the initial displacement \( u_0(x) \) and initial velocity \( u_1(x) \). The source term \( F(x,t) \) is given.
The problem as stated is in general too difficult to solve analytically and has to be approximated. It becomes manageable if we think of $t$ as a parameter and project all data functions as functions of $x$ into the finite dimensional subspace

$$\mathcal{M} = \text{span}\{\phi_1, \ldots, \phi_N\}.$$ 

Here $\phi_n$ is the $n$th eigenfunction of $Lu \equiv u''(x)$, i.e.,

$$\phi_n(x) = \sin \lambda_n x, \quad \lambda_n = \frac{n\pi}{b} \quad \text{and} \quad \mu_n = -\lambda_n^2.$$ 

The projection is the orthogonal projection with respect to the inner product for which the $\{\phi_n\}$ are orthogonal, i.e., the inner product

$$\langle f, g \rangle = \int_0^b f(x)g(x)dx.$$ 

As we learned early on, the projection onto $\mathcal{M}$ with an orthogonal basis is

$$PF(x, t) = \sum_{j=1}^N \beta_j(t)\phi_j(x)$$ 

where

$$\beta_j(t) = \frac{\langle F(x, t), \phi_j(x) \rangle}{\langle \phi_j(x), \phi_j(x) \rangle}.$$ 

Remember, $t$ is considered a parameter and $\beta_j$ will have to depend on $t$ because $F$ in general will change with $t$. The other data functions have the simpler projections

$$Pu_0(x) = \sum_{j=1}^N c_j\phi_j(x) \quad \text{with} \quad c_j = \frac{\langle u_0, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle}$$ 

and

$$Pu_1(x) = \sum_{j=1}^N d_j\phi_j(x) \quad \text{with} \quad d_j = \frac{\langle u_1, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle}.$$ 

We now try to solve

$$u_{Nxx} - \frac{1}{c^2}u_{Ntt} = PF(x, t)$$

$$u_N(0, t) = u_N(b, t) = 0$$

144
\begin{align*}
  u_N(x, 0) &= Pu_0(x) \\
  u_{N,t}(x, 0) &= Pu_1(x).
\end{align*}

The reason that this problem is manageable is that it has a solution \( u_N(x, t) \) which for fixed \( t \) also belongs to \( \mathcal{M} \). Thus we look for a solution of the form

\[
  u_N(x, t) = \sum_{j=1}^{N} \alpha_j(t) \phi_j(x).
\]

We substitute into the wave equation for \( u_N \), use that \( \phi''_i(x) = -\lambda_i^2 \phi_i(x) \) and collect terms. We find that

\[
  \sum_{j=1}^{N} \left[ -\lambda_i^2 \alpha_j(t) - \frac{1}{c^2} \alpha''_j(t) - \beta_j(t) \right] \phi_j(x) = 0.
\]

But as orthogonal functions the \( \{\phi_i\} \) are linearly independent so that this equation can only hold if the coefficient of each \( \phi_j \) vanishes. Thus the function \( \alpha_i(t) \) must satisfy the linear constant coefficient second order differential equation

\[
  \alpha''_i(t) + c^2 \lambda_i^2 \alpha_i(t) = -c^2 \beta_i(t).
\]

Moreover, the initial conditions \( u_N = Pu_0 \) and \( u_{N,t} = Pu_1 \) at \( t = 0 \) require that

\[
  \alpha_i(0) = c_i
\]

and

\[
  \alpha'_i(0) = d_i.
\]

It follows that

\[
  \alpha_i(t) = \gamma_i \cos \lambda_i ct + \delta_i \sin \lambda_i ct + \alpha_{ip}(t)
\]

where the coefficients \( \{\gamma_i, \delta_i\} \) can only be determined after the particular integral \( \alpha_{ip}(t) \) is known.

Problems of this type arise routinely in connection with diffusion, electrostatics and wave motion. They often can be solved on finite dimensional subspaces defined by eigenfunctions of the spatial operator of the partial differential equation. For a concrete application of the above development for a driven string we refer to the problem section of this module.
Module 18 - Homework

1) Consider

\[ Lu \equiv u'' \]

defined on \( \mathcal{M} = \{ u \in C^2[0, 1] : u(0) = 0, u'(1) = 0 \} \). Modify the proofs of this module to prove that:

- i) The eigenvalues of \( L \) must be real.
- ii) The eigenvalues of \( L \) must be strictly negative.
- iii) The eigenfunctions corresponding to distinct eigenvalues must be orthogonal with respect to the usual \( L^2[0,1] \) inner product.
- iv) Which results change when the space \( \mathcal{M} \) is changed to

\[ \mathcal{M} = \{ u \in C^2[0, 1] : u'(0) = u'(1) = 0 \} \]?

2) Compute the eigenvalues and eigenvectors of

\[ Lu \equiv u'' = \mu u \]

\[ u(0) = u(1), \quad u'(0) = u'(1). \]

3) Compute the eigenvalues of

\[ Lu \equiv u'' = \mu u \]

\[ u(0) = u'(0), \quad u(1) = -u'(1). \]

(It suffices to set up the equation which has to be solved to find the eigenvalues. You will not be able to find them in closed form.)

4) Consider the problem

\[ Lu \equiv u'' - u' = \mu u \]

\[ u(0) = u(1) = 0. \]

- i) Compute the eigenvalues and eigenvectors of \( L \).

This operator \( L \) is not of the form required by the theorems of this module. But it can be brought into the correct form as outlined in the following instructions:
ii) Find a function $\phi(t)$ such that

$$\left[e^{\phi(t)}u'(t)\right]' = [u'' - u']e^{\phi(t)} = \mu ue^{\phi(t)}.$$ 

iii) What orthogonality is predicted by the theorem of the module for eigenfunctions corresponding to distinct eigenvalues?

iv) Verify by direct computation that the eigenfunctions are orthogonal in the correct inner product.

5) Consider the problem of a vibrating string which is oscillated at $x = 0$. The model is:

$$u_{xx} - u_{tt} = 0$$

$$u(0, t) = F_0 \cos \delta t$$

$$u(1, t) = 0$$

$$u(x, 0) = F_0 (1 - x)$$

$$u_t(x, 0) = 0.$$ 

Find an approximate solution. When does resonance occur?

**Hint:** Reformulate the problem for $w(x, t) = u(x, t) - F_0 (1 - x) \cos \delta t$ and apply the ideas outlined in the module to find an approximation $w_N(x, t)$. 

147
Topics: The number system and the complex numbers

The development of the number system:

1. Positive integers (natural in counting)
   We can add and multiply.

2. All integers
   We can add, multiply and subtract.

3. Ordered pairs of integers \((a, b)\), \(b \neq 0\)
   Rules for addition and multiplication:

   \[ (a, b) + (c, d) = (ad + bc, bd) \]

   \[ (a, b)(c, d) = (ac, bd) \]

   These ordered pairs make up the rational numbers and are usually written in the form

   \[ (a, b) = \frac{a}{b} \]

Now we can add, subtract, multiply and divide.

We can enumerate the rational numbers, that is, we can put all rational numbers into a 1-1 correspondence with the positive integers. For example, if we write the array

\[
\begin{array}{cccccccc}
0/1 & 0/2 & 0/3 & 0/4 & 0/5 & 0/6 & 0/7 \\
1/1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & \cdots \\
2/1 & 2/2 & 2/3 & 2/4 & 2/5 & \cdots & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
m/1 & m/2 & \cdots & \cdots & \cdots & \cdots & \cdots \\
\end{array}
\]

then every rational number \(p/q\) appears in this array, and appears only once (where we distinguish between \(1/1, 2/2\) etc.). If we identify the rational number \(p/q\) with the number \(n\) of entries which lie in the triangle with corners \((0/1), (0/p + q - 1), (p + q - 1/1)\) plus the number of entries on the line from \(0/(p + q)\) to \(p/q\), then we have a unique function which maps \(p/q\) to \(n\) and any \(n\) to a unique \(p/q\). Hence there are as many rational numbers as
there are positive integers. We say that the rational numbers are countable. Similarly, the rational numbers in any finite interval are countable since there are infinitely many but not more than all rational numbers.

Integers = \((a, 1)\) are embedded in rational numbers.

But, not all numbers are rational numbers.

There is no \(p/q\) such that \((p/q)^2 = 2\), for if there were a \(p/q\) which we may assume to have no common factors, then \(p^2 = 2q^2\). \(2q^2\) is even which implies that \(p\) is even, so that \(p = 2p'\). But then \(2p'^2 = q^2\) which makes \(q\) even contrary to the assumption that \(p\) and \(q\) have no common factor. Hence \(\sqrt{2}\) is not rational.

4. Real numbers = limits of all sequences of rational numbers
   = rational and irrational numbers

   Between any two rational numbers there is an irrational number. For example, if \(a\) and \(b\) are rational then \(a(1 - 1/\sqrt{2}) + b/\sqrt{2}\) is irrational and between \(a\) and \(b\).

   Between any two irrational numbers there is a rational number because we can approximate any irrational number by a rational number from above or below.

**Theorem:** All the rational numbers on the interval \([0, 1]\) can be covered with open intervals such that the sum of the length of these intervals is arbitrary small.

**Proof:** We know we can enumerate the rationals, i.e., each rational can be labeled with an integer \(n\). Let \(I_n\) be an interval of length \(\ell(I_n) = \epsilon/2^n\) centered at the \(n\)th rational. Then all rationals will be covered by open intervals and

\[
\sum_{n=1}^{\infty} \ell(I_n) = \epsilon \sum_{n=1}^{\infty} 2^{-n} = \epsilon.
\]

Thus the rationals take up no space in the interval \([0, 1]\) so the irrationals must fill up the interval. Yet between two of one kind there is one of the other kind. This should convince you that the reals are very complicated.

5. Complex numbers:
Ordered pairs of real numbers and rules:

\[(a, b) + (c, d) = (a + c, b + d)\]

\[(a, b)(c, d) = (ac - bd, ad + bc)\]

\[\left(ac - bd, ad + bc\right) = (1, 0) \rightarrow (c, d) = \left(a/(a^2 + b^2), -b/(a^2 + b^2)\right)\]

real numbers: \((a, 0)\) are embedded in complex numbers.

Common notation

\[(a, b) = a + ib \quad \text{where } i^2 = -1.\]

Geometric representation:

\[(a, b) = \text{vector in } \mathbb{R}_2.\]

Addition: This is the same as vector addition in \(\mathbb{R}_2\) (component-wise addition).

Multiplication: No analog for vectors in \(\mathbb{R}_2\).

Two elementary functions of a complex variable:

Complex variable: \(z = x + iy\)

i) \(f(z) = \bar{z} = x - iy\)

Properties: \(\overline{z_1 + z_2} = \overline{z_1} + \overline{z_2}\)

\(\overline{z_1z_2} = \overline{z_1} \overline{z_2}\)

ii) \(f(z) = |z| = \sqrt{x^2 + y^2}\) (Euclidean length of the vector \((x, y) \in \mathbb{E}_2\)).

Properties: \(|z_1z_2| = |z_1||z_2|\)

\[z\bar{z} = |z|^2.\]

Triangle inequality:

\[|z_1 + z_2| \leq |z_1| + |z_2|.\]

This result follows immediately from the triangle inequality for vectors in \(\mathbb{E}_2\).

Reverse triangle inequality:

\[||z_1| - |z_2|| \leq |z_1 + z_2|.\]

This result follows from

\[|z_1| = |z_1 - z_2 + z_2| \leq |z_1 - z_2| + |z_2|.\]
**Definition:** $e^{i\theta} = \cos \theta + i \sin \theta$ for any real $\theta$.

From the trigonometric addition formulas follows: $e^{i(\theta + \phi)} = e^{i\theta} e^{i\phi}$.

The polar form of a complex number:

$$z = |z| e^{i\theta}.$$ 

**Definition:** $|z|$ is the magnitude (modulus) of $z$

$\theta$ is the argument of $z = \arg(z)$

$\arg(z)$ is not uniquely defined because adding any multiple of $2\pi$ to $\theta$ does not change the point in the complex plane. We say: $\arg(z)$ is “a multiple valued function.”

We can choose a “branch” of this multiple valued function by restricting the argument to a specific interval of length $2\pi$. This makes the function single valued. The most common branch is the principal value branch defined next.

**Definition:** The principal value of the argument of $z$, denoted by $\text{Arg } z$, is the value of the argument restricted to $(-\pi, \pi]$.

For example, $\text{Arg } -1 = \pi$, $\lim_{n \to \infty} \text{Arg}(-1 - i/n) = -\pi$. $\text{Arg } 0$ is not defined. Away from the negative $x$-axis $\text{Arg } z$ is a nice continuous function, but it jumps from $\pi$ to $-\pi$ as we cross from the half plane $y > 0$ to $y < 0$ across the negative real axis.

Geometric interpretation of complex multiplication:

If $z_1 = |z_1| e^{i\theta_1}$ and $z_2 = |z_2| e^{i\theta_2}$ then

$$z_1 z_2 = |z_1| |z_2| e^{i(\theta_1 + \theta_2)}.$$ 

In words: Magnitudes multiply, arguments add.

The fact that arguments add on multiplication allows us to take roots of complex numbers.

**Definition:** $w = \sqrt[n]{z} \equiv (z)^{1/n}$ for a given complex number $z$ means that we are looking for all complex numbers $w$ such that

$$w^n = z.$$ 

Calculation of roots of $z$:

Given a complex number $z$ we express it in polar form

$$z = |z| e^{i(\theta + 2\pi k)}$$
where \( k \) is any integer and where \( \theta \) is a conveniently chosen angle for the ray through \( z \) (maybe an angle \( \in (-\pi, \pi] \) i.e., \( \theta = \text{Arg} \, z \)). Then

\[
    w_k = \sqrt[n]{|z|} e^{i \left( \frac{\theta + 2\pi k}{n} \right)}, \quad k = 0, 1, 2, \ldots, n-1
\]

yields \( n \) complex numbers \( \{w_k\} \) with the property that \((w_k)^n = z\). It is readily verified that for any other integer \( k \) we obtain no new points in the complex plane. For example, if \( k = -1 \) then

\[
    e^{i \left( \frac{\theta - 2\pi}{n} \right)} = e^{i \left( \frac{\theta - 2\pi + 2\pi}{n} \right)} = e^{i \left( \frac{\theta + 2(n-1)\pi}{n} \right)}
\]

so that

\[
    w_{-1} = w_{n-1}.
\]

**Example:** \(-1 = e^{i(\pi + 2k\pi)}\) so that \( w = \sqrt{-1} \) has the solutions

\[
    w_0 = e^{i \frac{\pi}{2}} = i
\]

\[
    w_1 = e^{i \frac{3\pi}{2}} = -i
\]

**Note:** Complex numbers are not ordered.

\[
    z_1 < z_2 \quad \text{makes no sense.}
\]

Sets in the complex plane:

\[
    S = \{z : |z - a| = r\} \quad \text{is a circle about} \ a \ \text{of radius} \ r;
\]

\[
    z = a + re^{i\theta} \quad \text{defines the points on} \ S.
\]

\[
    S = \{z : |z - a| < R\} \quad \text{is the open disk around} \ a \ \text{of radius} \ R
\]

where \( z = a + re^{i\theta} \) with \( r < R \) defines the points in \( S \).
Module 19 - Homework

1) Use the ordered pair notation and the rules for computing with ordered pairs to show that for every non-zero \((a, b)\) there is a unique \((c, d)\) such that \((a, b)(c, d) = (1, 0)\).

2) Put into polar form

\[
z = \frac{-1 + \sqrt{3}i}{2 + 2i}
\]

3) Compute \((-1)^{1/5}\).

4) Compute the following limits or show that they do not exist:

\[
\begin{align*}
\lim_{n \to \infty} \text{Arg}\left(7 + \frac{(-1)^n}{n}i\right) \\
\lim_{n \to \infty} \text{Arg}\left(-7 + \frac{(-1)^n}{n}i\right) \\
\lim_{n \to \infty} \text{Arg}\left(-7 - \frac{1}{n}i\right) \\
\lim_{n \to \infty} \text{Arg}\left(-7 + \frac{(-1)^n}{n!}i\right) \\
\lim_{n \to \infty} \text{Arg}\left(7 + \frac{(-1)^n}{n}\right)
\end{align*}
\]

5) Find the value of \(\text{arg}(-1 - i)\) which belongs to the interval \((121, 121 + 2\pi]\).

6) Plot \(S = \{z : |z - 1| = \text{Re} z + 1\}\)

7) Given \(z\) with \(|z| \neq R\) and \(R > 0\) find \(z'\) and \(q\) with \(z' \neq z\) such that

\[
|z - w| = q|z' - w| \quad \text{for all } w \text{ such that } |w| = R.
\]

Where is \(z'\) if \(|z| < R\), where is \(z'\) if \(|z| > R\)?

8) i) Let \(D\) be a complex number. Let \(d_0\) and \(d_1\) be the two square roots of \(D\). Show that \(d_0 = -d_1\).

ii) Show that the quadratic formula for \(az^2 + bz + c = 0\) holds for complex \(a, b, c\).
We already have discussed a few functions of a complex variable $z = x + iy$, such as

- $f(z) = z$
- $f(z) = \bar{z}$
- $f(z) = |z|$
- $f(z) = \text{Arg } z$.

We shall now take a broader look at complex valued functions of a complex variable. The common notation is

$$w = f(z) = u(x, y) + iv(x, y).$$

We think of $f$ as taking a point $z = x + iy$ in the $z$-plane to a point $w = u + iv$ in the $w$-plane. Later on we shall worry about the image of sets in the $z$-plane under the transformation, as well as about the inverse of the mapping which takes sets in the $w$-plane to the $z$-plane.

Properties of $f$ which do not link the real and imaginary parts are readily examined with the calculus for real valued functions. For example, if $\{z_n\}$ is a sequence of points in the $z$-plane such that

$$\lim_{n \to \infty} f(z_n) = A, \quad \lim_{n \to \infty} g(z_n) = B$$

then

$$\lim_{n \to \infty} [f(z_n)g(z_n)] = AB$$

and

$$\lim_{n \to \infty} \left[\frac{f(z_n)}{g(z_n)}\right] = \frac{A}{B}, \quad \text{provided } B \neq 0.$$  

Something essentially new is introduced when we talk about differentiation of $f$ because as we shall see there is now a link between $u$ and $v$.

**Definition:** The derivative of $f$ at $z$ is

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}$$

provided this limit exists and is independent of how $\Delta z$ is chosen.
While the definition of the derivative is familiar from the calculus for real valued functions the independence of the limiting value of how \( \Delta z \to 0 \) means that we can try to find \( f'(z) \) for choices like

\[
\begin{align*}
\Delta z &= \Delta x \\
\Delta z &= i\Delta y \\
\Delta z &= \Delta t^\alpha + i\Delta t^\beta & \text{where} \Delta t \to 0 \text{ and } \alpha, \beta > 0.
\end{align*}
\]

**Applications:**

1) Suppose \( f(z) = z \) then

\[
\frac{f'(z) = \lim_{\Delta z \to 0} \frac{z + \Delta z - z}{\Delta z} = 1.}
\]

2) Suppose \( f(z) = \bar{z} \) then for

\[
\Delta z = \Delta x : \frac{\bar{z} + \Delta x - \bar{z}}{\Delta x} = 1; \quad z = i\Delta y : \frac{\bar{z} + iy - \bar{z}}{i\Delta y} = -1.
\]

There cannot be the same limiting value for all \( \Delta z \), so \( f(z) = \bar{z} \) is not differentiable anywhere.

3) Suppose \( f(z) = \text{Arg}(z) \), then for \( z = re^{i\theta} \) and \( \Delta z = re^{i(\theta + \Delta \theta)} - re^{i\theta} \)

\[
\lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} = \lim_{\Delta \theta \to 0} \frac{\theta + \Delta \theta - \theta}{re^{i\theta}(e^{i\Delta \theta} - 1)} = \lim_{\Delta \theta \to 0} \frac{\Delta \theta}{re^{i\theta}[\cos \Delta \theta + i \sin \Delta \theta - 1]} = \frac{1}{iz}
\]

as may be seen, e.g., from l’Hospital’s rule by differentiating with respect to \( \Delta \theta \). But if \( \Delta z = (r + \Delta r)e^{i\theta} - re^{i\theta} \) then \( f(z + \Delta z) - f(z) = 0 \), so the limit depends on \( \Delta z \) and \( f \) is not differentiable.

If \( f \) and \( g \) are differentiable then a manipulation of limits as in the real case will yield the familiar formulas

\[
\frac{d}{dz} \frac{f(z) + g(z)}{g(z)} = f'(z) + g'(z), \quad (f(z)g(z))' = f'(z)g(z) + f(z)g'(z)
\]

\[
\frac{f'(z)g(z) - f(z)g'(z)}{g^2(z)} = \left( \frac{f(z)}{g(z)} \right)' = f'(z)g(z) - f(z)g'(z)
\]

\[
(f(g(z)))' = f'(g(z))g'(z).
\]

**Application:**

\[
\frac{d}{dz} z^n = nz^{n-1}.
\]
We use induction:

\[(zz)' = z'z + zz' = 2z,\]

hence the formula is true for \(n = 2\); suppose the formula is true for arbitrary \(k\) then

\[(z^{k+1})' = (z^k z)' = k z^{k-1} z + z^k z' = (k + 1)z^k.\]

Hence the formula is true for \(k + 1\) and thus for all integers \(n\). It follows that polynomials \(P_m(z), Q_n(z)\) and rational functions \(P_m(z)/Q_n(z)\) are differentiable.

**The Cauchy-Riemann equations:**

In general it is not feasible to take the limit of the difference quotient to determine whether \(f\) is differentiable and to find its derivative. A more mechanical method is based on the following observation. Suppose that \(f\) is differentiable then necessarily

\[
f'(z) = \lim_{\Delta x \to 0} \frac{u(x + \Delta x, y) + iv(x + \Delta x, y) - (u(x, y) + iv(x, y))}{\Delta x} = \frac{\partial u(x, y)}{\partial x} + i \frac{\partial v(x, y)}{\partial x}
\]

and

\[
f'(z) = \lim_{i \Delta y \to 0} \frac{u(x, y + i \Delta y) + iv(x, y + i \Delta y) - (u(x, y) + iv(x, y))}{i \Delta y} = -i \frac{\partial u(x, y)}{\partial y} + \frac{\partial v(x, y)}{\partial y}.
\]

Since two complex numbers are the same only if the real and imaginary parts are the same we see that if \(f\) is differentiable at the point \(z\) then necessarily

\[
\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y},
\]

\[
\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.
\]

These two equations are the Cauchy-Riemann equations and must be satisfied if \(f\) is differentiable at a point. Without proof we state the condition under which the Cauchy-Riemann equations guarantee that \(f\) is differentiable.

**Theorem:** Suppose the partial derivatives of \(u\) and \(v\) exist and are continuous in a neighborhood of the point \(z\). If they satisfy the Cauchy-Riemann equations at \(z\) then \(f\) is differentiable at \(z\).

**Examples:** \(f(z) = x - iy\) was already shown not to be differentiable anywhere. We find that \(u_x = 1, u_y = 0, v_x = 0, v_y = -1\). Hence the partial derivatives are continuous everywhere
but satisfy the Cauchy-Riemann equation $u_x = v_y$ nowhere. $f(z) = |z| = \sqrt{x^2 + y^2}$ has continuous non-zero derivatives $u_x$ and $u_y$ for all $z \neq 0$, but $v_x = v_y = 0$ so the Cauchy-Riemann equations hold nowhere.

The Cauchy-Riemann equation can be transformed into polar coordinates since

$$x = r \cos \theta \quad \text{and} \quad y = r \sin \theta$$

where $r = |z|$ and $\theta = \arg(z)$. However, instead of a formal transformation let us derive the Cauchy-Riemann equations in polar coordinates directly from the difference quotient. If we express $z$ in polar coordinates then $z = re^{i\theta}$ and

$$f(z) = u(r, \theta) + iv(r, \theta).$$

If $\Delta z = (r + \Delta r)e^{i\theta} - re^{i\theta}$ then $\Delta z \to 0$ if and only if $\Delta r \to 0$ and

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{u(r + \Delta r, \theta) + iv(r + \Delta r, \theta) - i(u(r, \theta) + iv(r, \theta))}{\Delta r \exp(i\theta)}$$

and

$$\lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{1}{\exp(i\theta)} \frac{\partial u(r, \theta)}{\partial r} + i \frac{\partial v(r, \theta)}{\partial r}.$$

Similarly, if $\Delta z = r \exp(i(\theta + \Delta \theta)) - r \exp(i\theta) = r \exp(i\theta) [\cos \Delta \theta - 1 + i \sin \Delta \theta]$ then

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{u(r, \theta + \Delta \theta) + iv(r, \theta + \Delta \theta) - i(u(r, \theta) + iv(r, \theta))}{r \exp(i\theta) [\cos \Delta \theta - 1 + i \sin \Delta \theta]}.$$

For $\Delta \theta = 0$ we obtain formally $0/0$, so by l’Hospital’s rule with differentiation with respect to $\Delta \theta$ we obtain

$$\lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{\partial u}{\partial \theta} + i \frac{\partial v}{\partial \theta}.$$

Again, these two limits must be the same so that

$$\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta}$$

and

$$\frac{\partial v}{\partial r} = -\frac{1}{r} \frac{\partial u}{\partial \theta}.$$
The above theorem can be restated. If the partial derivatives with respect to $r$ and $\theta$ exist and are continuous, and if the Cauchy-Riemann equations hold then again we may conclude that $f$ is differentiable. Note that if $z = x + iy$ then we may take

$$f'(z) = u_x(x, y) + iv_x(x, y)$$

and if $z = r \exp(i\theta)$ then we may take

$$f'(z) = \exp(-i\theta)(u_r + iv_r).$$

The Cauchy-Riemann equations have a very important consequence. Suppose that $f(z) = u(x, y) + iv(x, y)$ is differentiable. And suppose also that $u$ and $v$ have continuous second partial derivatives. Then

$$u_{xx}(x, y) + u_{yy}(x, y) = 0$$

$$v_{xx}(x, y) + v_{yy}(x, y) = 0.$$  

This property follows immediately from the Cauchy-Riemann equations. The order of differentiation for smooth functions can be reversed so that adding

$$u_{xx} = v_{yx}$$

$$u_{yy} = -v_{xy} = -v_{yx}$$

shows that $u_{xx} + u_{yy} = 0$. The result for $v$ follows analogously. The equation

$$Lw(x, y) \equiv w_{xx} + w_{xx} = 0$$

is known as Laplace’s equation and functions which satisfy this equation are called harmonic. Laplace’s equation arises in the context of steady state heat transfer, potential flow of fluids and electrostatics. The fact that smooth $u$ and $v$ are harmonic means that we obtain a catalog of solutions for Laplace’s equation from the differentiable complex functions. Moreover, we shall learn later that once differentiable complex functions are in fact infinitely differentiable so that continuous first partial derivatives satisfying the Cauchy-Riemann equations characterize harmonic functions $u$ and $v$. 

158
Finally, suppose that we have a harmonic function \( u(x, y) \). Then \( u \) is the real part of a differentiable complex valued function \( f(z) \) and we can use the Cauchy-Riemann equations to find the so-called harmonic conjugate \( v(x, y) \). Suppose, for example, that

\[
  u(x, y) = x^3 - 3xy^2 + y.
\]

It is straightforward to verify that \( u \) is harmonic. Then

\[
  \frac{\partial v}{\partial y} = \frac{\partial u}{\partial x} = 3x^2 - 3y^2
\]

so that

\[
  v(x, y) = 3x^2y - y^3 + g(x).
\]

From the second Cauchy-Riemann equation follows

\[
  \frac{\partial v}{\partial x} = 6xy + g'(x) = -\frac{\partial u}{\partial y} = 6xy - 1
\]

so that \( g(x) = -x + c \) for an arbitrary constant \( c \). Hence

\[
  f(z) = (x^3 - 3xy^2 + y) + i(3x^2y - y^3 - x) + c.
\]

We recognize that

\[
  f(z) = z^3 - iz + c
\]

is another way of writing this function.

Since the real and imaginary parts of any differentiable complex valued function have to be harmonic, and since polynomials \( P_n(z) \) are differentiable we see that the null space of the partial differential operator

\[
  Lw \equiv w_{xx}(x, y) + w_{yy}(x, y)
\]

contains infinitely many linearly independent elements (e.g., \( \text{Re } z^n \) and \( \text{Im } z^n \) for any \( n \)).

The real and imaginary parts of a differentiable complex valued function of \( z \) have additional importance beyond solving Laplace’s equation. They are used to define new orthogonal coordinate systems in the plane. Consider

\[
  f(z) = u(x, y) + iv(x, y)
\]
where $f$ is differentiable. Then

$$u(x, y) = \text{constant}$$

$$v(x, y) = \text{constant}$$

define curves in the $x$-$y$ plane. For example

$$u(x, y) = u(x_0, y_0)$$

$$v(x, y) = v(x_0, y_0)$$

would be two curves passing through the point $(x_0, y_0)$. Suppose that we can parametrize these curves so that for $R(t) = (x(t), y(t))$ and $T(s) = (x(s), y(s))$ we have $u(x(t), y(t)) = u(x_0, y_0)$ and $v(x(s), y(s)) = v(x_0, y_0)$. Then

$$\frac{d}{dt} u(x(t), y(t)) = ux' + uy' = \langle \nabla u, R'(t) \rangle = 0$$

and

$$\frac{d}{dt} v(x(s), y(s)) = ux' + uy' = \langle \nabla v, T'(s) \rangle = 0$$

which state that the tangents $R'$ and $T'$ to the curves $u = \text{constant}$ and $v = \text{constant}$ are orthogonal (in the Euclidean sense) to the gradients of $u$ and $v$, i.e., that the vectors $\nabla u$ and $\nabla v$ are perpendicular to the level sets $u = \text{constant}$ and $v = \text{constant}$. But from the Cauchy-Riemann equations we find

$$\langle \nabla u(x_0, y_0), \nabla v(x_0, y_0) \rangle = 0.$$

Hence the curves $u = \text{constant}$ and $v = \text{constant}$ cross at right angles. For example, if

$$f(z) = z^2 = (x + iy)^2 = x^2 - y^2 + 2ixy$$

then the level sets through the point, say, $(2, 1)$, are given by

$$x^2 - y^2 = 3$$

$$2xy = 4.$$
It is simple enough to graph these curves and observe that they cross at right angles. Hence the real and imaginary parts of smooth complex functions can be used to introduce new orthogonal coordinate systems. We shall conclude this discussion of the Cauchy-Riemann equations and their applications by considering the following problem. Suppose we need to find a root of a nonlinear equation

\[ f(x) = 0. \]

In general this can only be done numerically. By far the most efficient method is Newton’s method – if it works. We start with an initial guess \( x^0 \) (which generally must be close to the solution \( x^* \)) and iteratively generate a sequence \( \{x^k\} \) of improved estimates. Specifically, given \( x^k \) we find \( x^{k+1} \) as the solution of the linear equation

\[ Lx = 0 \]

where

\[ Lx \equiv f'(x^k)(x^{k+1} - x^k) + f(x^k). \]

Note that \( y = Lx \) is just the equation of the tangent to \( f \) at \( x^k \). It is known that Newton’s method is equally applicable to vector systems. For example, if we are to solve

\[ u(x, y) = 0 \]
\[ v(x, y) = 0 \]

then we generate a sequence \( \{(x^k, y^k)\} \) by solving for \( (x^{k+1}, y^{k+1}) \) the matrix equation

\[ L \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} u_x(x^k, y^k) & u_y(x^k, y^k) \\ v_x(x^k, y^k) & v_y(x^k, y^k) \end{pmatrix} \begin{pmatrix} (x - x^k) \\ (y - y^k) \end{pmatrix} + \begin{pmatrix} u(x^k, y^k) \\ v(x^k, y^k) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \]

Consider now the problem of finding a root of

\[ f(z) = 0 \]

where \( f \) is a differentiable function of the complex variable \( z \). This problem is, of course, resolvable by solving for the zeroes of \( u \) and \( v \) simultaneously as just outlined. The question
now is: Can the problem also be solved as a scalar problem, albeit in complex arithmetic, by computing the solution \( z^{k+1} \) of the linear approximation

\[
Lz \equiv f'(z^k)(z - z^k) + f(z^k) = 0.
\]

Let us write \( z^k = x^k + iy^k \) and \( f'(z^k) = u_x(x^k, y^k) + iv_x(x^k, y^k) \) and isolate the real and imaginary parts of \( Lz \). We find that \( z^{k+1} = x^{k+1} + iy^{k+1} \) solves

\[
\begin{pmatrix}
  u_x(x^k, y^k) & -v_x(x^k, y^k) \\
  v_x(x^k, x^k) & u_x(x^k, y^k)
\end{pmatrix}
\begin{pmatrix}
  (x - x^k) \\
  (y - y^k)
\end{pmatrix}
\begin{pmatrix}
  u(x^k, y^k) \\
  v(x^k, y^k)
\end{pmatrix} = \begin{pmatrix}
  0 \\
  0
\end{pmatrix}.
\]

But by the Cauchy-Riemann equations \(-v_x = u_y\) and \(u_x = v_y\) so that the equation \( Lz = 0 \) is exactly the same as the real system \( L \begin{pmatrix} x \\ y \end{pmatrix} = 0 \). Thus Newton’s method for a single scalar equation is valid for real as well as for complex variables.
Module 20 - Homework

1) Let \( f(z) = e^{e^z}, \ z = re^{i\theta}, \ \theta = \text{Arg}(z). \)
   
i) Verify the Cauchy-Riemann equations in Cartesian and polar coordinates.  
      Where do they hold?  
   ii) Find \( f'(z). \)

2) i) Show that \( g(x, y) = \sin x \cosh y \) is the real part of a differentiable function \( f(z). \) Find 
      \( f(z). \)  
   ii) Show that \( h(x, y) = \sin x \cosh y \) is the imaginary part of a differentiable function 
      \( f(z). \) Find \( f(z). \)

3) Let \( x = r \cos \theta \) and \( y = r \sin \theta. \)
   i) Show that \( \frac{\partial x}{\partial r} = \frac{\partial r}{\partial x}. \)
   ii) Transform the Cauchy-Riemann equations

\[
\begin{align*}
  u_x &= v_y \\
  u_y &= -v_x
\end{align*}
\]

into polar coordinates with a formal change of variable.

4) Prove or disprove: Newton’s method in complex arithmetic for

\[
f(z) \equiv z^2 + 1 = 0
\]

cannot converge for any real initial guess \( z^0. \)

5) Let \( f(z) = z^3. \)
   i) Find the angle between the positive \( x \)-axis and the tangent to \( \text{Re} \ f(z) = \text{Re} \ f(1+i) \) 
   at \( z = 1 + i. \)
   ii) Find the tangent between the positive \( x \)-axis and the tangent to 

\[
\text{Im} \ f(z) = \text{Im} \ f(i + 1) \text{ at } z = 1 + i. \text{ Verify that \( \text{Re} \ f(z) \) and \( \text{Im} \ f(z) \) cross at right angles at } z = 1 + i.
\]

6) Use the Cauchy-Riemann equations in polar coordinates to derive a single second order equation for \( u \) and \( v. \) The resulting equation is Laplace’s equation in polar coordinates.
7) Let \( f(z) = u + iv \) be a differentiable function such that \( u \) and \( v \) are harmonic. Let \( D \) be an open set in the \( z \)-plane and let \( D' \) be the image of \( D \) in the \( w \)-plane. Finally, suppose that

\[
\phi(u, v)
\]

is harmonic in \( D' \). Use the chain rule to show that

\[
\phi(u(x, y), v(x, y)) \quad \text{is harmonic in} \quad D.
\]
Now that we have clarified the meaning of differentiability of a complex valued function of a complex variable we may begin to extend the elementary functions of the real calculus to the complex plane. Such an extension is not necessarily unique. For example, both

\[ f(z) = z \quad \text{and} \quad f(z) = \bar{z} \]

may be thought of as the extension of the identity function \( f(x) = x \) to the complex plane. However, we shall always require that the complex valued function be differentiable at all \( x \) where \( f \) as a function of \( x \) is differentiable. It turns out that this makes the extension unique. For example, \( f(z) = \bar{z} \) would not be an admissible extension of \( f(x) = x \) to the complex plane.

**The complex exponential:**

**Definition:**

\[ e^z = e^x e^{iy} = e^x [\cos y + i \sin y] . \]

We observe that if \( z = x \) then \( e^z = e^x \). In addition

\[ \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y} = e^x \cos y \]
\[ \frac{\partial u}{\partial y} = \frac{\partial v}{\partial x} = e^x \sin y \]

so that the partial derivatives are continuous everywhere and satisfy the Cauchy Riemann equations. Hence

\[ e^z = e^x e^{iy} = e^x \cos y + i e^x \sin y \]

is the correct extension of the exponential to the whole complex plane.

**Properties of the complex exponential:**

i) \( \frac{d}{dx} e^z = u_x + iv_x = e^z \) which we see by inspection

ii) \( e^{a+b} = e^a e^b \) for any complex numbers \( a \) and \( b \)

This result is not obvious and can be shown as follows:
Define $f(z) = e^z e^{c-z}$ for constant $c$ then by the product and chain rule $f'(z) = e^z e^{c-z} - e^z e^{c-z} = 0$. This implies that $u_x = u_y = v_x = v_y = 0$ so that $f(z)$ is constant. Thus $f(z) = f(c)$ or $e^z e^{c-z} = e^c$. The result follows if we set $c = a + b$ and $z = b$.

Note that this last result was shown without recourse to any trigonometric addition formulas. It implies

$$e^{i(\theta + \phi)} = \cos(\theta + \phi) + i \sin(\theta + \phi)$$

$$= (\cos \theta + i \sin \theta)(\cos \phi + i \sin \phi).$$

Equating real and imaginary parts we obtain

$$\cos(\theta + \phi) = \cos \theta \cos \phi - \sin \theta \sin \phi$$

$$\sin(\theta + \phi) = \cos \theta \sin \phi + \sin \theta \cos \phi.$$ 

Similarly,

$$(e^{i\theta})^n = (e^{in\theta})$$

so that

$$(\cos \theta + i \sin \phi)^b = \cos b\theta + i \sin b\theta.$$ 

By equating again the real and imaginary parts we see that $\cos n\theta$ and $\sin n\theta$ can always be expressed as functions of $\cos \theta$ and $\sin \theta$.

The geometry of $w = e^z$:

$$u + iv = e^x[\cos y + i \sin y]$$

can be interpreted in the vector sense

$$(u, v) = e^x(\cos y, \sin y)$$

and implies the following:

The lines $y = \text{constant}$ in the $z$-plane are mapped to rays in the $w$-plane, but we cannot get to the origin in the $w$-plane. The lines $x = \text{constant}$ in the $z$-plane are mapped to circles of radius $e^x$ in the $w$-plane.

Complex trigonometric functions:

If we define

$$\cos z = \frac{e^{iz} + e^{-iz}}{2}, \quad \sin z = \frac{e^{iz} - e^{-iz}}{2i}$$
then it follows from the definition of the complex exponential that \( \cos z \) and \( \sin z \) are the standard trigonometric functions when \( z = x \). Moreover, since the exponential now is known to be differentiable it follows from the chain rule that \( \cos z \) and \( \sin z \) are differentiable. Moreover,

\[
\frac{d}{dz} \cos z = \frac{ie^{iz} - e^{-iz}}{2} = -\sin z
\]

and similarly,

\[
(\sin z)' = \cos z.
\]

A little bit of algebra shows that

\[
\sin^2 z + \cos^2 z = 1
\]

just as in the real case.

Analogously we define the hyperbolic cosine and sine by:

\[
\cosh z = \frac{e^z + e^{-z}}{2} \quad \text{and} \quad \sinh z = \frac{e^z - e^{-z}}{2}
\]

which implies that

\[
(\cosh z)' = \sinh z \quad \text{and} \quad (\sinh z)' = \cosh z.
\]

Algebra again shows that for all \( z \)

\[
\cosh^2 z - \sinh^2 z = 1.
\]

However, in the complex plane the distinction between the trigonometric and hyperbolic functions is blurred because

\[
\cos iy = \cosh y
\]

\[
\sin iy = i \sinh y.
\]

In particular, this shows that the trigonometric functions are not bounded in the complex plane. For example, we can always find a \( z \) such that, e.g.,

\[
\sin z = 7 \quad \text{and} \quad \cos z = -3 + i.
\]
The logarithm of a complex variable:

In contrast to the exponential and the trigonometric functions the logarithm is considerably more complicated.

**Definition:** \( w = \log z (\equiv \ln z) \) denotes all those numbers \( w \) which satisfy

\[
e^w = z.
\]

The definition simply states that the log is the inverse of the exponential. To compute \( w \) we express \( z \) in polar form

\[
z = re^{i\theta} \text{ where } r = |z| \text{ and } \theta = \arg(z).
\]

Since \( w = u + iv \) we find

\[
e^{u+iv} = e^u e^{iv} = re^{i\theta}.
\]

This implies that \( e^u = r \) and \( v = \arg(z) \). Since \( e^u > 0 \) it follows that \( \log z \) cannot be defined for \( z = 0 \). But for \( z \neq 0 \) \( \arg(z) \) is not uniquely defined but has infinitely many values which differ by multiples of \( 2\pi \). Hence \( \log z \) is a multiple valued function

\[
\log z = u + iv = \log r + i \arg(z)
\]

where \( \log r \) denotes the real log of a positive number. It is a consequence of the definition of \( \log z \) that

\[
e^{\log z} = e^{\log r + i \arg(z)} = re^{i \arg(z)} = z
\]

\[
\log e^z = \log e^x e^{iy} = \log |e^x| + iy + 2\pi ki = z + 2\pi ki.
\]

Hence the nice symmetry of one function being the inverse of the other is lost when we define the log in the above form. Moreover, the log no longer is a function in the usual sense and writing

\[
\log(z + \Delta z) - \log z
\]

would not make sense because \( \log z \) is a whole set of numbers. Yet we wanted to extend the real logarithm differentiably to the complex plane, if possible. We had learned before that the multiple valued function \( \arg(z) \) can be made a standard function if we pick a branch.
where \( \arg(z) \) is constrained to lie in a prescribed interval of length \( 2\pi \). We shall agree to choose the principal branch \( \text{Arg}(z) \) and define accordingly the principal branch of the logarithm by

\[
\log z = \log |z| + i \text{Arg}(z) = \log r + i\theta
\]

where \( \theta \in (-\pi, \pi] \). We observe that this notation is consistent with our earlier usage since

\[
\log x = \log |x| + i0 \quad \text{for} \ x \ \text{real and positive}.
\]

It it easy to see that \( \log z \) is defined for all \( z \neq 0 \). Moreover, away from the negative axis \( x < 0 \) the function is differentiable because the Cauchy-Riemann equations in polar coordinates

\[
\frac{\partial u}{\partial r} = \frac{1}{r} \frac{\partial v}{\partial \theta} = \frac{1}{r}
\]

and

\[
\frac{\partial v}{\partial r} = \frac{1}{r} \frac{\partial u}{\partial \theta} = 0
\]

hold. Across the negative axis \( \log z \) shows a jump, e.g.,

\[
\lim_{n \to \infty} \log(-5 - i/n) = \log 5 - i\pi
\]

\[
\lim_{n \to \infty} \log(-5 + i/n) = \log 5 + i\pi.
\]

But away from the negative axis the function is differentiable and satisfies

\[
\frac{d}{dz} \log z = \frac{1}{e^{i\theta}} \left[ \frac{\partial u}{\partial r} + i \frac{\partial v}{\partial r} \right] = \frac{1}{z}.
\]

**The complex power function** \( z^\alpha \):

For real \( x > 0 \) and any \( \alpha \) the function \( x^\alpha \) is shorthand notation for

\[
x^\alpha = e^{\alpha \log x}
\]

which is perfectly well defined and straightforward to handle. Analogously, we have

**Definition:** For \( z \neq 0 \) and any \( \alpha \)

\[
z^\alpha = e^{\alpha \log z}.
\]
Note that this definition implies that in general $z^\alpha$ is multiple valued because $\log z$ is multiple valued. For example, we already have discussed the roots of $z$. Recall, for $z = |z|e^{i\theta}$

$$w = z^{1/n}$$

has the $n$ values

$$w_k = |z|^{1/n}e^{i\left(\frac{\theta + 2\pi k}{n}\right)}, \quad k = 0, 1, 2, \ldots, n - 1.$$  

According to the above definition we have

$$z^{1/n} = e^{\frac{1}{n}\log z} = e^{\frac{1}{n}[\log |z| + i\arg(z) - i\arg(z) + i\arg(z)]}$$

$$= |z|^{1/n}e^{-i\frac{\arg(z)}{n}} = |z|^{1/n}e^{i\left(\frac{\theta + 2\pi k}{n}\right)}$$

where $\theta$ is any one particular value of $\arg(z)$. For example

$$i^i = e^{i\log i} = e^{i[\log 1 + i(\pi/2 + 2\pi k) = e^{-1/2 + 2\pi k}}$$

for $k = 0, \pm 1, \pm 2, \ldots$ Again, as a multiple valued function we cannot talk about derivatives for the complex power function; we would again have to restrict ourselves to a single valued branch of the logarithm, like the principal value branch. However, we shall not pursue this topic any further.
Module 21 - Homework

1) Plot the image of the unit square \((0, 1) \times (0, 1)\) in the \(z\)-plane under the transformation \(w = e^z\).

2) Show that the function

\[
f(z) = \frac{1}{2} \left( z + \frac{1}{z} \right)
\]

i) maps a circle \(|z| = r, r \neq 1\) onto an ellipse in the \(w\)-plane.

ii) maps the unit circle \(|z| = 1\) onto the real interval \([-1, 1]\) in the \(w\)-plane.

3) Show that \(\sin z = 0\) if and only if \(z = n\pi\) for any integer \(n\).

Show that \(\cos z = 0\) if and only if \(z = \pi/2 + n\pi\) for any integer \(n\).

4) Evaluate: \(\log i, \log(3 + i), \log(3 - i), z^0\) for any \(z \neq 0, (1 + i)^{(1-i)}\).

5) Find all \(z\) such that \(\sin z = 7 - 3i\).

6) Find all \(z\) such that \(\cosh z = .5\).

7) Prove or disprove:

\[
\log z_1 z_2 = \log z_1 + \log z_2
\]

\[
\Log z_1 z_2 = \Log z_1 + \Log z_2
\]
In contrast to real integrals in $\mathbb{R}^2$ where line and area integrals occur, the theory of a complex variable only knows line integrals where a function of a complex variable is integrated over some curve in the $z$-plane. In this brief introduction to complex variables we shall consider only curves which consist of smooth segments where there is tangent to the curve except at isolated points where the curve may have a corner. Such a curve is usually called a contour. It need not loop back on itself, and it may cross itself. If the contour does not cross itself and ends where it started we speak of a simple closed curve.

Each segment of the curve will be described parametrically by

$$z(t) = x(t) + iy(t), \quad a \leq t \leq b$$

where $x'$ and $y'$ exist on $(a, b)$. For example,

$$z(t) = z_0 t + (1 - t)z_1, \quad t \in [0, 1]$$

describes a straight line from $z_0$ to $z_1$,

$$z(t) = a \cos t + ib \sin t, \quad a, b > 0, \quad t \in [0, 2\pi]$$

describes a simple closed curve starting at $z(0)$ counterclockwise around the ellipse

$$\left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 = 1$$

back to $z(0)$.

We remark that it is particularly useful in this setting to identify the contour in the complex plane with the curve in $\mathbb{R}^2$ traced out by the position vector

$$R(t) = (x(t), y(t)).$$

We know from particle dynamics, for example, that $R'(t)$ is a vector parallel to the tangent to the curve at the point $R(t)$ which points in the direction of motion ($R'(t)$ is the velocity vector).
Frequently it is useful to describe the curve in polar coordinates

\[ z(t) = r(t)e^{i\theta(t)}. \]

For example,

\[ z_1(t) = te^{it} \quad t \in [0, 8\pi] \]
describes a spiral from the origin to the point \((8\pi + i0)\) which crosses the \(y\)-axis four times.

Suppose we have a contour \(\Gamma_+\) in the complex plane which is piecewise smooth and given parametrically by \(z(t), a \leq t \leq b\). Then the contour integral over \(\Gamma_+\) is defined as

\[
\int_{\Gamma_+} f(z)dz = \int_a^b f(z(t))z'(t)dt.
\]

The real and imaginary parts of this integral are simple real integrals of a real variable \(t\) and are readily computed with the standard calculus. In particular, if we reverse the direction of travel so that we travel along \(\Gamma_-\) where \(t\) goes from \(b\) to \(a\) then

\[
\int_{\Gamma_-} f(z)dz = \int_b^a f(z(t))z'(t)dt = -\int_a^b f(z(t))z'(t)dt = -\int_{\Gamma_+} f(z)dz.
\]

In other words, reversing the direction of travel changes the algebraic sign of the contour integral.

The following example shows that the direction of integration can also be built into the parametrization of \(z\). Suppose that

\[ f(z) = 1/z, \]

that \(\Gamma_+\) is the unit circle and that the parametrization is

\[ z(t) = e^{it} \quad t \in [0, 2\pi]. \]

Then we move around the unit circle in the counterclockwise direction and obtain

\[
\int_{\Gamma_+} f(z)dz = \int_0^{2\pi} e^{-it}(ie^{it})dt = 2\pi i.
\]

If

\[ z(t) = e^{-it} \quad t \in [0, 2\pi] \]
then we move around the unit circle in the clockwise direction and obtain

\[ \int_{\Gamma} f(z)dz = \int_{0}^{2\pi} e^{it}(-ie^{-it})dt = -2\pi i. \]

Suppose that we have a simple closed curve. Then this contour separates the complex plane into two components, one of which is bounded while the other is unbounded. If we traverse the path in the direction such that the bounded component always is to our left then we say that we traverse the contour in the positive direction. If the path is near circular this would correspond to going around it in the counterclockwise direction. If the bounded component is on the right as the path is traversed we travel in the negative direction. If not stated otherwise all simple closed curves are traversed in the positive direction.

**Cauchy’s Theorem:**

**Definition:** A complex valued function of a complex variable is analytic in an open set \( D \) if it has a derivative at every point of \( D \).

It follows that if \( f \) is analytic in \( D \) then the Cauchy-Riemann equations have to hold at every point of \( D \). Let us now suppose that \( D \) is an open set such that

i) any two points in \( D \) can be connected with a contour in \( D \) (so \( D \) is not the union of disjoint sets). We call such an open set a DOMAIN.

ii) \( D \) does not have any holes in it. We call such a set a SIMPLY CONNECTED DOMAIN.

We usually visualize a simply connected domain \( D \) as a (perhaps highly) distorted disk but it may be unbounded such as a wedge, a quarter, a half or the whole complex plane.

**Cauchy’s integral theorem:**

Let \( f \) be analytic in a simply connected domain. Let \( \Gamma \) denote a a closed curve in \( D \). Then

\[ \int_{\Gamma} f(z)dz = 0. \]

The proof of this theorem is accessible when \( \Gamma \) is the boundary of a rectangle but becomes quite involved for general closed curves. We shall not reproduce the proof but attempt to make the result plausible by interpreting the real and imaginary parts of the integral as the
work done moving in a conservative force field. Consider
\[
\text{Re} \int_{\Gamma} f(z)dz = \text{Re} \int_{\Gamma} (u(x,y) + iv(x,y))(x'(t) + iy'(t))dt
= \int_{\Gamma} \langle (u(x(t), y(t)), -v(x(t), y(t)), R'(t)) \rangle dt.
\]
This integral states that we are moving along the curve traced out by the position vector \( R(t) \) under the influence of a force \( F = (F_1, F_2) = (u(x, y), -v(x, y)) \). The force is conservative if
\[ F(x, y) = \nabla \phi(x, y) \]
for some potential \( \phi \). Such a \( \phi \) can be found if
\[
\frac{\partial F_1}{\partial y} = \frac{\partial F_2}{\partial x} \quad \text{and the derivatives are continuous.}
\]
For the above force this relationship is
\[
\frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.
\]
This is just one of the Cauchy-Riemann equations which holds everywhere in \( D \) because \( f \) is assumed to be analytic. But if \( F = \nabla \phi \) then
\[
\int_{\Gamma} \frac{d}{dt} \phi(x(t), y(t))dt = \int_{\Gamma} \langle \nabla \phi(x, y), R'(t) \rangle dt = 0
\]
because \( R(t) \) describes a closed curve. The second Cauchy-Riemann equation is used in a similar fashion to show that
\[
\text{Im} \int_{\Gamma} f(z)dz = 0.
\]
In many applications the closed curve \( \Gamma \) and the function \( f \) are given explicitly. If \( f \) is differentiable everywhere inside and on \( \Gamma \) then one can argue that \( \Gamma \) can be enclosed in a slightly larger simply connected domain to which Cauchy’s theorem applies. Hence the key always is to check whether \( f \) is differentiable everywhere inside and on \( \Gamma \) before applying Cauchy’s theorem.

**Applications:**

1) Let \( P_N(z) \) be a polynomial of degree \( N \), then
\[
\int_{\Gamma} P_N(z)dz = 0
\]
around any closed curve in the complex plane.

2) Let $\Gamma$ be any closed curve in the upper half plane $y > 0$ then

$$\int_{\Gamma} \frac{1}{\sin z} \, dz = 0$$

because $\sin z$ is analytic everywhere and $\sin z = 0$ only for $z = n\pi$.

A common application of Cauchy’s theorem is the transformation of a contour integral around a complicated closed curve into an integral over a simpler closed curve where the integration is easier to carry out.

Let us illustrate the process for the following problem. Let $\Gamma$ denote a simple closed curve which encloses the origin. Suppose we wish to compute

$$\int_{\Gamma} \frac{1}{z} \, dz$$

where the integral, as usual, is to be evaluated in the positive sense. Cauchy’s theorem does NOT predict a value of zero for this integral because any simply connected domain containing $\Gamma$ would have to include the origin and $f$ is definitely not differentiable at $z = 0$.

However, let $|z| = R$ be a circle which encloses $\Gamma$. Let us connect $\Gamma$ and $|z| = R$ and travel around the domain between the two closed curves on the path $\Gamma$ indicated in the drawing.

The function $f(z) = 1/z$ is differentiable everywhere in the interior to our path and on the path. So by Cauchy’s theorem

$$\int_{\Gamma} \frac{1}{z} \, dz = 0.$$ 

Since we traverse the connection between $\Gamma$ and the circle twice in opposite directions the contributions from these segments cancel. Since the circle is traversed in the negative direction we transpose and find

$$\int_{\Gamma} \frac{1}{z} \, dz = \int_{|z|=R} \frac{1}{z} \, dz.$$ 

The second integral is easy to evaluate. We set $z = Re^{it}, \, dz = iRe^{it} \, dt, \, t \in [0, 2\pi]$ and find that

$$\int_{\Gamma} \frac{1}{z} \, dz = 2\pi i$$

for any simple closed curve which loops once around the origin.
Cauchy’s theorem applies to analytic functions. If a function is not analytic in the domain enclosed by $\Gamma$ then, as the last example shows, one has to integrate the contour integral explicitly. The function may, of course, still integrate to zero. For example, suppose that there exists an $F(z)$ such that

$$\frac{d}{dz} F(z) = f(z) \quad \text{for all } z \text{ on } \Gamma.$$ 

Then

$$\int_{\Gamma} f(z(t))z'(t)dt = \int_{\Gamma} \frac{d}{dt} F(z(t))z'(t)dt = F(z(b)) - F(z(a))$$

so that in this case the integral is zero if $z(a) = z(b)$. For example, for $z(t) = \text{Re}^{it}$ we find

$$\int_{|z| = R} z^n dz = \frac{z^{n+1}}{n+1} \bigg|_{z(0)}^{z(2\pi)} = 0 \quad \text{for } n \neq -1.$$ 

It is instructive to look again at

$$f(z) = \frac{1}{z}.$$ 

We have seen that

$$\frac{d}{dz} \log z = \frac{1}{z} \quad \text{everywhere except on the negative } x\text{-axis.}$$

If the circle of radius $R$ is written as

$$\Gamma_1 \cup \Gamma_2$$

where $\Gamma_1$ is the semi-circle in the upper half plane and $\Gamma_2$ is the semi-circle in the lower half plane then

$$\int_{|z| = R} \frac{1}{z} dz = \int_{\Gamma_1} \frac{1}{z} dz + \int_{\Gamma_2} \frac{1}{z} dz$$

$$= \log(\text{Re}^{i\pi}) - \log(\text{Re}^{i0}) + \log(\text{Re}^{i0}) - \lim_{\Delta \theta \to 0} \log(\text{Re}^{i(\pi - \Delta \theta)})$$

$$= i\pi - (-i\pi) = 2\pi i.$$ 

This way we simply have avoided integrating across the discontinuity of the antiderivative $F(z)$ of $f(z)$. 

177
Module 22 - Homework

1) Let $f(z)$ be continuous on $\Gamma$.
   i) Show that
   \[ \left| \int_{\Gamma} f(z) \, dz \right| \leq \int_{\Gamma} |f(z)| \, |dz|. \]
   ii) Suppose that $|f(z)| < M$ on $\Gamma$. Show that
   \[ \left| \int_{\Gamma} f(z) \, dz \right| \leq M l(\Gamma) \]
   where $l(\Gamma)$ is the length of $\Gamma$.

2) Let $f(z) = \text{Re} \, z^3$. Compute
   \[ \int_{|z|=1} f(z) \, dz \]

3) Let $\Gamma_1$ and $\Gamma_2$ be the arcs of the circle of radius 4 from $z = -1 - 3i$ to $z = -5 + i$ and from $z = -5 + i$ to $z = -1 - 3i$ so that the circle is traversed in the positive direction.
   i) Compute
   \[ \int_{\Gamma_1} z^3 \, dz, \quad \int_{\Gamma_2} z^3 \, dz \]
   ii) \[ \int_{\Gamma_1} \frac{1}{z} \, dz, \quad \int_{\Gamma_2} \frac{1}{z} \, dz \]
   iii) \[ \int_{\Gamma_1} \text{Log} \, z \, dz, \quad \int_{\Gamma_2} \text{Log} \, z \, dz \]

4) Integrate $f(z) = 1/z$ around the contour given by
   \[ z(t) = (2 + \cos t)e^{i2t}, \quad 0 < t < 2\pi. \]
Let $\Gamma$ be a simple closed curve and suppose that $f$ is analytic inside $\Gamma$ and on $\Gamma$. Let us consider the function $g(z)$ defined by

$$g(z) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(s) ds}{s - z},$$

where $z$ is an arbitrary but fixed point inside $\Gamma$. We can rewrite this function in the form

$$g(z) = \frac{1}{2\pi i} \left[ \int_{\Gamma} \frac{f(z)}{s - z} ds + \int_{\Gamma} \frac{f(s) - f(z)}{s - z} ds \right].$$

Since

$$\int_{\Gamma} \frac{ds}{s - z} ds = 2\pi i$$

we see that

$$g(z) = f(z) + \frac{1}{2\pi i} \int_{\Gamma} \frac{f(s) - f(z)}{s - z} ds.$$

Cauchy’s theorem allows us to shrink the contour $\Gamma$ to a circle $|s - z| = \epsilon$ without changing the value of the integral. But as $\epsilon \to 0$

$$\frac{f(s) - f(z)}{s - z} \to f'(z)$$

so that

$$\left| \int_{\Gamma} \frac{f(s) - f(z)}{s - z} ds \right| \simeq |f'(z)| \left| \int_{\Gamma} ds \right| \leq |f'(z)| \int_{|z - s| = \epsilon} |ds| = |f'(z)| 2\pi \epsilon.$$

Since this has to hold for all $\epsilon$ no matter how small, it follows that

$$\int_{\Gamma} \frac{f(s) - f(z)}{s - z} ds = 0,$$

and we have

**Cauchy’s Integral Formula:**

$$f(z) = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(s)}{s - z} ds.$$
where \( \Gamma \) is any simple closed curve around \( z \) and where \( f \) is analytic inside and on \( \Gamma \). Formal differentiation under the integral shows that
\[
f^{(n)}(z) = \frac{n!}{2\pi i} \int \frac{f(s)}{(s - z)^{n+1}} \, ds.
\]
The argument can be made rigorous with an \( \epsilon - \delta \) argument so that we can conclude: If \( f \) is analytic in a domain \( D \) then \( f \) is infinitely differentiable at any point of \( D \). In particular, this implies that if \( f \) is analytic then \( u \) and \( v \) are harmonic without further hypotheses on the differentiability of the Cauchy-Riemann equations. A simple consequence of the formula for the derivatives is the following result. Suppose that on a circle of radius \( R \) around \( z \) the analytic function \( f \) is bounded, i.e.

\[
|f(s)| \leq M \quad \text{on } |s - z| = R
\]
then

\[
|f^{(n)}(z)| \leq \frac{n!M}{R^n}.
\]
Suppose that \( f \) is analytic in the whole complex plane. Such a function is called an entire function. If, in addition, the function is also bounded by same constant \( M \) for all \( z \) then we may take an arbitrarily large circle in our estimate

\[
|f'(z)| \leq \frac{M}{R}
\]
and conclude that \( f'(z) = 0 \) for all \( z \) so that a bounded entire function must be a constant.

**Fundamental Theorem of Algebra:** Every non-constant polynomial \( P_N(z) \) has a root.

**Proof:** We may write \( P_N(z) = z^N(1 + a_{N-1}/z + \cdots + a_0/z^N) \) which shows that for \( |z| > R_0 \), where \( R_0 \) is sufficiently large, we can assert that \( |P_N(z)| > |z|^N/2 \). Let us suppose that \( P_N(z) = 0 \) does not have a solution inside the circle of radius \( R_0 \). Then the function

\[
f(z) = \frac{1}{P_N(z)}
\]
is differentiable and does not blow up inside the circle. Hence \( |f(z)| < K \) for some constant \( K \) for all \( z \) in the circle. Since outside the circle

\[
|f(z)| \leq \frac{2}{R_0^N}
\]
it follows that $f(z)$ is entire and bounded in the whole complex plane. This would imply that $f$ is constant, but this cannot be because by hypothesis, $P_N(z)$ is not a constant. It follows that $P_N(z)$ must have a root $z_1$ inside the circle. One can show by algebraic manipulation that for any integer $k$ the expression $z^k - z_1^k$ can be factored into

$$z^k - z_1^k = (z - z_1^k)Q_{k-1}(z)$$

where $Q_{k-1}(z)$ is a polynomial of degree $k - 1$. This allows us to “deflate” $P_N(z)$:

$$P_N(z) = P_N(z) - P_N(z_1) = (z - z_1)Q_{N-1}(z).$$

We then establish the existence of a root for $Q_{N-1}$ and so on until we have $N$ roots for $P_N$.

Cauchy’s integral formula has many other important applications. For example, since

$$f(z) = u + iv = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(s)}{s - z} ds,$$

it can be shown that the real part of the integral is a solution of

$$u_{xx} + u_{yy} = 0 \quad \text{inside } \Gamma$$

when $u$ is a given function on $\Gamma$. Thus we have a formula for the solution of the potential problem in a simply connected domain $D$ when the potential is given on the boundary of $D$. For a circle the famous Poisson formula results. Here we shall follow another tack and use Cauchy’s integral formula to find a series representation for $f$.

**Theorem:** Let $f$ be analytic in the annulus $D = \{z : r < |z - z_0| < R\}$ then we can expand $f$ in terms of a so-called Laurent series

$$f(z) = \sum_{k=-\infty}^{\infty} a_k (z - z_0)^k \quad \text{for all } z \in D.$$

**Proof:** Let $z$ be an arbitrary point in $D$. Let $\Gamma_1$ and $\Gamma_2$ be two circles around $z_0$ inside the annulus with $z$ in between them. If we integrate around $z$ along the path indicated in class then it follows from Cauchy’s integral formula that

$$f(z) = \frac{1}{2\pi i} \int_{\Gamma_2} \frac{f(s)}{s - z} ds - \frac{1}{2\pi i} \int_{\Gamma_1} \frac{f(s)}{s - z} ds.$$
We can formally rewrite this identity as

\[ f(z) = \frac{1}{2\pi i} \int_{\Gamma_2} \frac{f(s)}{(s - z_0)(1 - \frac{z - z_0}{s - z_0})} \, ds - \frac{1}{2\pi i} \int_{\Gamma_1} \frac{f(s)}{(z - z_0)(z - z_0 - 1)} \, ds. \]

On \( \Gamma_2 \): \(|(z - z_0)/(s - z_0)| < 1 \) and on \( \Gamma_1 \) we have \(|(s - z_0)/(z - z_0)| < 1 \) so that on \( \Gamma_2 \):

\[
\frac{1}{1 - \left( \frac{z - z_0}{s - z_0} \right)} = \sum_{n=0}^{\infty} \left( \frac{z - z_0}{s - z_0} \right)^n
\]

and on \( \Gamma_1 \):

\[
\frac{1}{1 - \left( \frac{s - z_0}{z - z_0} \right)} = \sum_{n=0}^{\infty} \left( \frac{s - z_0}{z - z_0} \right)^n.
\]

Assuming that integration and summation can be interchanged we see that

\[ f(z) = \sum_{k=-\infty}^{\infty} a_k (z - z_0)^k \]

where

\[ a_k = \frac{1}{2\pi i} \int_{\Gamma_2} \frac{f(s)}{(s - z_0)^k} \, ds \quad \text{for } k = 0, 1, 2, \ldots \]

and

\[ a_{-k} = \frac{1}{2\pi i} \int_{\Gamma_1} f(s)(s - z_0)^k \, ds \quad \text{for } k = 1, 2, 3, \ldots \]

The infinite series is known as a Laurent series for the function \( f \). In practice, the series is rarely found by actually integrating around the indicated circles. Rather, the existence of such a series is important and the basis for actual calculations. However, we do observe that if \( f(z) \) is analytic everywhere inside and on the larger contour \( \Gamma_2 \) then

\[ a_{-k} = 0 \quad \text{for } k = 1, 2, \ldots \]

and, by the Cauchy integral formula,

\[ a_k = \frac{f^{(k)}(z_0)}{k!} \quad \text{for } k = 0, 1, 2, \]

so that the Laurent series becomes the Taylor series for \( f \). In particular, the result implies that the Taylor series for \( f \) around \( z_0 \) has a radius of convergence equal to the distance from \( z_0 \) to the nearest singularity of \( f \).
Suppose that \( f \) is analytic in the annulus \( r < |z - z_0| < R \). Then \( f \) can be expanded in a Laurent series

\[
f(z) = \cdots + \frac{a_{-k}}{(z - z_0)^k} + \cdots + \frac{a_{-1}}{z - z_0} + a_0 + a_1(z - z_0) + \cdots
\]

where

\[
a_k = \frac{1}{2\pi i} \int_{\Gamma} \frac{f(z)}{(z - z_0)^{k+1}} \, dz \quad \text{for } k = 0, \pm 1, \pm 2,
\]

and where \( \Gamma \) is any positively oriented simple closed contour in the annulus which encloses \( z_0 \). Let us take for granted that the convergence properties of the Laurent series allow a term by term integration. If \( \Gamma \) is a simple closed curve around \( z_0 \) in the annulus then

\[
\int_{\Gamma} f(z) \, dz = \sum_{k=-\infty}^{\infty} \int_{\Gamma} a_k(z - z_0)^k \, dz = 2\pi i a_{-1}
\]

since all other terms of the series integrate to zero. However, the coefficient \( a_{-1} \) is generally not known, and if one had to find it from its integral representation then one may as well integrate \( f \) directly.

The usefulness of the Laurent expansion increases if the point \( z_0 \) is the center of a disk and \( f \) is analytic at every other point of the disk. Hence from now on we shall assume that

\[
f \text{ is analytic for } 0 < |z - z_0| < R.
\]

If \( f \) is not analytic at \( z_0 \) then we say that \( z_0 \) is an isolated singularity of \( f \). If we now consider the Laurent series of \( f \) in the punctured disk \( 0 < |z - z_0| < R \) we have

**Definition:** The coefficient \( a_{-1} \) is the residue of \( f \) at \( z_0 \).

We use the notation

\[
a_{-1} = \text{Res}(f, z_0).
\]

If \( f \) is analytic at \( z_0 \) then \( \text{Res}(f; z_0) = 0 \). In general, only those points will be of interest where \( f \) is singular. It is generally not difficult to find the points where \( f \) is singular,
although the residue may not always be easy to find. We had seen that if $\Gamma$ is a simple closed curve around $z_0$ in $|z - z_0| < R$ then

$$\int_{\Gamma} f(z)\,dz = 2\pi i \operatorname{Res}(f : z_0).$$

Suppose now that we need to find

$$\int_{\Gamma} f(z)\,dz$$

around some simple closed curve $\Gamma$ enclosing a domain $D$ where $f$ has $N$ isolated singularities at the points $z_j$, $j = 1, 2, \ldots, N$. Away from these points and on $\Gamma$, $f$ is assumed to be analytic. Around each $z_j$ there is an annulus where $f$ is analytic so that in this annulus $f$ can be written in terms of a Laurent series around $z_j$. Moreover, the integral around $\Gamma$ can be deformed into $N$ circles around the $\{z_j\}$ and each circle integrates into $2\pi i \operatorname{Res}(f; z_j)$.

Thus we have the so-called residue theorem:

Let $\Gamma$ be a simple closed positively oriented curve.

Assume that $f$ is analytic on $\Gamma$ and inside $\Gamma$ except at isolated points $\{z_j\}$, $j = 1, \ldots, N$. Then

$$\int_{\Gamma} f(z)\,dz = 2\pi i \sum_{j=1}^{N} \operatorname{Res}(f : z_j).$$

It is not obvious at this time that this theorem is useful because we appear to need to know the Laurent series before we know the residue. However, it usually is possible to classify the singularities of $f$ and to find the residues with simple calculations. Suppose that

$$f(z) = \sum_{k=-\infty}^{\infty} a_k(z - z_0)^k \quad \text{for} \quad 0 < |z - z_0| < R$$

then we have

**Definition:** If $a_{-m} = 0$ for all $m > M$, $a_{-M} \neq 0$ for $M \geq 1$ then $f$ has a pole of order $M$ at $z_0$. If $a_m = 0$ for all $m < M$, $a_M \neq 0$ for $M \geq 1$ then $f$ has a zero of order $M$ at $z_0$.

It can be shown that if $f$ has a pole of order $M$ then there exists an analytic function $g$ such that

$$f(z) = \frac{g(z)}{(z - z_0)^M}$$
where \( g(z_0) \neq 0 \). Similarly, if \( f \) has a zero of order \( M \) then

\[
f(z) = (z - z_0)^M g(z)
\]

where \( g(z) \) is analytic and does not vanish at \( z_0 \). For example, the function

\[
f(z) = \frac{1}{(z^2 - 1)} = \frac{1}{(z - 1)(z + 1)}
\]

has a pole of order 1 (also called a simple pole) at \( z = \pm 1 \). Similarly, the function

\[
f(z) = \frac{\sin z}{\cos z - 1}
\]

has a simple pole at \( z = 0 \) which may be inferred from the Taylor series

\[
\sin z = z - z^3/3! + z^5/5! \cdots
\]

\[
\cos z = 1 - z^2/2! + z^4/4! \cdots
\]

and writing

\[
f(z) = \frac{z(1 - z^2/3! + z^4/5! - \cdots)}{z^2(-1/2! + z^2/4! - \cdots)} = \frac{1}{z} g(z)
\]

Let us now suppose that \( f \) has a pole of order \( m \) at \( z_0 \). Then

\[
(z - z_0)^m f(z) = a_{-m} + a_{-m+1}(z - z_0) + \cdots + a_{-1}(z - z_0)^{m-1} + a_0(z - z_0)^m + \cdots
\]

from which follows that

\[
\lim_{z \to z_0} \frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)] = (m-1)!a_{-1}.
\]

For example, the function

\[
f(z) = \frac{z^3 - 2iz + 5}{(z^2 - 2z + 1)}
\]

has a pole of order 2 at \( z = 1 \). Hence the residue is

\[
a_{-1} = \lim_{z \to 1} \frac{d}{dz} \left[ \frac{(z - 1)^2(z^3 - 2iz + 5)}{z^2 - 2z + 1} \right] = 3 - 2i.
\]

In many instances only simple poles appear in which case

\[
a_{-1} = \lim_{z \to z_0} (z - z_0) f(z).
\]
Not all singularities are poles of order $m$. For example, the function

$$f(z) = \frac{z}{z}$$

has a so-called removable singularity at $z = 0$. Its Laurent series in $0 < |z - 0| < \infty$ is the one term expansion

$$f(z) = 1 \quad \text{on } 0 < |z|.$$  

On the other end of the spectrum are those functions whose Laurent expansion does not terminate at some finite $-m$. For example, since

$$e^w = \sum_{n=0}^{\infty} \frac{w^n}{n!} \quad \text{for all } w,$$

it follows that

$$e^{1/z} = \sum_{n=0}^{\infty} \frac{1}{n! z^n} \quad \text{on } |z| > 0.$$  

When there are non-zero coefficients in the Laurent expansion for infinitely many negative integers then the singularity is called an essential singularity.
Module 23/Module 24 - Homeworks

1) Find the Taylor series for the following functions around the given point $z_0$ and give the radius of convergence.
   i) $f(z) = e^z$, $z_0 = 0$
   ii) $f(z) = \sin z$, $z_0 = 0$
   iii) $f(z) = \cos z$, $z_0 = 0$
   iv) $f(z) = \log z$, $z_0 = 1$
   v) $f(z) = 1/(1 + z^2)$, $z_0 = 1$.

2) Find the Laurent expansion of
   \[ f(z) = \frac{1}{z(z + 1)} \]
   i) which is valid for $0 < |z| < 1$,
   ii) which is valid for $|z| > 1$.
   iii) Find Res($f; 0$).

3) Evaluate
   \[ \int_{|z|=2} \frac{\sin^2 z}{(i + z)^2} \, dz. \]

4) Let $D$ be the triangle in the complex plane with vertices at $-1 + 5i$, $-3 + i$, $1 - i$. Let $\Gamma$ be the positively oriented boundary of $D$. Compute
   i) $\int_{\Gamma} \frac{e^z}{(z^2 + 49)^2(z - 1)} \, dz$
   ii) $\int_{\Gamma} \frac{e^z}{(z^2 + 49)^2z} \, dz$

5) Let $f$ and $g$ be analytic on $|z - z_0| < r$ and suppose that $f(z_0) = g(z_0) = 0$. Prove l'Hospital’s rule:
   \[ \lim_{z \to z_0} \frac{f(z)}{g(z)} = \lim_{z \to z_0} \frac{f'(z)}{g'(z)} \]
   provided the limit exists.
Topics: Application of the residue theorem

Complex contour integrals occur in the computation of inversion formulas for Fourier and Laplace transforms. But they may also be used to evaluate certain definite real integrals.

1) Integration of trigonometric functions:

Suppose we are to compute

\[ \int_0^{2\pi} F(\sin \theta, \cos \theta) d\theta. \]

Since

\[ \sin \theta = \frac{e^{i\theta} - e^{-i\theta}}{2i} \quad \text{and} \quad z = e^{i\theta} \quad \text{for any } z \text{ on } |z| = 1 \]

we find that

\[ \sin z = \frac{z - 1/z}{2i}, \]

and analogously

\[ \cos z = \frac{z + 1/z}{2}. \]

Moreover, from \( z = e^{i\theta} \) follows that

\[ dz = ie^{i\theta} d\theta = iz d\theta. \]

Hence the original integral is a parametrization of the integral

\[ \int_{|z|=1} F\left(\frac{z - 1/z}{2i}, \frac{z + 1/z}{2}\right) \frac{dz}{iz}. \]

But the complex integral is equal to \( 2\pi i \) times the residues of the integrand inside the unit circle.

**Illustration:** Suppose \( b \) is real and \( |b| < 1 \). Then

\[ \int_0^{2\pi} \frac{d\theta}{1 + b \sin \theta} = \int_{|z|=1} \frac{1}{1 + \frac{b(z - 1/z)}{2i}} \frac{dz}{iz} = 2 \int_{|z|=1} \frac{dz}{2iz + bz^2 - b}. \]

The roots of the denominator for \( |b| < 1 \) are

\[ z_1 = -\frac{i (1 + \sqrt{1 - b^2})}{b} \]
and

\[ z_2 = -\frac{i(1 - \sqrt{1 - b^2})}{b}. \]

By inspection \(|z_1| > 1\) so the simple pole at \(z_1\) is outside the unit circle and does not influence the integral. If we set \(\alpha^2 = 1 - b^2\) then we can write

\[ z_2 = i\frac{\sqrt{1 - \alpha}}{\sqrt{1 + \alpha}} \]

which tells us that the integrand has a simple pole inside the unit circle at \(z_2\). We compute with l’Hospital’s rule

\[
\text{Res}(f, z_2) = \lim_{z \to z_2} \frac{z - z_2}{2iz + bz^2 - b} = \frac{1}{2i + 2bz_2} = \frac{1}{2i\sqrt{1 - b^2}}
\]

where \(f\) denotes the integrand \(1/(2iz + bz^2 - b)\). Thus

\[
\int_0^2 \frac{1}{1 + b\sin \theta} d\theta = 2(2\pi i \text{ Res}(f : z_2)) = \frac{2\pi}{\sqrt{1 - b^2}}.
\]

The integral is real so the computation would have to be wrong if the final result were not real.

2) Integration of rational functions along the line.

Let us consider the improper integral

\[
\int_{-\infty}^{\infty} \frac{P_M(x)}{Q_N(x)} \, dx
\]

where \(P\) and \(Q\) are polynomials of degree \(M\) and \(N\). We shall assume that \(Q_N(x) \neq 0\) for any \(x \in (-\infty, \infty)\) and

\[ N \geq M + 2. \]

These conditions assure that the integral exist and that it may be obtained from its principal value

\[
\int_{-\infty}^{\infty} \frac{P_M(x)}{Q_N(x)} \, dx = \lim_{R \to \infty} \int_{-R}^{R} \frac{P_M(x)}{Q_N(x)} \, dx.
\]

Let us now consider the domain \(D\) in the upper half plane bounded by the semi-circle \(z = Re^{i\theta}, \theta \in [0, \pi]\) and the \(x\)-axis. Let \(\Gamma\) be the positively oriented boundary of \(D\). Then

\[
\int_{\Gamma} \frac{P_M(z)}{Q_N(z)} \, dz = 2\pi i \sum \text{Res}(f : z_j)
\]
where the residues have to be found at the poles of \( f(z) = P_M(z)/Q_N(z) \). As in the proof of the fundamental theorem of algebra we know from

\[
Q_N(z) = z^N(a_N + a_{N-1}/z + \cdots + a_0/z^N)
\]

that the roots of \( Q_N(z) = 0 \) have to lie inside some circle with radius \( R_0 \). Hence if in our contour integral \( R \) is sufficiently large then \( \Gamma \) will enclose all singularities of \( f \) in the upper half plane. If we write

\[
\int_\Gamma f(z)dz = \int_{-R}^R f(x)dx + \int_\gamma f(z)dz = 2\pi i \sum \text{Res}(f:z_j)
\]

where \( \gamma \) is the semi-circle then it follows that the left hand side is independent of \( R \) for all \( R > R_0 \) since all residues are now inside \( D \). Finally, by hypothesis we have on

\[
\left| \frac{P_M(z)}{Q_N(z)} \right| > \frac{K}{|z|^2}
\]

for sufficiently large \( R \) and some constant \( K \).

It follows that

\[
\left| \int_\gamma f(z)dz \right| < \left( \frac{K}{R^2} \right) 2\pi R \to 0 \quad \text{as} \quad R \to \infty.
\]

Thus

\[
\int_{-\infty}^{\infty} \frac{P_M(x)}{Q_N(x)} dx = 2\pi i \sum \text{Res} \left( \frac{P_M(z)}{Q_N(z)}, z_j \right)
\]

where the \( z_j \) are the roots of \( Q_N(z) = 0 \) in the upper half plane.

**Illustration:** Consider

\[
\int_{-\infty}^{\infty} \frac{1}{1 + x^6} \, dx.
\]

In this case \( M = 0 \) and \( N = 6 \) and \( Q \) does not vanish on the line. Hence the formula for rational functions applies. We need the zeros of \( Q \). We know that the equation

\[
1 + z^6 = 0
\]

has the six solutions

\[
z_k = e^{i(\pi + 2\pi k)/6}, \quad k = 0, 1, 2, 3, 4, 5.
\]

The first three roots lie in the upper half plane. The other three are the complex conjugates.

There are three simple poles in the upper half plane and the residues are

\[
\lim_{z \to z_k} \left( \frac{(z - z_k)}{1 + z^6} \right) = \frac{1}{6z_k^5}.
\]
It follows that
\[
\int_{-\infty}^{\infty} \frac{1}{1 + x^6} \, dx = \frac{2\pi i}{6} \left[ e^{-5\pi i/6} + e^{-5\pi i/2} + e^{-25\pi i/6} \right] \\
= \frac{2\pi i}{6} \left[ e^{-5\pi i/6} + e^{-\pi i/2} - e^{5\pi i/6} \right] \\
= \frac{2\pi i}{6} \left[ -2i \sin(5\pi/6) - i \right] = \frac{2\pi}{3}.
\]