Numerical Analysis

Shing Tak Lam

May 11, 2022

Contents

1 Polynomial interpolation

Definition 1.1 (Fundamental Lagrange polynomial)
Suppose $x_0, \dots, x_n \in [a, b]$ distinct, $i \in \{0, \dots, n\}$, then the i-th fundamental Lagrange polynomial is

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_i}{x_j - x_i}$$

Proposition 1.2.

$$\ell_i(x_j) = \delta_{ij}$$

.

Definition 1.3 (Nodal polynomial)
Suppose $x_0, \dots, x_n \in [a, b]$ distinct, $i \in \{0, \dots, n\}$, then the nodal polynomial is

$$\omega(x) = \prod_{i=0}^{n}(x - x_i)$$

Proposition 1.4.

$$\ell_i(x) = \frac{\omega(x)}{\omega'(x_i)(x - x_i)}$$

Theorem 1.5. Suppose $f : [a, b] \to \mathbb{R}$, $x_0, \ldots, x_n \in [a, b]$ distinct. Then there exists unique $p \in \mathcal{P}_n$ such that $p(x_i) = f(x_i)$ for all i.

Proof. Let

$$p(x) = \sum_{i=0}^{n} f(x_i)\ell_i(x)$$

Then this satisfies the property required. On the other hand, if p and q are both polynomials which satisfy the required property, then $p - q$ has degree at most n and $n + 1$ roots, so must be identically zero. □

Definition 1.6 (Divided difference)

Suppose $f : [a, b] \to \mathbb{R}$, $x_0, \ldots, x_k \in [a, b]$ distinct. Then the divided difference $f[x_0, \ldots, x_k]$ is the leading coefficient of the polynomial $p_k \in \mathcal{P}_k$ which interpolates f at those points.

Theorem 1.7 (Newton formula)**.** Suppose $f : [a, b] \to \mathbb{R}$, $x_0, \ldots, x_n \in [a, b]$ distinct, $p_n \in \mathcal{P}_n$ interpolates f at those points. Then it can be written in Newton form

$$p_n(x) = \sum_{k=0}^{n} f[x_0, \ldots, x_k] \prod_{i=0}^{k-1}(x - x_i)$$

Proof. By induction. $n = 0$ is trivial. Note that p_n and p_{n+1} agree on $x_0, \ldots, x_n$ and has degree (at most) $n + 1$, so we have that

$$p_{n+1}(x) - p_n(x) = A_{n+1} \prod_{i=0}^{n}(x - x_i)$$

Suffices to show $A_{n+1} = f[x_0, \ldots, x_{n+1}]$. By considering the degree x^{n+1} term on the left and right hand sides, and using uniqueness we get the required result. □

Theorem 1.8 (Recurrence relation for divided differences)**.** Suppose $x_0, \ldots, x_k \in [a, b]$ distinct, with $k \geq 1$, we have that

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}$$

Proof. Let $q_0, q_1 \in \mathcal{P}_{k-1}$ be polynomials that interpolate f at $x_0, \ldots, x_{k-1}$ and $x_1, \ldots, x_k$ respectively. Then let

$$p(x) = \frac{x - x_0}{x_k - x_0} q_1(x) + \frac{x_k - x}{x_k - x_0} q_0(x)$$

Then p interpolates f at $x_0, \ldots, x_k$, and computing the leading coefficients on both sides we get the required result. □

Definition 1.9 (Horner form)
For a polynomial $p(x) = a_n x^n + \cdots + a_0$, the Horner form of the polynomial is

$$a_0 + x(a_1 + (a_2 + x(a_3 + \cdots + x(a_{n-1} + xa_n))))$$

1.1 Error bounds

Definition 1.10 (Interpolation error)
Suppose $f : [a, b] \to \mathbb{R}$, $p_n \in \mathcal{P}_n$ interpolates f at $x_0, \ldots, x_n \in [a, b]$ distinct, the interpolation error is

$$e_n(x) = f(x) - p_n(x)$$

Theorem 1.11. Suppose $p_n \in \mathcal{P}_n$ interpolates f at $x_0, \ldots, x_n$. Then for any $x \notin (x_i)$, we have that

$$e_n(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x]\omega(x)$$

Proof. Suppose p_{n+1} interpolates f at $x_0, \ldots, x_n, x_{n+1} = x$. Then noting that $p_{n+1}(x) = f(x)$ in the Newton form gives the required result. □

Lemma 1.12. Suppose $g \in C^k[a, b]$ has $k + \ell$ distinct zeroes. Then $g^{(k)}$ has at least ℓ distinct zeroes in $[a, b]$.

Proof. By Rolle and induction. □

Theorem 1.13. Suppose $x_0, \ldots, x_k \in [a, b]$ distinct, and $a = \min_i x_i$, $b = \max_i x_i$, $f \in \mathbb{C}^k[a, b]$. Then there exists $\xi \in (a, b)$ such that

$$f[x_0, \ldots, x_k] = \frac{1}{k!} f^{(k)}(\xi)$$

Proof. Suppose $p \in \mathcal{P}_k$ interpolates f at $x_0, \ldots, x_k$. Then $e = f - p$ has at least $k + 1$ distinct zeroes in $[a, b]$, so by Rolle's theorem, $f^{(k)} - p^{(k)}$ must have a root $\xi \in (a, b)$. But $p^{(k)} \equiv k! f[x_0, \ldots, x_k]$. □

Theorem 1.14. Suppose $f \in C^{n+1}[a, b]$, and $p_n \in \mathcal{P}_n$ interpolates f at $x_0, \ldots, x_n \in [a, b]$ distinct. Then for every $x \in [a, b]$, there exists $\xi \in [a, b]$ such that

$$e_n(x) = f(x) - p_n(x) = \frac{1}{(n+1)!}\omega(x) f^{(n+1)}(\xi)$$

Proof. If $x = x_i$ for some i, then both sides are zero, and we are done. Otherwise,

$$e_n(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x]\omega(x) = \frac{1}{(n+1)!}\omega(x) f^{(n+1)}(\xi)$$

from the previous theorems. □

Corollary 1.15. For all x, we have that

$$|e_n(x)| = |f(x) - p_n(x)| \leq \frac{1}{(n+1)!}|\omega(x)|\left\|f^{(n+1)}\right\|_\infty$$

Corollary 1.16. For *any* set Δ of $n+1$ interpolation points, p_Δ interpolating polynomial for f in Δ, we have that

$$\left\|e_\Delta\right\|_\infty = \left\|f - p_\Delta\right\|_\infty \leq \frac{1}{(n+1)!}\left\|\omega_\Delta\right\|_\infty \left\|f^{(n+1)}\right\|_\infty$$

1.2 Chebyshev polynomials

Definition 1.17 (Chebyshev polynomial)
The Chebyshev polynomial of degree n on $[-1, 1]$ is defined by

$$T_n(x) = \cos(n \arccos(x))$$

Proposition 1.18. T_n has maximum absolute value 1, and alternating signs.

Proposition 1.19. T_n has n distinct zeroes at

$$x_k = \cos\left(\frac{2k-1}{2n}\pi\right) \quad \text{for} \quad k = 1, \dots, n$$

Lemma 1.20. The Chebyshev polynomials satisfies the recurrence relation

$$\begin{aligned} T_0(x) &\equiv 1 \\ T_1(x) &\equiv x \\ T_{n+1}(x) &\equiv 2xT_n(x) - T_{n-1}(x) \end{aligned}$$

Proof. Substitute $x = \cos(\theta)$ into $\cos((n+1)\theta) - \cos((n-1)\theta) = 2\cos(\theta)\cos(n\theta)$. □

Corollary 1.21. T_n has degree n, and leading coefficient 2^{n-1}.

Theorem 1.22. Let $\gamma_n = 2^{-(n-1)}$. Then among all monic polynomials with degree n, $\gamma_n T_n$ has the smallest L^∞ norm over $[-1, 1]$. That is,

$$\inf_{p \in \mathcal{P}_n \text{ monic}} \left\|p\right\|_\infty = \gamma_n \left\|T_n\right\|_\infty$$

Proof. Suppose $q \in \mathcal{P}_n$ monic, with $\left\|q\right\|_\infty < \gamma_n$. Consider $r = \gamma T_n - q$. Then $r \in \mathcal{P}_{n-1}$. Furthermore, at $t_k = \cos\left(\frac{\pi k}{n}\right)$, $k = 0, \dots n$, $\gamma_n T_n(t_k) = (-1)^k \gamma_n$. Since $\left\|q\right\|_\infty < \gamma_n$, we must have that $\text{sign}(r(t_k)) = \text{sign}(\gamma_n(T_k)) = (-1)^k$. But this means that r has at least n zeroes in $[-1, 1]$. Contradiction as $r \in \mathcal{P}_{n-1}$. □

Corollary 1.23. For a set of n interpolating points Δ, we have that

$$\frac{1}{2^n} \leq \|\omega_\Delta\|_\infty$$

Theorem 1.24. For $f \in C^{n+1}[-1,1]$, the best choice of approximation points is

$$\Delta = \left\{ \cos\left(\frac{2k+1}{2n+2}\pi\right) : k = 0, \ldots, n \right\}$$

which achieves the above bound, and we have that

$$\|e_\Delta\|_\infty = \|f - p_\Delta\|_\infty \leq \frac{1}{2^n(n+1)!}\|f^{(n+1)}\|_\infty$$

1.3 Orthogonal polynomials

Definition 1.25 (Inner product)
Let $w \in C[a,b]$, $w > 0$. Then we have an inner product on $C[a,b]$ defined by

$$\langle f, g \rangle = \langle f, g \rangle_w = \int_a^b f(x)g(x)w(x)\mathrm{d}x$$

Definition 1.26 (n-th orthogonal polynomial)
$Q_n \in \mathcal{P}_n$ is an n-the degree orthogonal polynomial if for all $p \in \mathcal{P}_{n-1}$, $\langle Q_n, p \rangle = 0$.

Lemma 1.27. There exists a unique orthonormal basis $Q_0, Q_1, Q_2, \ldots$ of monic polynomials such that $\deg(Q_n) = n$.

Proof. Existence follows by applying Gram-Schmidt to $1, x, x^2, \ldots$. For uniqueness, suppose we have Q_n and $\tilde{Q}_n$. Then we note that

$$\left\langle Q_n - \tilde{Q}_n, Q_n - \tilde{Q}_n \right\rangle = \left\langle Q_n, Q_n - \tilde{Q}_n \right\rangle - \left\langle \tilde{Q}_n, Q_n - \tilde{Q}_n \right\rangle = 0$$

Since $Q_n - \tilde{Q}_n$ has degree $n-1$. So $Q_n = \tilde{Q}_n$. □

Theorem 1.28 (Three term recurrence)**.** Monic orthogonal polynomials satisfy the relation

$$Q_{n+1}(x) = (x - a_n)Q_n(x) - b_n Q_{n-1}(x)$$

where $Q_{-1}(x) = 0$, $Q_0(x) = 1$ and

$$a_n = \frac{\langle xQ_n, Q_n \rangle}{\|Q_n\|^2} \quad \text{and} \quad b_n = \frac{\|Q_n\|^2}{\|Q_{n-1}\|^2}$$

Proof. Since the Q_i form an orthonormal basis, we have that

$$xQ_n(x) = \sum_{k=0}^{n+1} c_k Q_k(x) \quad \text{where} \quad c_k = \frac{\langle xQ_n, Q_k \rangle}{\|Q_k\|^2} = \frac{\langle Q_n, xQ_k \rangle}{\|Q_k\|^2}$$

Then we have the follwoing cases.

- $k = n+1$ gives $c_{n+1} = 1$.
- $k = n$ gives $c_n = a_n$ by definition.
- $k = n-1$ gives us that $\langle Q_n, xQ_{n-1}\rangle = \langle Q_n, Q_n + (xQ_{n-1} - Q_n)\rangle = \langle Q_n, Q_n\rangle$ as $xQ_{n-1} - Q_n \in \mathcal{P}_{n-1}$.
- $k \leq n-2$ has $xQ_k \in \mathcal{P}_{n-1}$, so $\langle Q_n, xQ_k\rangle = 0$.

This then gives us that $xQ_n(x) = Q_{n+1}(x) + a_nQ_n(x) + b_nQ_{n-1}(x)$. □

Proposition 1.29. Suppose Q_{n+1} is orthogonal to all $p_n \in \mathcal{P}_n$ on $[a, b]$. Then all of the zeroes of Q_{n+1} are distinct and lie within the interval (a, b).

Proof. Let k be the number of sign changes of Q_{n+1} in (a, b). Suppose for contradiction $k \leq n$. If $k = 0$, set $p_k = 1$, otherwise, let $p_k(x) = \prod_{i=1}^{k}(x - t_i)$ where the t_i are where Q_{n+1} changes signs. Then $\langle Q_{n+1}, p_k\rangle = 0$, as $p_k \in \mathcal{P}_k \leq \mathcal{P}_n$. On the other hand, by construction p_kQ_{n+1} does not change sign on (a, b), so

$$|\langle Q_{n+1}, p_k\rangle| = \left|\int_a^b Q_{n+1}(x)p_k(x)w(x)\mathrm{d}x\right| = \int_a^b |Q_{n+1}(x)p_k(x)|w(x)\mathrm{d}x > 0$$

Contradiction. So $k \geq n+1$. □

1.4 Least squares polynomial fitting

Theorem 1.30 (Least squares polynomial)**.** Suppose $Q_0, \ldots, Q_n$ are an orthogonal basis for $\mathcal{P}_n$, $f \in C[a, v]$, the least squares approximant $p \in \mathcal{P}_n$ for f is given by

$$p = \sum_{k=0}^{n} c_kQ_k \quad \text{where} \quad c_k = \frac{\langle f, Q_k\rangle}{\|Q_k\|^2}$$

and the error is given by

$$\|f - p\|^2 = \|f\|^2 - \sum_{k=0}^{n} \frac{\langle f, Q_k\rangle^2}{\|Q_k\|^2} = \|f\|^2 - \|p\|^2$$

Proof. Since the Q_k form a basis, for $c = (c_0, \ldots, c_n)$, let $p_c \in \mathcal{P}_n$ where

$$p_c = \sum_{k=0}^{n} c_kQ_k$$

Then define the function $F : \mathbb{R}^{n+1} \to \mathbb{R}$ by

$$F(c) = \langle f - p_c, f - p_c\rangle = \left\langle f - \sum_{k=0}^{n} c_kQ_k, f - \sum_{k=0}^{n} c_kQ_k \right\rangle = \|f\|^2 - 2\sum_{k=0}^{n} c_k\langle f, Q_k\rangle + \sum_{k=0}^{n} c_k^2\|Q_k\|^2$$

This is a quadratic in each c_k, hence convex, so the minima is achieved when

$$\frac{\partial F(c)}{\partial c_k} = -2\langle f, Q_k\rangle + 2c_k\|Q_k\|^2 = 0$$

Substituting gives the required result. The expression for the error is given by this and orthogonality. □

Theorem 1.31 (Parseval)**.** Suppose we have a compact interval $[a, b]$ for which we are approximating in. Then

$$\sum_{k=0}^{\infty} \frac{\langle f, Q_k \rangle^2}{\|Q_k\|^2} = \|f\|^2$$

Proof. By the Weierstrass approximation theorem,

$$\lim_{n\to\infty} \inf_{p\in\mathcal{P}_n} \|f - p\|^2 \to 0$$

□

2 Approximation of linear functionals

Definition 2.1 (Linear functional)
Given a real vector space V, we call the elements of the dual space $V^* = \mathrm{Hom}(V, \mathbb{R})$ a linear functional.

Definition 2.2 (Interpolating formula)
Given a linear functional $\lambda : C^{n+1}[a, b] \to \mathbb{R}$, distinct interpolating points $x_0, \dots, x_n \in [a, b]$, we define the interpolating formula

$$\lambda(f) \approx \sum_{i=0}^{n} \lambda(\ell_i) f(x_i)$$

Definition 2.3 (Exact)
Given a linear functional $\lambda : C^{n+1}[a, b] \to \mathbb{R}$, points $x_0, \dots, x_n \in [a, b]$ distinct, the approximation

$$\lambda(f) \approx \sum_{i=0}^{n} a_i f(x_i)$$

is exact on $\mathcal{P}_n$ if for all $p \in \mathcal{P}_n$, the above is an equality.

Proposition 2.4. An approximation is exact on $\mathcal{P}_n$ if and only if it is interpolating.

Proof. By definition, an interpolating formula is exact. Conversely, considering the basis ℓ_i of $\mathcal{P}_n$, we get that $a_i = \lambda(\ell_i)$. □

2.1 Numerical integration

Definition 2.5 (Quadrature)
For a weight function $w > 0$, we have the quadrature

$$\lambda(f) = \int_a^b f(x) w(x) \mathrm{d}x \approx \sum_{i=0}^{n} a_i f(x_i)$$

with nodes (x_i) and weights (a_i).

Proposition 2.6. No quadrature rule with $n + 1$ nodes is exact on $\mathcal{P}_m$ for $m \geq 2n + 2$.

Proof. Let $p(x) = \prod(x - x_i)^2 \in \mathcal{P}_{2n+2}$. Then $\lambda(p) > 0$, but any quadrature will be zero. □

Theorem 2.7. Suppose a quadrature with nodes $x_0, \ldots, x_n$ is exact (i.e. interpolating) on $\mathcal{P}_n$. Then it is exact on $\mathcal{P}_{2n+1}$ if and only if $x_0, \ldots, x_n$ are the zeroes of the $(n+1)$-st orthogonal polynomial Q_{n+1}.

Proof. Suppose a quadrature with nodes $x_0, \ldots, x_n$ is exact for all $p \in \mathcal{P}_{2n+1}$, let $Q_{n+1}(x) = \prod(x - x_i) \in \mathcal{P}_{n+1}$, taking any $q_n \in \mathcal{P}_n$, we find that

$$\langle Q_{n+1}(x), q_n(x)\rangle = \int_a^b Q_{n+1}(x)q_n(x)w(x)\mathrm{d}x = \sum_{i=0}^{n} a_i Q_{n+1}(x_i)q_n(x_i) = 0$$

So Q_{n+1} is orthogonal to all $q_n \in \mathcal{P}_n$. On the other hand, suppose Q_{n+1} has zeroes at $x_0, \ldots, x_n$. Given any $p_{2n+1} \in \mathcal{P}_{2n+1}$, we have $q_n, r_n \in \mathcal{P}_n$ such that

$$p_{2n+1} = Q_{n+1}q_n + r_n$$

Since Q_{n+1} is orthogonal to q_n, we have that

$$I(p_{2n+1}) = \int_a^b p_{2n+1}(x)w(x)\mathrm{d}x = \int_a^b r_n(x)w(x)\mathrm{d}x = I(r_n)$$

On the other hand, since $Q_{n+1}(x_i) = 0$ for all i, we have that

$$\sum_{i=0}^{n} a_i p_{2n+1}(x_i) = \sum_{i=1}^{n} a_i s_n(x_i) = I(s_n)$$

since the approximation is exact on $\mathcal{P}_n$. □

Definition 2.8 (Gaussian quadrature)
A quadrature with $n+1$ nodes and is exact on $\mathcal{P}_{2n+1}$ is called Gaussian quadrature.

2.2 Approximation error

Definition 2.9 (Approximation error)
Given a linear functional λ, and an approximation formula

$$\lambda(f) \approx \sum_{i=0}^{n} a_i f(x_i)$$

define the approximation error

$$e_\lambda(f) = \lambda(f) - \sum_{i=0}^{n} a_i f(x_i)$$

Definition 2.10 (Peano kernel)
Let $g_t(x) = (x-t)_+^n = \begin{cases} (x-t)^n & x \geq t \\ 0 & x < t \end{cases}$. Then the Peano kernel for a linear functional λ is

$$K_\lambda(t) = \lambda(g_t)$$

Theorem 2.11 (Peano kernel theorem (General functional))**.** Suppose λ is a linear functional on $C^{n+1}[a,b]$ such that we can exchange λ and $\int_a^b$. Furthermore, suppose λ vanishes on $\mathcal{P}_n$. Then we have an integral representation

$$\lambda(f) = \frac{1}{n!}\int_a^b K_\lambda(t) f^{(n+1)}(t)\mathrm{d}t$$

Proof. Consider the Taylor series of $f \in C[a,b]$ with integral remainder

$$f(x) = \sum_{k=0}^{n} \frac{1}{n!}(x-a)^n f^{(n)}(a) + R(x) \quad \text{where} \quad R(x) = \frac{1}{n!}\int_a^x (x-t)^n f^{(n+1)}(t)\mathrm{d}t$$

Note that we can also write

$$q_n(x) = \sum_{k=0}^{n} \frac{1}{n!}(x-a)^n f^{(n)}(a) \quad \text{and} \quad R(x) = \frac{1}{n!}\int_a^b (x-t)_+^n f^{(n+1)}(t)\mathrm{d}t$$

Since λ vanishes on $\mathcal{P}_n$, $\lambda(q_n) = 0$. So interchanging λ and $\int_a^b$ we have

$$\lambda(f) = \lambda(R) = \frac{1}{n!}\int_a^b K_\lambda(t) f^{(n+1)}(t)\mathrm{d}t$$

□

Proposition 2.12. Let Λ_0 be the set of linear functionals on $C^{n+1}[a,b]$ spanned by

$$\mu(f) = f^{(k)}(x) \quad \text{for} \quad 0 \le k \le n, x \in [a,b]$$

and

$$\mu(f) = \int_a^x f(t)w(t)\mathrm{d}t \quad \text{for} \quad x \in [a,b]$$

Then for any $\lambda \in \Lambda_0$, we can exchange λ and $\int_a^b$.

Theorem 2.13. Suppose $\lambda \in \Lambda_0$, $\lambda(f) \approx \sum_{i=0}^{m} a_i f(x_i)$ is an approximation which is exact on $\mathcal{P}_n$. Then the error functional satisfies

$$|e_\lambda(f)| \le c_\lambda \left\| f^{(n+1)} \right\|_\infty \quad \text{where} \quad c_\lambda = \frac{1}{n!}\left\| K_{e_\lambda} \right\|_1$$

Furthermore, equality is achieved for some $f \in C^{n+1}[a,b]$.

Proof.

$$|e_\lambda| = \frac{1}{n!}\left| \int_a^b K_{e_\lambda}(t) f^{(n+1)}(t)\mathrm{d}t \right| \le \frac{1}{n!}\left\| K_{e_\lambda} \right\|_1 \left\| f^{(n+1)} \right\|_\infty$$

Equality holds if we take (a sequence of functions converging to) the function f_0 with $f_0^{(n+1)}(t) = \mathrm{sign}(K_{e_\lambda}(t))$.

□

3 Ordinary differential equations

3.1 Single step methods

Definition 3.1 (Single step method)
For a first order differential equation

$$y' = f(t, y) \quad 0 \leq t \leq T$$

and time step $t_n = nh$, a single step method is

$$y(t_{n+1}) \approx y_{n+1} = \phi(t_n, y_n)$$

That is, y_{n+1} depends only on t_n, h and y_n.

Definition 3.2 (Euler method)
The Euler method is

$$y_{n+1} = y_n + hf(t_n, y_n)$$

Definition 3.3 (Convergence)
Fix $T > 0$, and suppose for all $h > 0$, we have a sequence $y_n = y_{n,h}$ for $0 \leq n \leq \lfloor T/h \rfloor$. Then we say the method converges if

$$\max_n \left\| y_n - y(t_n) \right\| \to 0$$

as $h \to 0$.

Theorem 3.4. Suppose f is λ-Lipschitz in the second argument (as in the statement of Picard-Lindelöf), and y is C^2. Then there exists c_0 such that the error $e_n = y(t_n) - y_n$ satisfies $\left\| e_n \right\| \leq c_0 h$. In particular, the Euler method converges.

Proof. Expanding y about t_n we get that

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{1}{2}h^2 y''(\tau_n)$$

where $\tau_n \in (t_n, t_{n+1})$. Subtracting the Euler method from this, and defining $c = \frac{1}{2}\left\| y'' \right\|_\infty$, we get that

$$\left\| e_{n+1} \right\| \leq \left\| e_n \right\| + h\left\| f(t_n, y(t_n)) - f(t_n, y_n) \right\| + ch^2 \leq (1 + \lambda h)\left\| e_n \right\| + ch^2$$

Inductively, we have that

$$\left\| e_{n+m} \right\| \leq (1 + \lambda h)^m \left\| e_n \right\| + ch^2 \sum_{i=0}^{m-1} (1 + \lambda h)^i$$

Since $e_0 = 0$, setting $n = 0$ in the above, we get that

$$\left\| e_n \right\| \leq ch^2 \sum_{i=0}^{n-1} (1 + \lambda h)i = ch^2 \frac{(1 + \lambda h)^n - 1}{(1 + \lambda h) - 1} \leq \frac{ch}{\lambda}(1 + \lambda h)^n \leq \frac{ce^{\lambda T}}{\lambda} h$$

since $1 + \lambda h \leq e^{\lambda h}$ and $nh \leq T$. □

Definition 3.5 (Local truncation error)
The local truncation error of a numerical method $y_{n+1} = \phi_h(t_n, y_0, \ldots, y_n)$ is the error of the method on the true solution, that is,

$$\eta_{n+1} = y(t_{n+1}) - \phi_h(t_n, y(t_0), \dots, y(t_n))$$

Definition 3.6 (Order)
The order of a method is the largest integer $p \geq 0$ such that

$$\eta_{n+1} = \mathcal{O}\left(h^{p+1}\right)$$

for all $h > 0$, $n \geq 0$ and f sufficiently smooth.

Definition 3.7 (Theta methods)
For $\theta \in [0, 1]$, methods of the form

$$y_{n+1} = y_n + h\left(\theta f(t_n, y_n) + (1-\theta) f(t_{n+1}, y_{n+1})\right)$$

are called theta methods.

Definition 3.8 (Implicit)
A method is implicit if for each time step we need to solve a system of algebraic equations to find the solution. Otherwise, the method is called explicit.

Proposition 3.9. If $\theta < 1$, then the theta method is implicit. If $\theta = 1$, we recover the Euler method.

Remark 3.10. $\theta = 0$ is called the backwards Euler method, and $\theta = 1/2$ is called the trapezoidal rule.

Proposition 3.11. The local truncation error of the theta method is

$$\left(\theta - \frac{1}{2}\right) h^2 y''(t_n) + \left(\frac{1}{2}\theta - \frac{1}{3}\right) h^3 y'''(t_n) + \mathcal{O}\left(h^4\right)$$

Thus the theta method has order 1, except the trapezoidal rule has order 2.

3.2 Multistep methods

Definition 3.12 (Multistep method)
For $s \geq 1$, we say that

$$\sum_{m=0}^{s} a_m y_{n+m} = h \sum_{m=0}^{s} b_m f_{n+m}$$

where $a_s = 1$ and $f_{n+m} = f(t_{n+m}, y_{n+m})$ is an s-step method.

Proposition 3.13. The method is implicit if $b_s \neq 0$, and explicit if $b_s = 0$.

Theorem 3.14. A multistep method has order $p \geq 1$ if and only if

$$\sum_{m=0}^{s} a_m = 0 \quad \text{and} \quad \sum_{m=0}^{s} m^k a_m = k \sum_{m=0}^{s} m^{k-1} b_m \text{ for } k = 1, \dots, p$$

Proof. Substituting the exact solution and expanding into the the Taylor series about t_n, we have that

$$\begin{aligned} \sum_{m=0}^{s} a_m y(t_{n+m}) - h \sum_{m=0}^{s} b_m y'(t_{n+m}) &= \sum_{m=0}^{s} a_m \sum_{k=0}^{\infty} \frac{(mh)^k}{k!} y^{(k)}(t_n) - h \sum_{m=0}^{s} b_m \sum_{k=1}^{\infty} \frac{(mh)^{k-1}}{(k-1)!} y^{(k)}(t_n) \\ &= \left(\sum_{m=0}^{s} a_m \right) y(t_n) + \sum_{k=1}^{\infty} \frac{h^k}{k!} \left(\sum_{m=0}^{s} m^k a_m - k \sum_{m=0}^{s} m^{k-1} b_m \right) y^{(k)}(t_n) \end{aligned}$$

For the method to be order p, it is necessary and sufficient for the coefficients of the h^k to be zero for $k \leq p$. □

Definition 3.15 (Characteristic polynomials)
Given a s-step method, define the characteristic polynomials

$$\rho(w) = \sum_{m=0}^{s} a_m w^m \quad \text{and} \quad \sigma(w) = \sum_{m=0}^{s} b_m w^m$$

Theorem 3.16. The multistep method is order $p \geq 1$ if and only if

$$\rho(e^z) - z\sigma(e^z) = \mathcal{O}\left(z^{p+1}\right)$$

Proof. Expanding into Taylor series, we have that

$$\begin{aligned} \rho(e^z) - z\sigma(e^z) &= \sum_{m=0}^{s} a_m e^{mz} - z \sum_{m=0}^{s} b_m e^{mz} \\ &= \sum_{m=0}^{s} a_m \sum_{k=0}^{\infty} \frac{m^k z^k}{k!} - z \sum_{m=0}^{s} b_m \sum_{k=0}^{\infty} \frac{m^k z^k}{k!} \\ &= \left(\sum_{m=0}^{s} a_m \right) + \sum_{k=1}^{\infty} \frac{z^k}{k!} \left(\sum_{m=0}^{s} m^k a_m - k \sum_{m=0}^{s} m^{k-1} b_m \right) \end{aligned}$$

and the result follows by the previous theorem. □

Definition 3.17 (Convergence)
For the multistep method, define the errors of the initial steps and the method respectively:

$$\hat{e}(h) = \max_{0 \leq i < s} \left\| y(t_i) - y_i \right\| \quad \text{and} \quad e(h) = \max_{0 \leq i \leq N} \left\| y(t_i) - y_i \right\|$$

We say that a method is convergent if for every ODE $y' = f(t, y)$ where f is Lipschitz in the second argument, if $h \to 0$ and $\hat{e}(h) \to 0$, then $e(h) \to 0$.

Definition 3.18 (Root condition)
For a polynomial p, we say that p satisfies the root condition if all roots have modulus at most 1, and the roots with modulus 1 are simple.

Theorem 3.19 (Dahlquist equivalence). The multistep method is convergent if and only if it is order $p \geq 1$ and ρ satisfies the root condition.

Proposition 3.20. For an arbitrary degree s polynomial satisfying the root condition and has $\rho(1) = 0$, define

$$\sigma(z) = \frac{\rho(w)}{\log(w)} + \begin{cases} \mathcal{O}\left(|w-1|^{s+1}\right) & \text{implicit method} \\ \mathcal{O}\left(|w-1|^{s}\right) & \text{explicit method} \end{cases}$$

Then this defines a multistep method.

Definition 3.21 (Backwards differentiation formula)
A backwards differentiation formula is a s-step, order s multistep method with $\sigma(w) = w^s$. That is,

$$\sum_{m=0}^{s} a_m y_{n+m} = h f_{n+s}$$

Lemma 3.22. The characteristic polynomial ρ of a BDF has the form

$$\rho(w) = \sum_{k=1}^{s} \frac{1}{k} w^{s-k}(w-1)^k$$

Proof. Setting $w = e^z$, we need to show that

$$\rho(e^z) - z\sigma(e^z) = \mathcal{O}\left(z^{s+1}\right)$$

which becomes

$$\rho(w) - w^s \log(w) + \mathcal{O}\left(|w-1|^{s+1}\right)$$

expanding in Taylor series about 1 gives the required result. □

3.3 Runge-Kutta methods

Definition 3.23 (Explicit Runge-Kutta scheme)
An s-stage Runge-Kutta scheme is a method of the form

$$y_{n+1} = y_n + h \sum_{i=1}^{s} b_i k_i$$

where

$$k_i = f\left(t_n + c_i h, y_n + h\sum_{j=1}^{i-1} a_{ij}k_j\right)$$

Definition 3.24 (Runge-Kutta methods)
A general s-stage Runge-Kutta scheme is a method of the form

$$y_{n+1} = y_n + h\sum_{i=1}^{s} b_i k_i$$

where

$$k_i = f\left(t_n + c_i h, y_n + h\sum_{j=1}^{s} a_{ij}k_j\right)$$

an explicit method has $a_{ij} = 0$ for $i \leq j$.

3.4 Stiffness and stability

Definition 3.25 (Stiff ODE)
An ODE $y' = f(t, y)$ is stiff if (for some numerical methods) we need to reduce h for stability beyond the requirements for accuracy.

Definition 3.26 (Linear stability domain)
Suppose a numerical method with constant h, applied to the ODE $y' = \lambda y; y(0) = 1$ generates the sequence (y_n). We call the set

$$\mathcal{D} = \left\{z = \lambda h : \lim_{n\to\infty} y_n = 0\right\}$$

the linear stability domain of the method.

Definition 3.27 (A-stable)
A numerical method is A-stable if

$$\{z : \mathrm{Re}(z) < 0\} \subseteq \mathcal{D}$$

Proposition 3.28. The set of $\lambda \in \mathbb{C}$ such that $y(t) = e^{\lambda t} \to 0$ as $t \to \infty$ is $\{z : \mathrm{Re}(z) < 0\}$. Thus a numerical method is A-stable if and only if the numerical solution exhibits the same behaviour.

Remark 3.29. If a method is A-stable, then we can just set the step size to fit the accuracy requirements and we do not need to decrease it further for stability.

Proposition 3.30. For a multistep method with characteristic polynomials ρ, σ, $z = \lambda h$ is in the linear stability domain if and only if the roots of the characteristic equation

$$p(x) = \rho(x) - z\sigma(x) = \sum_{m=0}^{s} a_m x^m - z \sum_{m=0}^{s} b_m x^m = 0$$

are less than one in modulus.

Proof. $z = \lambda h \in \mathcal{D}$ if the sequence y_n which is the solution to the recurrence relation

$$\sum_{m=0}^{s} a_m y_{n+m} = \lambda h \sum_{m=0}^{s} b_m y_{n+m}$$

satisfies $y_n \to 0$. □

Proposition 3.31. ∂D can be parametrised by the curve $z(t) = \frac{\rho(e^{it})}{\sigma(e^{it})}$

Proof. If $z \in \partial D$, then the charactertistic equation has a root with modulus one, say e^{it}. Substituting and rearranging gives the required result. □

Theorem 3.32 (Second Dahlquist barrier). No multistep method of order $p \geq 3$ is A-stable.

Remark 3.33. The trapezoidal rule has $p = 2$ and is A-stable.

Definition 3.34 (A_0-stable)
A numerical method is A_0 stable if we have $\alpha > 0$ such that

$$\left\{-re^{i\theta} : \theta \in (-\alpha, \alpha)\right\} \subseteq \mathcal{D}$$

Theorem 3.35. All convergent BDF methods (i.e. order ≤ 6) are A_0-stable.

Proposition 3.36. No explicit Runge-Kutta method is A_0-stable. Hence there are no A-stable RK methods.

3.5 Implementation

Definition 3.37 (Milne device)
The Milne device consists of a pair of multistep methods of the same order, one explicit (predictor, P) and one implicit (corrector, C).

Proposition 3.38. Suppose the predictor has truncation error (say)

$$y(t_{n+1}) - y_{n+1}^P = c_P h^{p+1} y^{(p+1)}(t_n) + \mathcal{O}\left(h^{p+2}\right)$$

and the corrector has truncation error (say)

$$y(t_{n+1}) - y_{n+1}^C = c_C h^{p+1} y^{(p+1)}(t_n) + \mathcal{O}\left(h^{p+2}\right)$$

Then we have that

$$h^{p+1} y^{(p+1)}(t_n) \approx \frac{y_{n+1}^C - y_{n+1}^P}{c_C - c_P}$$

and

$$y(t_{n+1}) - y_{n+1}^C \approx \frac{c_C}{c_C - c_P} \left(y_{n+1}^C - y_{n+1}^P\right)$$

Definition 3.39 (Embedded RK)
An embedded RK contains a s-stage (explicit) RK method y_n and a $s+m$ stage (explicit) RK method $\tilde{y}_n$, where the first s stages of y_n and $\tilde{y}_n$ are the same. Then we have the error estimate

$$y(t_{n+1}) - y_{n+1} \approx \tilde{y}_{n+1} - y_{n+1}$$

4 Numerical linear algebra

4.1 Sparse and band matrices

Definition 4.1 (Sparse matrix)
A matrix A is sparse if nearly all elements are zero.

Definition 4.2 (Band matrix)
A matrix A is a band matrix with bandwidth r if $a_{ij} = 0$ for all $|i - j| > r$.

4.2 LU factorisation

Definition 4.3 (LU factorisation)
For a nonsingular matrix A, the LU factorisation of A is

$$A = LU$$

where L is lower triangular and has diagonal entries one, and U is upper triangular.

Proposition 4.4. Suppose $A = LU$, l_k is the k-th column of L, and u_k^T is the k-th row of U. Let $A = A^{(0)}$ and define

$$A^{(k)} = A^{(k-1)} - l_k u_k^\mathsf{T}$$

Then u_k^T is the k-th row of $A^{(k-1)}$ and $a_{kk}^{(k-1)} \cdot l_k$ is the k-th column of $A^{(k-1)}$.

Definition 4.5 (Column pivoting)
In each stage of the LU factorisation (i.e. suppose we already have $A^{(k-1)}$), exchange two rows of $A^{(k-1)}$ such that the element with the largest magnitude in the k-th column is at the (k, k) position. The result is

$$PA = LU \iff A = P^\mathsf{T} LU$$

where P is a permutation matrix (which is orthogonal).

Proposition 4.6. If column pivoting is used to obtain $A = P^{\mathsf{T}}LU$, then every element of L has modulus at most one.

Proof. Immediate from $a_{kk}^{(k-1)} \cdot l_k$ being the k-th column of $A^{(k-1)}$. □

Definition 4.7 (Strictly regular)
A square matrix A is strictly regular is the leading submatrices are all nonsingular.

Theorem 4.8. A has an LU factorisation if and only if it is strictly regular.

Theorem 4.9. The LU factorisation, if it exists, is unique.

Corollary 4.10. A strictly regular matrix A has a unique factorisation $A = LDU$ where L and U have unit diagonals, and D is diagonal.

Corollary 4.11. A strictly regular symmetric matrix has a unique factorisation $A = LDL^{\mathsf{T}}$.

Definition 4.12 (Symmetric positive definite)
A matrix A is SPD if it is symmetric and positive definite.

Theorem 4.13. Let $A \in \mathrm{Mat}_n(\mathbb{R})$ be symmetric, it is positive definite if and only if it has a LDL^{T} factorisation, where all of the diagonal elements of D are positive.

Proof. Suppose such a factorisation exists. Then it is clear that A is SPD. On the other hand, suppose A is positive definite. Then A is strictly regular, so has a LDL^{T} factorisation, and clearly the diagonal elements are all positive. □

Definition 4.14 (Cholesky factorisation)
A SPD matrix has a factorisation

$$A = \tilde{L}\tilde{L}^{\mathsf{T}}$$

where $\tilde{L}$ is a lower triangular matrix.

Definition 4.15 (Strictly diagonally dominant)
A matrix A is strictly diagonally dominant by rows if for all i,

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}|$$

Theorem 4.16. If A is strictly regular by rows, then it is strictly regular.

Theorem 4.17. Suppose $A = LU$. Then all leading zeroes in the rows of A to the left of the diagonal are inherited by L. Similarly, all leading zeroes in the columns of A above the diagonal are inherited by U.

Corollary 4.18. If A is a band matrix with bandwidth r, then so are L and U.

4.3 QR factorisation

Definition 4.19 (QR factorisation)
The QR factorisation of a $m \times n$ matrix A is $A = QR$, where Q $m \times m$ orthogonal, and R $m \times n$ upper triangular.

Theorem 4.20. Every matrix A has a QR factorisation. If A is square and nonsingular, then a factorisation $A = QR$ where the diagonal entries of R are positive is unique.

Proof. For existence we will consider three different algorithms in this section. For uniqueness, let $A = QR$ be nonsingular. Then $A^\top A = R^\top R$ is SPD, so has a unique Cholesky decompositiion $A = \tilde{L}\tilde{L}^\top$, with $\tilde{L}$ having a positive main diagonal. So $R^\top = \tilde{L}$ is unique. □

Proposition 4.21. Suppose A square nonsingular. Then by running the Gram-Schmidt algorithm on the columns of A we obtain a QR factorisation.

Definition 4.22 (Givens rotations)
Given p, q, a, b, define the Givens rotation

$$\Omega_{a,b}^{[p,q]} = \begin{pmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & c & & s & & \\ & & & \ddots & & & \\ & & -s & & c & & \\ & & & & & \ddots & \\ & & & & & & 1 \end{pmatrix}$$

where $c = \frac{a}{\sqrt{a^2+b^2}}$ and $s = \frac{b}{\sqrt{a^2+b^2}}$.

Proposition 4.23.

$$\Omega_{a,b}^{[p,q]} \begin{pmatrix} x_1 \\ \vdots \\ a \\ \vdots \\ b \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ \sqrt{a^2+b^2} \\ \vdots \\ 0 \\ \vdots \\ x_n \end{pmatrix}$$

Proposition 4.24. Suppose A is an $m \times n$ matrix, $1 \le p \le q \le m$, $\tilde{A} = \Omega_{a,b}^{[p,q]} A$, where $a = a_{pp}$ and $b = a_{qp}$. Then $\tilde{a}_{qp} = 0$. Furthermore, all other rows are not changed.

Theorem 4.25. For any matrix A, there exists a sequence of Givens rotations such that

$$R = \left(\Omega^{[m-1,m]}\right) \cdots \left(\Omega^{[2,m]} \cdots \Omega^{[2,3]}\right) \left(\Omega^{[1,m]} \cdots \Omega^{[1,2]}\right) A$$

is upper triangular.

Definition 4.26 ((Householder) Reflection)
Given a nonzero vector $u \in \mathbb{R}^n$, reflection in u has matrix

$$H_u = I - \frac{2}{\|u\|^2} u u^{\mathsf{T}}$$

Proposition 4.27. For any vectors $a, b \in \mathbb{R}^n$, with $\|a\| = \left\|b\right\|$, let $u = a - b$. Then $H_u a = b$.

Corollary 4.28. For any nonzero vector a, $u = a \mp \|a\| e_i$ has $H_u a = \mp \|a\| e_i$.

Remark 4.29. We prefer $-$ for calculations by hand, $+$ for numerical computations for stability reasons.

Theorem 4.30. For any matrix A, there exists a sequence of Householder reflections such that

$$R = H_{n-1} \cdots H_2 H_1 A$$

is an upper triangular matrix.

Proof. By recursion. H_1 mapping the first column to $\|a\| e_1$ means $H_1 A$ has as the first column $\|a\| e_1$.

Suppose the first $k-1$ columns of $C = H_{k-1} \cdots H_1 A$ are upper triangular. Let c be the k-th column of C. Let $\gamma^2 = \sum_{i=k}^{m} c_i^2$, $u = c - \gamma e_k$. Then the last $m-k$ entries of $H_u c$ are zero. □

4.4 Least squares

Definition 4.31 (Ordinary least squares)
Given $A \in \mathbb{R}^{m \times n}$, $y \in \mathbb{R}^m$, we wish to find

$$c^* = \arg\min_{c \in \mathbb{R}^n} \lVert Ac - y \rVert$$

Theorem 4.32. $c^* \in \mathbb{R}^n$ is a solution to the OLS problem if and only if $A^\mathsf{T}(Ac^* - y) = 0$.

Proof. If c^* is a solution, then it minimises the quadratic form

$$F(x) = \langle Ac - y, Ac - y \rangle = c^\mathsf{T} A^\mathsf{T} Ac - 2c^\mathsf{T} A^\mathsf{T} y + y^\mathsf{T} y$$

Then $\nabla F = 2A^\mathsf{T}Ac - A^\mathsf{T}y = 0$ at $c = c^*$.
Conversely, if $A^\mathsf{T}(Ac^* - y) = 0$. Let $c = c^* + d$, and consider the quadratic form

$$\begin{aligned} G(d) &= \lVert Ac - y \rVert^2 \\ &= \langle Ac - y, Ac - y \rangle \\ &= \langle Ad + (Ac^* - y), Ad + (Ac^* - y) \rangle \\ &= \lVert Ad \rVert^2 + 2 \left\langle A^\mathsf{T}(Ac^* - y), d \right\rangle + \lVert Ac^* - y \rVert^2 \\ &= \lVert Ad \rVert^2 + \lVert Ac^* - y \rVert^2 \end{aligned}$$

Then d minimises G if and only if $G(d) \in \ker(A)$. In particular, $d = 0$, so $c = c^*$ is a minimiser of F. □

Corollary 4.33. If $\ker(A) = 0$, then the minimiser is unique.

Proof. From proof of the above theorem. □

Corollary 4.34. c^* is optimal if and only if $Ac^* - y$ is orthogonal to all the columns of A, or equivalently, Ac^* is the projection of y onto the image of A.

Definition 4.35 (Normal equations)
The normal equations are

$$A^\mathsf{T} Ac^* = A^\mathsf{T} y$$

where $A^\mathsf{T}A$ is known as the Gram matrix, and c^* is the normal solution.

Proposition 4.36. If A has linearly independent columns, and thus $\ker(A) = 0$, then $A^\mathsf{T}A$ is invertible, and the solution is given by

$$c^* = (A^\mathsf{T}A)^{-1}A^\mathsf{T}y$$

where $(A^\mathsf{T}A)^{-1}A^\mathsf{T}$ is the Penrose-Moore pseudoinverse of A.

Proposition 4.37. Suppose $A = QR$ where Q orthogonal and R upper triangular. Then

$$\|Ac - y\| = \|Q^\mathsf{T}(Ac - y)\| = \|Rc - Q^\mathsf{T} y\|$$

Proposition 4.38. Suppose $A = QR$, where $\operatorname{rank}(R) = \operatorname{rank}(A) = n$. Then the bottom $m - n$ rows of R are zero, and a solution can be found by considering the first n equations of

$$Rc = Q^\mathsf{T} y$$

which is nonsingular.