Pierson Guthrey
pguthrey@iastate.edu

1 Taylor Series

1.1 Important Maclaurin Series

Trigonometric Functions

$$\sin(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - ...$$

$$\cos(x) = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} x^{2n} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - ...$$

Hyperbolic Functions

$$\sinh(x) = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!} = x + \frac{x^3}{3!} + \frac{x^5}{5!} - ...$$

$$\cosh(x) = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n)!} = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} - ...$$

Exponential Function

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + ...$$

Natural Logarithm (for $|x| < 1$)

$$log(1-x) = -\sum_{n=1}^{\infty} \frac{x^n}{n} = -x - \frac{x^2}{2} - \frac{x^3}{3} - ...$$

$$log(1+x) = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{x^n}{n} = x - \frac{x^2}{2} + \frac{x^3}{3} - ...$$

Geometric Series (for $|x| < 1$)

$$\frac{1}{1-x} = \sum_{n=0}^{\infty} x^n = 1 + x + x^2 + x^3 + ...$$

Binomial Series (for $|x| < 1$, $\alpha \in \mathbb{C}$)

$$(1+x)^\alpha = \sum_{n=0}^{\infty} \left(\begin{array}{c} \alpha \\ n \end{array} \right) x^n, \qquad \left(\begin{array}{c} \alpha \\ n \end{array} \right) = \frac{\alpha(\alpha-1)...(\alpha-n+1)}{n!}$$

This includes the square root series for $\alpha = \frac{1}{2}$ and the infinite geometric series for $\alpha = -1$.

$$(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x - \frac{1}{8}x^2 + ...$$

2 Linear Algebra Fundamentals

2.1 Matrix Adjoints

A is Hermitian if $A^* = A$ and symmetric if $A^T = A$

- Eigenvectors of A for distinct eigenvalues are orthogonal
- The algebraic and geometric multiplicities are the same.
- A has a spectral decomposition

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

2.2 Norms

2.2.1 Vector Norms $\|\cdot\|_p$

$$\|x\|_p = \left(\sum |x_i|^p\right)^{\frac{1}{p}}$$

Sum Norm, $p = 1$

$$\|x\|_{sum} = \|x\|_1 = \sum |x_i|$$

Euclidean Norm, $p = 2$

$$\|x\|_{Euclidean} = \|x\|_2 = \sqrt{\sum |x_i^2|}$$

Maximum Norm, $p = \infty$

$$\|x\|_{max} = \|x\|_\infty = max\left\{|x_i|\right\}$$

All finite dimensional norms behave similarly with respect to convergence.

Examples: $\|x\|_2 \leq \|x\|_1$ and $\|x\|_1 \leq \sqrt{n}\|x\|_2$.

2.2.2 Matrix Norms

Matrix norms have the property $\|AB\| \leq \|A\| \|B\|$.
The first variety is the Induced Norms

$$\|A\|_p = \sup_{x\neq 0, x\in\mathbb{C}^n} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\|x\|_p = 1, x\in\mathbb{C}^n} \|Ax\|_p$$

- $p = 1$ leads to the maximum absolute column sum

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

- $p = 2$ leads to the largest singular value of A, ie the square root of the largest eigenvalue of the positive semi-definite matrix A^*A

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^*A)} = \sigma_{\max}(A)$$

- $p = \infty$ leads to the maximum absolute row sum

$$\|A\|_\infty = \max_{1\leq i\leq m} \sum_{j=1}^{n} |a_{ij}|$$

Fun facts

- For any induced norm, $\|A^r\|_p \geq \rho(A)^r$, where equality holds for $p = 2$.

2.2.3 Schatten Norms

The Schatten norms arise when applying the $p-$norm to the vector of singular values of a matrix. Let $\{\sigma_i\}$ be the singular values of A.

$$\|A\|_p = \left(\sum_{i=1}^{\min(m,n)} \sigma_i^p \right)^{\frac{1}{p}}$$

Schatten norms are unitarily invariant. That is, for any unitary matrices U, V,

$$\|A\|_p = \|UAV\|_p$$

- For $p = 1$, we have the **Nuclear norm** also known as the **Trace norm**

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

- $p = 2$. The Frobenius norm is the finite dimensional analog of the Hilbert Schmidt norm

$$\|A\|_F = \|A\|_2 = \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}|^2 \right)^{\frac{1}{2}}$$

  with the interesting property

$$\|A\|_F = \sqrt{\operatorname{tr}(A^*A)} = \sqrt{\sum_{i=1}^{\min(m,n)} \sigma_i^2}$$

- For $p = \infty$, we have the **spectral norm**

$$\|A\|_* = \max_{1 \le i \le \min(m,n)} \sigma_i$$

2.2.4 Matrix Determinants

Useful properties of matrix determinants are

- $\det(AB) = \det(A)\det(B)$
- $\det(A^{-1}) = \det(A)^{-1}$

2.2.5 Condition Numbers

The condition number κ measures the sensitivity of a function with respect to an argument. That it is the maximum relative change.
In general, this is

$$\kappa(f) = \frac{(f(x + \Delta x) - f(x))}{f(x)} \frac{x}{\Delta x} = \frac{(f(x + \Delta x) - f(x))}{\Delta x} \frac{x}{f(x)}$$

For a differentiable function, this is

$$\kappa(f) = \frac{x f'(x)}{f(x)}$$

For $Ax = b$,

$$\kappa(A) = \frac{\left\|A^{-1}\delta b\right\|}{\|\delta b\|} \frac{\|b\|}{\|A^{-1}b\|}$$

Which for any consistent norm (ie, norms with $\kappa(A) \geq 1$) becomes

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

- If $\|\cdot\| = \|\cdot\|_2$, then
$$\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}$$
If A is normal, then
$$\kappa(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$$
If A is unitary, then
$$\kappa(A) = 1$$

In general, for a differentiable function $f(x)$,

$$\kappa(f) = \frac{\|J(x)\|}{\|f(x)\|} \|x\|$$

2.3 Positive Definite Matrices

For a matrix A, the following are equivalent

- All of its eigenvalues are positive
- It may form an inner product such as $\langle \vec{\mathbf{x}}, A\vec{\mathbf{y}} \rangle$
- It is the Gram matrix of linearly independent vectors
- It's leading principle minors are all positive
- It has a unique Cholesky decomposition

2.4 Matrix Calculus

If f is a scalar function of many variables, then

$$\partial f = \frac{\partial f}{\partial \vec{\mathbf{x}}} = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \frac{\partial f}{\partial x_3} \right)$$

One could also take the directional derivative in the direction $\vec{\mathbf{u}}$

$$\partial_{\vec{\mathbf{u}}} f = \partial f \cdot \vec{\mathbf{u}}$$

The derivative of a vector function is written as

$$\frac{\partial \vec{\mathbf{y}}}{\partial \vec{\mathbf{x}}} = (J(\vec{\mathbf{y}}))(\vec{\mathbf{x}}) \qquad J_{ij} = \frac{\partial y_i}{\partial x_j}$$

- $\frac{\partial \vec{\mathbf{x}}}{\partial \vec{\mathbf{x}}} = I$
- $\frac{\partial A\vec{\mathbf{x}}}{\partial \vec{\mathbf{x}}} = A$
- $\frac{\partial \vec{\mathbf{x}}^T A}{\partial \vec{\mathbf{x}}} = A^T$
- $\frac{\partial \vec{\mathbf{u}}^T v}{\partial \vec{\mathbf{x}}} = \frac{\partial \vec{\mathbf{u}}}{\partial \vec{\mathbf{x}}} \vec{\mathbf{v}} + \frac{\partial \vec{\mathbf{v}}}{\partial \vec{\mathbf{x}}} \vec{\mathbf{u}}$

- $\frac{\partial \vec{\mathbf{x}}^T x}{\partial \vec{\mathbf{x}}} = 2\vec{\mathbf{x}}$
- $\frac{\partial \vec{\mathbf{u}}^T A v}{\partial \vec{\mathbf{x}}} = \frac{\partial \vec{\mathbf{u}}}{\partial \vec{\mathbf{x}}} A \vec{\mathbf{v}} + \frac{\partial \vec{\mathbf{v}}}{\partial \vec{\mathbf{x}}} A^T \vec{\mathbf{u}}$
- $\frac{\partial \vec{\mathbf{x}}^T A x}{\partial \vec{\mathbf{x}}} = Ax + A^T x$
- $\frac{\partial^2 \vec{\mathbf{x}}^T A x}{(\partial \vec{\mathbf{x}})^2} = A + A^T$

3 Common Theorems

3.1 Reminders

- The mean value theorem for a differentiable $f(x)$ is

  $$f(x) - f(y) = f'(\theta x + (1-\theta)y)(x-y)$$

  for some $\theta \in [0,1]$.

- For polynomial division, dividing a polynomial of degree n $p(x)$ by the monomial $(x-a)$ leaves you with a polynomial of degree $n-1$ and

  $$p(x) = (x-a)q(x) + r(x)$$

  That is, $q(x), r(x) \in \mathbb{P}_{n-1}$. $q(x)$ and $r(x)$ are unique. Note that $p(a) = r(a)$.

3.2 Fredholm Alternative , Linear Algebra Version

For $A \in \mathbb{C}^{n\times m}$ and $b \in \mathbb{C}^{m\times 1}$,

- Either $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ has a solution $\vec{\mathbf{x}}$
- OR: $A^T \vec{\mathbf{y}} = 0$ has a solution $\vec{\mathbf{y}}$ with $\vec{\mathbf{y}}^T \vec{\mathbf{b}} \neq 0$.

That is, $A\vec{\mathbf{x}} = \vec{\mathbf{b}}$ has a solution if and only if for any $\vec{\mathbf{y}}$ s.t. $A^T\vec{\mathbf{y}} = 0$, $\vec{\mathbf{y}}^T\vec{\mathbf{b}} = 0$.

4 Polynomials

4.1 Special Polynomials

4.1.1 Legendre Polynomials

Defined by

$$\frac{d}{dx}\left((1-x^2)\frac{d}{dx}P_n(x)\right) + n(n+1)P_n(x) = 0$$

Also by Rodrigues' formula

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{(dx)^n}\left((x^2-1)^n\right)$$

or by the recurrence relation

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

They are orthogonal in the sense that

$$\int_{-1}^{1} P_m(x)P_n(x)dx = \frac{2}{2n+1}\delta_{mn}$$

- The first few are

$$\begin{aligned} P_0 &= 1 \\ P_1 &= x \\ P_2 &= \frac{1}{2}(3x^2 - 1) \\ P_3 &= \frac{1}{2}(5x^3 - 3x) \end{aligned}$$

- These arise when you carry out the GS process on the canonical polynomial basis $1, x, x^2, ...$
- They are antisymmetric

$$P_n(-x) = (-1)^n P_n(x)$$

4.1.2 Chebyshev Polynomials

The Chebyshev polynomials of the first kind are defined as solutions to

$$(1-x^2)y'' - xy' + n^2 y = 0$$

$$(1-x^2)y'' - 3xy' + n(n+2)y = 0$$

by the recurrence relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad T_0(x) = 1, T_1(x) = x$$

These also have the important trigonometric definition

$$T_n(x) = \cos(n \arccos(x)) = \cosh(n\operatorname{arccosh}(x))$$

In other words, these satisfy

$$T_n(\cos(\theta)) = \cos(n\theta)$$

They are orthogonal in the sense

$$\int_{-1}^{1} \frac{T_n(x)T_m(x)}{\sqrt{1-x^2}}dx = \begin{cases} 0 & n \neq m \\ \pi & n = m = 0 \\ \frac{\pi}{2} & n = m > 0 \end{cases}$$

The Chebyshev polynomials of the second kind are defined by

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x) \qquad U_0(x) = 1, U_1(x) = 2x$$

These satisfy

$$U_n(\cos(\theta)) = \frac{\sin((n+1)\theta)}{\sin(\theta)}$$

They are orthogonal in the sense

$$\int_{-1}^{1} U_n(x)U_m(x)\sqrt{1-x^2}dx = \begin{cases} 0 & n \neq m \\ \frac{\pi}{2} & n = m \end{cases}$$

Pierson Guthrey
pguthrey@iastate.edu

4.1.3 Hermite Polynomials

The physicists' Hermite Polynomials are defined by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} e^{-x^2} = \left(2x - \frac{d}{dx}\right)^n 1$$

With the recursion relations

$$H_{n+1}(x) = 2xH_n(x) - H_n'(x)$$

and

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x)$$

These are orthogonal in the sense that

$$\int_{-\infty}^{\infty} H_m(x)H_n(x)e^{-x^2}dx = \sqrt{\pi}2^n n!\delta_{nm}$$

4.1.4 Bernstein Polynomials

Let the Bernstein basis polynomials be defined by

$$b_{\nu,n}(x) = \left(\begin{array}{c} n \\ \nu \end{array} \right) x^\nu (1-x)^{n-\nu} \qquad \nu = 0, ..., n$$

with coefficients β_ν. So the Bernstein polynomial of degree n is defined as

$$B_n(x) = \sum_{\nu=0}^{n} \beta_\nu b_{\nu,n}(x)$$

Now consider

$$B_n(f)(x) = \sum_{\nu=0}^{n} f\left(\frac{\nu}{n}\right) b_{\nu,n}(x)$$

It can be shown that

$$\lim_{n\to\infty} B_n(f)(x) = f(x)$$

and the convergence is uniform on $[0, 1]$.

- These are used in the proof of the Stone-Weierstrass approximation theorem.

4.1.5 Lagrange Polynomials

Given a distinct set $\{x_j\}$, the Lagrange basis polynomials are defined by

$$\ell_j(x) = \prod_{0\le m\le k, m\ne j} \frac{x - x_m}{x_j - x_m} \qquad 0 \le j \le k$$

We see that

$$\ell_j(x_m) = \delta_{jm}$$

- These interpolate a function $f(x)$ exactly when

$$L(x) = \sum_{j=0}^{n} \prod_{0\le m\le k, m\ne j} \frac{x - x_m}{x_j - x_m} f(x_j)$$

4.2 Polynomial Interpolation

Interpolation refers to approximations that are exact for certain parts of the data.

4.2.1 Using Lagrange Polynomials

In general, for n data points, polynomial interpolation has the form

$$p_n(x) = \sum_{i=0}^{n} f(x_i) \left(\prod_{0 \le j \le n, j \ne i} \frac{x - x_j}{x_i - x_j} \right)$$

which comes from solving the Vandermonde matrix problem

$$\begin{pmatrix} x_0^n & \dots & x_0^0 \\ \vdots & \ddots & \vdots \\ x_n^n & \dots & x_n^0 \end{pmatrix} \begin{pmatrix} a_n \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} x_0^n & \dots & 1 \\ \vdots & \ddots & \vdots \\ x_n^n & \dots & 1 \end{pmatrix} \begin{pmatrix} a_n \\ \vdots \\ a_0 \end{pmatrix} = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix}$$

These interpolants are exact at the data points.

4.2.2 Using Hermite Polynomials

Hermite interpolation involves a function and it's first m derivatives at n data points.

$$\|f(x) - H(x)\| \; = \frac{f^{(K)}(\eta)}{K!} \prod_{i=1}^{n} (x - x_i)^{k_i}$$

where $K = mn$ is the total number of data points. and $k_i = m - 1$.

4.2.3 Using Newton's Method

The Newton basis polynomials are defined as

$$N(x) = \sum_{j=0}^{k} a_j \eta_j(x)$$

With basis polynomials

$$\eta_0(x) = 1 \qquad\qquad \eta_j(x) = \prod_{j=0}^{j-1} (x - x_i) \qquad j > 0$$

And coefficients defined as the forward differences

$$a_j = [y_0, ..., y_j] = \frac{[y_1, ..., y_j] - [y_0, ..., y_{j-1}]}{x_j - x_0}$$

$$\begin{array}{lllll}
x_0 & y_0 = [y_0] & & & \\
 & & [y_0, y_1] & & \\
x_1 & y_1 = [y_1] & & [y_0, y_1, y_2] & \\
 & & [y_1, y_2] & & [y_0, y_1, y_2, y_3] \\
x_2 & y_2 = [y_2] & & [y_1, y_2, y_3] & \\
 & & [y_2, y_3] & & \\
x_3 & y_3 = [y_3] & & &
\end{array}$$

So then

$$N(x) = [y_k] + [y_k, y_{k-1}]\,(x - x_k) + ... + [y_k, y_0]\,(x - x_k)(x - x_{k-1})...(x - x_1)$$

If x_i are equally spaces with $x_i = x_0 + ih$, $x = s + x_0$, then $x - x_i = s - ih$

$$N(x) = [y_k] + [y_k, y_{k-1}]\,(s - kh) + ... + [y_k, y_0]\,(s - kh)(s - (k-1)h)...(s - h)$$

4.3 Approximation by Polynomial

The Stone-Weierstrass Theorem (also known as Weierstrass Approximation Theorem) states that every continuous function on a closed interval can be uniformly approximated by a polynomial function. That is, for every $\epsilon > 0$, there exists an n and a nth degree polynomial $p_n(x)$ such that

$$\|f - p\|_\infty < \epsilon$$

Consider the Bernstein polynomials on $[0, 1]$. Notice that

$$f(x) = f(x)(x + (1-x))^n = f(x)\sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) x^i (1-x)^{n-i}$$

So then

$$f(x) - (B_n f)(x) = \sum_{i=0}^{n} \left(f(x) - f\left(\frac{i}{n}\right)\right) \left(\begin{array}{c} n \\ i \end{array} \right) x^i (1-x)^{n-i}$$

and thus

$$|f(x) - (B_n f)(x)| \leq \sum_{i=0}^{n} \left| f(x) - f\left(\frac{i}{n}\right)\right| \left(\begin{array}{c} n \\ i \end{array} \right) x^i (1-x)^{n-i}$$

This is a continuous function on a compact interval, thus

- It's maximum will be achieved, say at $x = \eta$
- The function is uniformly continuous on the interval, so for all $\epsilon > 0$, there exists a $\delta > 0$ such that for all x, $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$

Let δ be such that $|f(\eta) - f(x)| < \frac{1}{2}\epsilon$. Now let N denote the index subset such that $\frac{i}{n}$ is within δ of η. So then

$$\begin{aligned} |f(x) - (B_n f)(x)| \leq & \sum_{i \in N} \frac{\epsilon}{2} \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} + \sum_{i \notin N} \left| f(\eta) - f\left(\frac{i}{n}\right)\right| \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \\ < & \frac{\epsilon}{2} + \sum_{i \notin N} \left| f(\eta) - f\left(\frac{i}{n}\right)\right| \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \end{aligned}$$

Now notice that

$$\begin{aligned} \sum_{i \notin N} \left| f(\eta) - f\left(\frac{i}{n}\right)\right| \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \leq & \sum_{i \notin N} \left| f(\eta) - f\left(\frac{i}{n}\right)\right| \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \frac{\left(\eta - \frac{i}{n}\right)^2}{\delta^2} \\ \leq & \sum_{i \notin N} \left(|f(\eta)| + \left| f\left(\frac{i}{n}\right)\right| \right) \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \frac{\left(\eta - \frac{i}{n}\right)^2}{\delta^2} \end{aligned}$$

Let M be such that $|f(x)| < M$, so

$$\sum_{i \notin N}^{n} \left| f(\eta) - f\left(\frac{i}{n}\right) \right| \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \leq 2M\delta^{-2} \sum_{i=1}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) \eta^i (1-\eta)^{n-i} \left(\eta - \frac{i}{n}\right)^2$$

Somehow we get this down to

$$|f(x) - (B_n f)(x)| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

4.4 Best Approximation

Let V be a Banach space and let $T \subset V$, an approximation p is a best approximation if for a given f,

$$\|f - p\| \ \leq \|f - \hat{p}\| \ \forall \, \hat{p} \in T$$

In other words, if

$$\|f - p\| \ = E_T(f) = \inf_{\hat{p} \in T} \|f - \hat{p}\|$$

- A best approximation exists if T is compact

Since finite dimensional linear spaces are compact, at least one best approximation is guaranteed to exist. This amounts to finding the coefficients $\{\alpha_i\}$ to minimize the distance

$$d(\alpha_1, ..., \alpha_n) = \|f - (\alpha_1 v_1 + ... + \alpha_n v_n)\|$$

4.4.1 pre-Hilbert Case

In the case that V is a pre-Hilbert space and $f \in U$, $p \in U$ the best approximation from U if and only if

$$\langle f - p, v \rangle = 0 \ \forall \ v \in U$$

This can be used to compute the best approximation using the normal equations. Pick $\{g_k\}$ such that $\operatorname{span}\{g_k\} = U$.

$$\left\langle f - \sum_{j=1}^{n} g_j, g_k \right\rangle = 0 \implies \sum_{j=1}^{n} \alpha_j \langle g_j, g_k \rangle = \langle f, g_k \rangle \ \forall \ k$$

If $\{g_k\}$ is orthonormal,

$$\alpha_k = \langle f, g_k \rangle \ \forall \ k$$

- Zero Theorem: If the set of polynomials $\{\psi_1, ..., \psi_n\}$ forms an orthonormal set on $[a, b]$, with respect to a weight function $\omega(x)$, then each of these polynomials has only simple real zeros, all of which lie in (a, b)

4.5 Least Squares

Given an inner product $\langle \cdot, \cdot \rangle$ and a set of functions $\{\phi_i\}_{i=0,\ldots,n}$, we see the best approximation of $f(x)$ in the norm given by this inner product. That is, we seek to choose $\{c_i\}$ such that for any other choice of sets $\{\hat{c}_i\}$,

$$\left\| f(x) - \sum_{i=0}^{n} c_i \phi_i(x) \right\|_{\mathcal{H}} \leq \left\| f(x) - \sum_{i=0}^{n} \hat{c}_i \phi_i(x) \right\|_{\mathcal{H}}$$

This is achieved by solving the system

$$\begin{pmatrix} \langle \phi_0, \phi_0 \rangle & \langle \phi_0, \phi_1 \rangle & \ldots & \langle \phi_0, \phi_n \rangle \\ \langle \phi_1, \phi_0 \rangle & \langle \phi_1, \phi_1 \rangle & & \\ \vdots & & \ddots & \\ \langle \phi_n, \phi_0 \rangle & \langle \phi_n, \phi_1 \rangle & & \langle \phi_n, \phi_n \rangle \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} \langle f, \phi_0 \rangle \\ \langle f, \phi_1 \rangle \\ \vdots \\ \langle f, \phi_n \rangle \end{pmatrix}$$

In the discrete case, we instead consider the system of equations

$$\begin{pmatrix} \phi_0(x_0) & \phi_1(x_0) & \ldots & \phi_n(x_0) \\ \phi_0(x_1) & \phi_1(x_1) & & \\ \vdots & & \ddots & \\ \phi_0(x_m) & \phi_1(x_m) & & \phi_n(x_m) \end{pmatrix}_{m \times n} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix}_{n \times 1} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_m) \end{pmatrix}_{m \times 1}$$

Which has the form $A\vec{\mathbf{c}} = \vec{\mathbf{f}}$. This has the solution

$$A^* A \vec{\mathbf{c}} = A^* \vec{\mathbf{f}} \implies \vec{\mathbf{c}} = (A^* A)^{-1} A^* \vec{\mathbf{f}}$$

5 Numerical Integration

5.1 Connection to Interpolation

Let $\{p_k(x)\}_{k=0}^n$ be a set of orthogonal polynomials on $[a, b]$ with respect to the inner product

$$\langle f, g \rangle = \int_a^b f(x) g(x) \omega(x) dx$$

The quadrature rule given by

$$I(f) = \int_a^b f(x) \omega(x) dx \approx Q(f) = \int_a^b \omega(x) \left(\sum_{j=1}^{n} f(x_j) L_j(x) \right) dx = \sum_{j=1}^{n} w_j f(x_j)$$

$$w_j = \int_a^b L_j(x) \omega(x) dx \qquad L_j(x) = \prod_{k=1, k \neq j}^{n} \frac{x - x_k}{x_j - x_k}$$

is exact for polynomials of degree at most $n - 1$.
If x_j, $j = 1, ..., n$ are the zeros of the highest degree polynomial $p_n(x)$, then the quadrature rule given above has exact degree $2n - 1$. This is because

$$f(x) = p_n(x) q(x) + r(x) \qquad q(x), r(x) \in \mathbb{P}_{n-1}(x)$$

So then

$$I(f) = I(p_n q + r) = \int_a^b q(x) p_n(x) \omega(x) dx + \int_a^b r(x) \omega(x) dx = \int_a^b r(x) \omega(x) dx = I(r) = Q(r)$$

since our formula is exact for $r(x)$. So then since $p_n(x_j) = 0$,

$$Q(r) = \sum_{j=1}^{n} w_j r(x_j) = \sum_{j=1}^{n} w_j (p_n(x_j) q(x_j) + r(x_j)) = Q(f)$$

Thus this method works for polynomials of degree $2n-1$. If we pick $f \in \mathbb{P}_{2n}$, then $q(x) \in \mathbb{P}_n$ and $r(x) \in \mathbb{P}_{n-1}$, so $\langle q, p_n \rangle \neq 0$, so

$$I(f) - Q(f) = I(qp_n + r) - Q(p_n q + r) = Q(r) + \langle q, p_n \rangle - Q(r) = \langle q, p_n \rangle \neq 0$$

5.2 Newton Cotes Formulas

The Newton Cotes formulas assume $n+1$ equidistant sampling points where n is the degree of the method. The closed formulas include the endpoints:

$$x_i = a + \frac{i}{n}(b-a) \qquad f_i = f(x_i) \qquad i = 0, ..., n$$

while the open formulas use

$$x_i = a + \frac{i}{n}(b-a) \qquad f_i = f(x_i) \qquad i = 1, ..., n-1$$

The methods are derived via the Lagrange Basis polynomials.

$$\int_a^b f(x)dx \approx I = \int_a^b L(x)dx = \int_a^b \left(\sum_{i=0}^{n} f(x_i) l_i(x) \right) dx = \sum_{i=0}^{n} f(x_i) \int_a^b l_i(x) dx$$

or for open rules

$$\int_a^b f(x)dx \approx I = \sum_{i=1}^{n-1} f(x_i) \int_a^b l_i(x) dx$$

These formulas can suffer from instability for large n unless they are converted into a composite rule.

5.2.1 Closed Schemes

- Trapezoid Rule:
  A degree 1 closed scheme is

  $$I = \frac{b-a}{2}(f_0 + f_1) \qquad h = b-a$$

  With error term

  $$E = -\frac{(b-a)^3}{12} f^{(2)}(\xi)$$

- Simpson's Rule:
  A degree 2 closed scheme is

  $$I = \frac{b-a}{6}(f_0 + 4f_1 + f_2) \qquad h = \frac{(b-a)}{2}$$

  With error term

  $$E = -\frac{(b-a)^5}{2880} f^{(4)}(\xi)$$

- Simpson's $\frac{3}{8}$ Rule:
  A degree 3 scheme is
$$I = \frac{b-a}{8}\left(f_0 + 3f_1 + 3f_2 + f_3\right) \qquad h = \frac{(b-a)}{3}$$
  With error term
$$E = -\frac{(b-a)^5}{6480} f^{(4)}(\xi)$$
- Boole's Rule :
  A degree 4 scheme is
$$I = \frac{b-a}{90}\left(7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4\right) \qquad h = \frac{(b-a)}{4}$$
  With error term
$$E = -\frac{(b-a)^7}{1935360} f^{(6)}(\xi)$$

5.2.2 Open Schemes

- Midpoint Rule:
  A degree 2 open scheme is
$$I = b-a\left(f_1\right) \qquad h = \frac{(b-a)}{2}$$
  With error term
$$E = \frac{(b-a)^3}{24} f^{(2)}(\xi)$$
- Trapezoid Method:
  A degree 3 open scheme is
$$I = \frac{b-a}{2}\left(f_1 + f_2\right) \qquad h = \frac{(b-a)}{3}$$
  With error term
$$E = \frac{(b-a)^3}{36} f^{(2)}(\xi)$$
- Milne's Rule:
  A degree 4 open scheme is
$$I = \frac{b-a}{4}\left(2f_1 - f_2 + 2f_3\right) \qquad h = \frac{(b-a)}{4}$$
  With error term
$$E = \frac{7(b-a)^5}{23040} f^{(4)}(\xi)$$

6 Iterative Schemes for $Ax = b$

Given an interative scheme where $A = L + D + R = M + N$
$$M\vec{\mathbf{x}}^{(k+1)} = N\vec{\mathbf{x}}^{(k)} + b$$

Thus if we have the amplification matrix $G = M^{-1}N$,
$$\vec{\mathbf{e}}^{(k+1)} = G\vec{\mathbf{e}}^{(k)}$$

By induction, $\vec{\mathbf{e}}^{(k+1)} = G^k \vec{\mathbf{e}}^{(0)}$. Perform the spectral decomposition of G,

$$\vec{\mathbf{e}}^{(k)} = R\Gamma^K R^{-1}\vec{\mathbf{e}}^{(0)}$$

If $\Gamma = \mathrm{diag}(\gamma_1, ..., \gamma_n)$ with $|\gamma_1| \geq ... \geq |\gamma_m|$ then we have convergence if $|\gamma_1| < 1.$ since then $G^K \to 0$ as $k \to \infty$.

6.1 Jacobi Iteration

$$M = D \qquad N = L + R = D - A \implies G = I - D^{-1}A$$

or alternatively,

$$Du_{k+1} = -(L+R)u_k + b$$

This method converges when $\rho\left(D^{-1}(L+R)\right) < 1$

- This method is guaranteed to converge when is strictly or irreducibly diagonally dominant. That is, when $|a_{ii}| > \sum\limits_{j=1, j\neq i}^{n} |a_{ij}|$

6.2 Gauss Seidel

$$M = L + D \qquad N = -R \implies G = -(L+D)^{-1}R$$

or alternatively,

$$(L+D)u_{k+1} = -Ru_k + b$$

This method converges when $\rho\left((L+D)^{-1}R\right) < 1$

- This method is guaranteed to converge when is strictly or irreducibly diagonally dominant.
- This method is guaranteed to converge when A is SPD

6.3 Successive Over-Relaxation (SOR)

Let ω be a parameter, and then

$$A = D + L + R \qquad M = \frac{1}{\omega}(D + \omega L) \qquad N = \frac{1}{\omega}((1-\omega)D - \omega R)$$

so then

$$G = \omega(D+\omega L)^{-1}\frac{1}{\omega}((1-\omega)D - \omega R) = (D+\omega L)^{-1}((1-\omega)D - \omega R)$$

That is,

$$(D + \omega L)u_{k+1} = ((1-\omega)D - \omega R)u_k + f$$

- If $A \in \mathbb{R}^{n\times n}$ is SPD, and $D + \omega L$ is nonsingular, then SOR converges for all $0 < \omega < 2$
- If $\omega = 1$, this method reduces to the Gauss Seidel method.

7 Matrix Decompositions

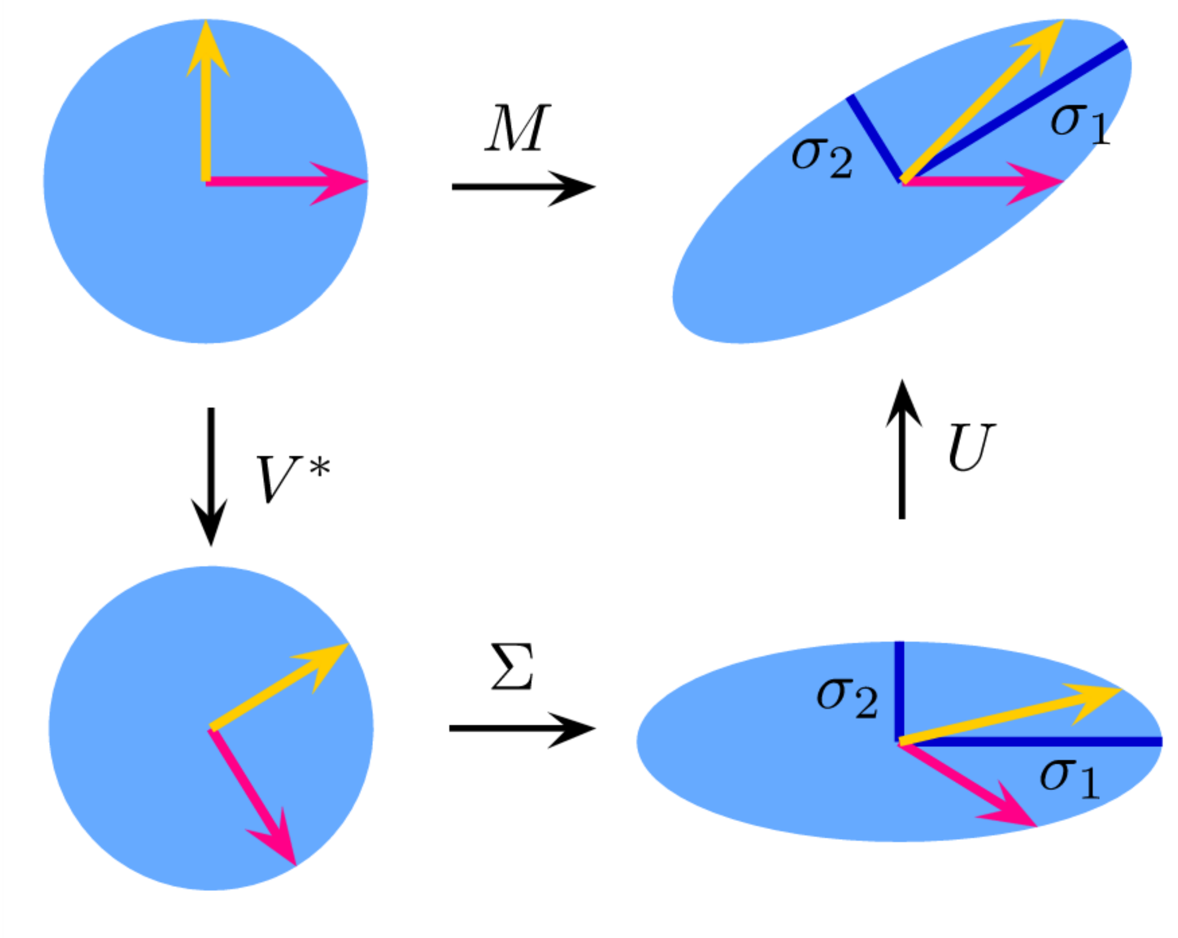
7.1 Singular Value Decomposition

The SVD of a $m \times n$ matrix is a factorization of the form

$$M = U\Sigma V^*$$

where

- U is an $m \times m$ unitary matrix made of the left-singular vectors of M, the eigenvectors of MM^*
- Σ is an $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, corresponding to the square roots of the non-zero eigenvalues of both M^*M and MM^*
- V is an $n \times n$ unitary matrix made of the right-singular vectors of M, the eigenvectors of M^*M



8 Nonlinear Solvers

8.1 Newton's Method

$$x_{k+1} = x_k - ((Jf)(x))^{-1} f(x_k)$$

8.2 Secant Method

$$x_{k+1} = x_k - f\left(x_k\right) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}$$

Letting $x_{k+1} = x^* + e_{k+1}$ for the solution x^*,

$$e_{k+1} = e_k - f\left(x^* + e_k\right) \frac{e_k - e_{k-1}}{f(x^* + e_k) - f(x^* + e_{k-1})}$$

We see that

$$e_{k+1} = \frac{f(x^* + e_k)e_k - f(x^* + e_{k-1})e_k - f\left(x^* + e_k\right)\left(e_k - e_{k-1}\right)}{f(x^* + e_k) - f(x^* + e_{k-1})}$$

Which simplifies to

$$e_{k+1} = \frac{f\left(x^* + e_k\right)e_{k-1} - f(x^* + e_{k-1})e_k}{f(x^* + e_k) - f(x^* + e_{k-1})}$$

We see that since $f(x^*) = 0$,

$$\begin{aligned} f(x^* + e_k) =& f'(x^*)e_k + \frac{1}{2}f''(x^*)(e_k)^2 + \frac{1}{6}f'''(\eta)(e_k)^3 \\ f(x^* + e_{k-1}) =& f'(x^*)e_{k-1} + \frac{1}{2}f''(x^*)(e_{k-1})^2 + \frac{1}{6}f'''(\eta)(e_{k-1})^3 \end{aligned}$$

So then

$$f(x^* + e_k) - f(x^* + e_{k-1}) = (e_k - e_{k-1})\left(f'(x^*) + \frac{1}{2}f''(x^*)(e_k + e_{k-1}) + \frac{1}{6}f'''(\eta)q(e_k, e_{k-1})\right)$$

And

$$\begin{aligned} f(x^* + e_k)e_{k-1} - f(x^* + e_{k-1})e_k =& \frac{1}{2}f''(x^*)(e_k^2 e_{k-1} - e_k e_{k-1}^2) + \frac{1}{6}f'''(\eta)(e_k^3 e_{k-1} - e_k e_{k-1}^3) \\ =& e_k e_{k-1}(e_k - e_{k-1})\left(\frac{1}{2}f''(x^*) + \frac{1}{6}f'''(\eta)(e_k + e_{k-1})\right) \end{aligned}$$

So

$$e_{k+1} = e_k e_{k-1} \frac{\left(\frac{1}{2}f''(x^*) + \frac{1}{6}f'''(\eta)(e_k + e_{k-1})\right)}{\left(f'(x^*) + \frac{1}{2}f''(x^*)(e_k + e_{k-1}) + \frac{1}{6}f'''(\eta)q(e_k, e_{k-1})\right)}$$

assuming $f''(x^*) \neq 0$

$$|e_{k+1}| \leq |e_k|\,|e_{k-1}| \left| \frac{\left(\frac{1}{2}f''(x^*) + \frac{1}{6}f'''(\eta)(e_k + e_{k-1})\right)}{\left(f'(x^*) + \frac{1}{2}f''(x^*)(e_k + e_{k-1}) + \frac{1}{6}f'''(\eta)q(e_k, e_{k-1})\right)} \right|$$

So if we assume small initial errors,

$$|e_{k+1}| \leq M\,|e_k|\,|e_{k-1}| \qquad M = \frac{1}{2}\frac{|f''(x^*)|}{|f'(x^*)|}$$

So assume $|e_{k+1}| \approx C\,|e_k|^p$, and $|e_k| \approx D\,|e_{k-1}|^p$. Then $|e_{k+1}| \approx CD^p\,|e_{k-1}|^{p^2}$. So then

$$CD^p\,|e_{k-1}|^{p^2} \approx D\,|e_{k-1}|^p\,|e_{k-1}|$$

So we want to solve

$$p^2 - p - 1 = 0 \implies p = \frac{1 \pm \sqrt{1+4}}{2} = \frac{1 \pm \sqrt{5}}{2}$$

So our method is of order $\frac{1+\sqrt{5}}{2}$.

9 ODE Solvers

We consider an ODE of the form

$$u'(t) = f(t,u)$$

9.1 Single Step Methods

Single step methods have the form

$$y_{i+1} = y_i + h\Phi(t,y;h), h = \delta t > 0$$

Given a solution $u(t)$, we define the truncation error of the method to be the difference between the exact and approximate solutions over a step of size h:

$$T(t,y;h) = \frac{1}{h}\left(y_{i+1} - u(t+h)\right)$$

or alternatively,

$$T(t,y;h) = \Phi(t,y;h) - \frac{1}{h}\left(u(t+h) - u(t)\right)$$

- A method is consistent if $T(t,y;h) \to 0$ uniformly as $h \to 0$. That is if

  $$\lim_{h\to 0} \|T(t,y;h)\|_\infty \to 0$$

  We have consistency if and only if $T(t,y;0) = f(t,y)$.

- The method is accurate of order p if for some vector norm, we have

  $$\|T(t,y;h)\| \ \leq Ch^p$$

  for some C. That is, as $h \to 0$,

  $$T(t,y;h) = \mathcal{O}(h^p)$$

  $p > 0$ implies consistency, and usually $p \geq 1$.

- The method is zero stable if there exists a constant C such that for any two sequences y_k and $\hat{y}_k$ with $y_0 \neq \hat{y}_0$,

  $$|y_k - \hat{y}_k| \leq C \max_{0\leq i\leq k-1} \{|y_i - \hat{y}_i|\} \text{ as } h \to 0$$

  Sufficient condition: A method is zero stable for any ode $y' = f(t,y)$ for LIschitz f if and only if the spectral radius of the iteration is bounded by 1 with equality holding only for simple eigenvalues.

9.1.1 Single Step Methods

Euler's method has the form

$$y_{i+1} = y_i + hf(t_i, y_i)$$

And is first order unless $f_t + f_y f = 0$

One can use Taylor expansion methods to improve the order of accuracy. We calculate the total derivatives of f,

$$\begin{aligned} f^{(0)} &= f(t,y) \\ f^{(1)} &= f_t(t,y) + f_y(t,y)f(t,y) \\ \vdots &= \vdots \\ f^{(k+1)} &= f_t^{(k)}(t,y) + f_y^{(k)}(t,y)f^{(k)}(t,y) \end{aligned}$$

Then using $u^{(k+1)}(t) = f^{(k)}(t, u(t))$, we can form an approximation of order p via

$$y_{i+1} = y_i + h\left(f^{(0)}(t,y) + \frac{1}{2}hf^{(1)}(t,y) + ... + \frac{1}{p!}h^{p-1}f^{(p-1)}(t,y)\right)$$

It can be shown that for this method,

$$\|T(t,y,h)\| \ \leq \frac{C_p}{(p+1)!}h^p$$

Again this may perform better if $f^{(p)} = 0$.

9.2 Runge-Kutta Methods

We could use multiple stages in our calculation, using

$$\begin{aligned} k_1(t,y) &= f(t,y) \\ k_2(t,y) &= f(t+\mu h, y + \mu h k_1) \\ y_{i+1} &= y_i + h\left(\alpha_1 k_1 + \alpha_2 k_2\right) \end{aligned}$$

In general, two stages offer at best an order 2 method, achieved by using the family

$$\alpha_1 + \alpha_2 = 1 \qquad \alpha_2 \mu = \frac{1}{2} \qquad \alpha_2 \neq 0 \text{ (arbitrary)}$$

To extend the idea of the two-stage methods to r stage methods, we use

$$\begin{aligned} k_1(t,y) &= f(t,y)i) \\ k_s(t,y;h) &= f\left(t + \mu_s h, y + h\sum_{j=1}^{s-1}\lambda_{sj}k_j\right) \\ y_{i+1} &= y_i + h\left(\sum_{s=1}^{r}\alpha_s k_s\right) \end{aligned}$$

For consistency we require $\sum_{s=1}^{r} \alpha_s = 1$, and it is natural to impose

$$\mu_s = \sum_{j=1}^{s-1}\lambda_{sj} \qquad s = 2, 3, ..., r$$

These methods can be summarize in the form of a Butcher Tableau as seen here:

$$\begin{array}{r|cccc}
\mu_1 = 0 & & & & \\
\mu_2 & \lambda_{21} & & & \\
\mu_3 & \lambda_{31} & \lambda_{32} & & \\
\vdots & \vdots & & \ddots & \\
\mu_s & \lambda_{s1} & \lambda_{s2} & \dots & \lambda_{s,s-1} \\
\hline
 & \alpha_1 & \alpha_2 & \dots & \alpha_s
\end{array}$$

So if $f(u) = Cu$ is linear,

$$\begin{pmatrix} 1 & & \\ \vdots & \ddots & \\ -\lambda_{s1} & \dots & 1 \end{pmatrix} \begin{pmatrix} k_1 \\ \vdots \\ k_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} Cy_i$$

and so

$$y_{i+1} = y_i + h \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_s \end{pmatrix}^T \begin{pmatrix} 1 & & \\ \vdots & \ddots & \\ -\lambda_{s1} & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} Cy_i$$

$$y_{i+1} = \left(1 + Ch \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_s \end{pmatrix}^T \begin{pmatrix} 1 & & \\ \vdots & \ddots & \\ -\lambda_{s1} & \dots & 1 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right) y_i$$

9.2.1 Stability

Runge Kutta Methods are absolute stable if the spectral radius of the amplification factor is strictly bounded by 1. That is, for linear scalar problems $u'(t) = Cu(t)$, if

$$|1 + ChG| < 1$$

This leads to a region of Ch for which the method is stable.

$$-1 < 1 + ChG < 1 \implies -2G^{-1} < Ch < 0$$

9.3 0

Linear Multistep Methods

We could develop a recurrence - like method in the form

$$\sum_{j=0}^{k} \alpha_j y_{n+j} = h \sum_{j=0}^{k} \beta_j f(t_{n+j}, y_{n+j})$$

With real $\{\alpha_i\}, \{\beta_i\}$, $\alpha_k \neq 0$ and α_0, β_0 are not both zero. If $\beta_k = 0$, then this method is explicit. Otherwise it is implicit. This is linear since it involves linear combinations of $\{y_i\}$.

- Implicit Euler Method

$$y_{n+1} = y_n + \frac{1}{2} h \left(f_{n+1} \right)$$

- The Implicit Trapezium method
$$y_{n+1} = y_n + \frac{1}{2}h\left(f_{n+1} + f_n\right)$$
- The Explicit 4-step Adams-Bashforth Method
$$y_{n+4} = y_{n+3} + \frac{1}{24}h\left(55f_{n+3} - 59_{n+2} + 37f_{n+1} - 9f_{n+3}\right)$$
- For a linear multi-step method that is consistent with the ODE, zero stability is necessary and sufficient for convergence.

9.4 Absolute Stability

Apply a given mehtod

A linear multi-step method is absolutely stable for a given Ch if and only if for that Ch, all of the roots of the stability polynomial $p(z, Ch)$ satisfy $|r_i| < 1$, $i = 1, .., k$.

- Absolutely stable methods are also zero stable.

9.5 Implicit Methods

We can also consider implicit Runge-Kutta methods

$$\begin{aligned} k_s &= f\left(t + \mu_s h, y + h\sum_{j=1}^{r}\lambda_{sj}k_j(t, y; h)\right) \qquad s = 1, 2, ..., r \\ y_{i+1} &= y_i + h\left(\sum_{s=1}^{r}\alpha_s k_s\right) \end{aligned}$$

For consistency we require $\sum\limits_{s=1}^{r}\alpha_s = 1$, and it is natural to impose

$$\mu_s = \sum_{j=1}^{s-1}\lambda_{sj} \qquad s = 2, 3, ..., r$$

The maximum order attainable by any Runge Kutta method is equal to the number of stages used for methods of less than order 4.