Part IB — Numerical Analysis

Theorems with proof

Based on lectures by G. Moore

Notes taken by Dexter Chua

Lent 2016

Polynomial approximation
Interpolation by polynomials. Divided differences of functions and relations to derivatives. Orthogonal polynomials and their recurrence relations. Least squares approximation by polynomials. Gaussian quadrature formulae. Peano kernel theorem and applications. [6]

Computation of ordinary differential equations
Euler's method and proof of convergence. Multistep methods, including order, the root condition and the concept of convergence. Runge-Kutta schemes. Stiff equations and A-stability. [5]

Systems of equations and least squares calculations
LU triangular factorization of matrices. Relation to Gaussian elimination. Column pivoting. Factorizations of symmetric and band matrices. The Newton-Raphson method for systems of non-linear algebraic equations. QR factorization of rectangular matrices by Gram-Schmidt, Givens and Householder techniques. Application to linear least squares calculations. [5]

Contents

0 Introduction

1 Polynomial interpolation

1.1 The interpolation problem

1.2 The Lagrange formula

Theorem. The interpolation problem has exactly one solution.

Proof. We define $p \in P_n[x]$ by

$$p(x) = \sum_{k=0}^{n} f_k \ell_k(x).$$

Evaluating at x_i gives

$$p(x_j) = \sum_{k=0}^{n} f_k \ell_k(x_j) = \sum_{k=0}^{n} f_k \delta_{jk} = f_j.$$

So we get existence.

For uniqueness, suppose $p, q \in P_n[x]$ are solutions. Then the difference $r = p - q \in P_n[x]$ satisfies $r(x_j) = 0$ for all j, i.e. it has $n + 1$ roots. However, a non-zero polynomial of degree n can have at most n roots. So in fact $p - q$ is zero, i.e. $p = q$. □

1.3 The Newton formula

Theorem (Recurrence relation for Newton divided differences)**.** For $0 \leq j < k \leq n$, we have

$$f[x_j, \cdots, x_k] = \frac{f[x_{j+1}, \cdots, x_k] - f[x_j, \cdots, x_{k-1}]}{x_k - x_j}.$$

Proof. The key to proving this is to relate the interpolating polynomials. Let $q_0, q_1 \in P_{k-j-1}[x]$ and $q_2 \in P_{k-j}$ satisfy

$$\begin{aligned} q_0(x_i) &= f_i && i = j, \cdots, k-1 \\ q_1(x_i) &= f_i && i = j+1, \cdots, k \\ q_2(x_i) &= f_i && i = j, \cdots, k \end{aligned}$$

We now claim that

$$q_2(x) = \frac{x - x_j}{x_k - x_j} q_1(x) + \frac{x_k - x}{x_k - x_j} q_0(x).$$

We can check directly that the expression on the right correctly interpolates the points x_i for $i = j, \cdots, k$. By uniqueness, the two expressions agree. Since $f[x_j, \cdots, x_k]$, $f[x_{j+1}, \cdots, x_k]$ and $f[x_j, \cdots, x_{k-1}]$ are the leading coefficients of q_2, q_1, q_0 respectively, the result follows. □

1.4 A useful property of divided differences

Lemma. Let $g \in C^m[a, b]$ have a continuous mth derivative. Suppose g is zero at $m + \ell$ distinct points. Then $g^{(m)}$ has at least ℓ distinct zeros in $[a, b]$.

Proof. This is a repeated application of Rolle's theorem. We know that between every two zeros of g, there is at least one zero of $g' \in C^{m-1}[a, b]$. So by differentiating once, we have lost at most 1 zeros. So after differentiating m times, $g^{(m)}$ has lost at most m zeros. So it still has at least ℓ zeros. □

Theorem. Let $\{x_i\}_{i=0}^n \in [a, b]$ and $f \in C^n[a, b]$. Then there exists some $\xi \in (a, b)$ such that

$$f[x_0, \cdots, x_n] = \frac{1}{n!} f^{(n)}(\xi).$$

Proof. Consider $e = f - p_n \in C^n[a, b]$. This has at least $n + 1$ distinct zeros in $[a, b]$. So by the lemma, $e^{(n)} = f^{(n)} - p_n^{(n)}$ must vanish at some $\xi \in (a, b)$. But then $p_n^{(n)} = n! f[x_0, \cdots, x_n]$ constantly. So the result follows. □

1.5 Error bounds for polynomial interpolation

Theorem. Assume $\{x_i\}_{i=0}^n \subseteq [a, b]$ and $f \in C[a, b]$. Let $\bar{x} \in [a, b]$ be a non-interpolation point. Then

$$e_n(\bar{x}) = f[x_0, x_1, \cdots, x_n, \bar{x}]\omega(\bar{x}),$$

where

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

Proof. We think of $\bar{x} = x_{n+1}$ as a new interpolation point so that

$$p_{n+1}(x) - p_n(x) = f[x_0, \cdots, x_n, \bar{x}]\omega(x)$$

for all $x \in R$. In particular, putting $x = \bar{x}$, we have $p_{n+1}(\bar{x}) = f(\bar{x})$, and we get the result. □

Theorem. If in addition $f \in C^{n+1}[a, b]$, then for each $x \in [a, b]$, we can find $\xi_x \in (a, b)$ such that

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

Proof. The statement is trivial if x is an interpolation point — pick arbitrary ξ_x, and both sides are zero. Otherwise, this follows directly from the last two theorems. □

Corollary. For all $x \in [a, b]$, we have

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

Lemma (3-term recurrence relation)**.** The Chebyshev polynomials satisfy the recurrence relations

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

with initial conditions

$$T_0(x) = 1, \quad T_1(x) = x.$$

Proof.

$$\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\cos(n\theta). \qquad \square$$

Theorem (Minimal property for $n \geq 1$)**.** On $[-1, 1]$, among all polynomials $p \in P_n[x]$ with leading coefficient 1, $\frac{1}{2^{n-1}}\|T_n\|$ minimizes $\|p\|_\infty$. Thus, the minimum value is $\frac{1}{2^{n-1}}$.

Proof. We proceed by contradiction. Suppose there is a polynomial $q_n \in P_n$ with leading coefficient 1 such that $\|q_n\|_\infty < \frac{1}{2^{n-1}}$. Define a new polynomial

$$r = \frac{1}{2^{n-1}}T_n - q_n.$$

This is, by assumption, non-zero.

Since both the polynomials have leading coefficient 1, the difference must have degree at most $n-1$, i.e. $r \in P_{n-1}[x]$. Since $\frac{1}{2^{n-1}}T_n(X_k) = \pm\frac{1}{2^{n-1}}$, and $|q_n(X_n)| < \frac{1}{2^{n-1}}$ by assumption, r alternates in sign between these $n+1$ points. But then by the intermediate value theorem, r has to have at least n zeros. This is a contradiction, since r has degree $n-1$, and cannot be zero. $\square$

Corollary. Consider

$$w_\Delta = \prod_{i=0}^{n}(x - x_i) \in P_{n+1}[x]$$

for any distinct points $\Delta = \{x_i\}_{i=0}^n \subseteq [-1, 1]$. Then

$$\min_\Delta \|\omega_\Delta\|_\infty = \frac{1}{2^n}.$$

This minimum is achieved by picking the interpolation points to be the zeros of T_{n+1}, namely

$$x_k = \cos\left(\frac{2k+1}{2n+2}\pi\right), \quad k = 0, \cdots, n.$$

Theorem. For $f \in C^{n+1}[-1, 1]$, the Chebyshev choice of interpolation points gives

$$\|f - p_n\|_\infty \leq \frac{1}{2^n}\frac{1}{(n+1)!}\|f^{(n+1)}\|_\infty.$$

2 Orthogonal polynomials

2.1 Scalar product

2.2 Orthogonal polynomials

Theorem. Given a vector space V of functions and an inner product $\langle \cdot, \cdot \rangle$, there exists a unique monic orthogonal polynomial for each degree $n \geq 0$. In addition, $\{p_k\}_{k=0}^n$ form a basis for $P_n[x]$.

Proof. This is a big induction proof over both parts of the theorem. We induct over n. For the base case, we pick $p_0(x) = 1$, which is the only degree-zero monic polynomial.

Now suppose we already have $\{p_n\}_{k=0}^n$ satisfying the induction hypothesis.

Now pick any monic $q_{n+1} \in P_{n+1}[x]$, e.g. x^{n+1}. We now construct p_{n+1} from q_{n+1} by the Gram-Schmidt process. We define

$$p_{n+1} = q_{n+1} - \sum_{k=0}^{n} \frac{\langle q_{n+1}, p_k \rangle}{\langle p_k, p_k \rangle} p_k.$$

This is again monic since q_{n+1} is, and we have

$$\langle p_{n+1}, p_m \rangle = 0$$

for all $m \leq n$, and hence $\langle p_{n+1}, p \rangle = 0$ for all $p \in P_n[x] = \langle p_0, \cdots, p_n \rangle$.

To obtain uniqueness, assume both $p_{n+1}, \hat{p}_{n+1} \in P_{n+1}[x]$ are both monic orthogonal polynomials. Then $r = p_{n+1} - \hat{p}_{n+1} \in P_n[x]$. So

$$\langle r, r \rangle = \langle r, p_{n+1} - \hat{p}_{n+1} \rangle = \langle r, p_{n+1} \rangle - \langle r, \hat{p}_{n+1} \rangle = 0 - 0 = 0.$$

So $r = 0$. So $p_{n+1} = \hat{p}_{n-1}$.

Finally, we have to show that $p_0, \cdots, p_{n+1}$ form a basis for $P_{n+1}[x]$. Now note that every $p \in P_{n+1}[x]$ can be written uniquely as

$$p = cp_{n+1} + q,$$

where $q \in P_n[x]$. But $\{p_k\}_{k=0}^n$ is a basis for $P_n[x]$. So q can be uniquely decomposed as a linear combination of $p_0, \cdots, p_n$.

Alternatively, this follows from the fact that any set of orthogonal vectors must be linearly independent, and since there are $n + 2$ of these vectors and $P_{n+1}[x]$ has dimension $n + 2$, they must be a basis. □

2.3 Three-term recurrence relation

Theorem. Monic orthogonal polynomials are generated by

$$p_{k+1}(x) = (x - \alpha_k)p_k(x) - \beta_k p_{k-1}(x)$$

with initial conditions

$$p_0 = 1, \quad p_1(x) = (x - \alpha_0)p_0,$$

where

$$\alpha_k = \frac{\langle xp_k, p_k \rangle}{\langle p_k, p_k \rangle}, \quad \beta_k = \frac{\langle p_k, p_k \rangle}{\langle p_{k-1}, p_{k-1} \rangle}.$$

Proof. By inspection, the p_1 given is monic and satisfies

$$\langle p_1, p_0\rangle = 0.$$

Using $q_{n+1} = xp_n$ in the Gram-Schmidt process gives

$$p_{n+1} = xp_n - \sum_{k=0}^{n} \frac{\langle xp_n, p_k\rangle}{\langle p_k, p_k\rangle} p_k$$

$$p_{n+1} = xp_n - \sum_{k=0}^{n} \frac{\langle p_n, xp_k\rangle}{\langle p_k, p_k\rangle} p_k$$

We notice that $\langle p_n, xp_k\rangle$ and vanishes whenever xp_k has degree less than n. So we are left with

$$\begin{aligned} &= xp_n - \frac{\langle xp_n, p_n\rangle}{\langle p_n, p_n\rangle} p_n - \frac{\langle p_n, xp_{n-1}\rangle}{\langle p_{n-1}, p_{n-1}\rangle} p_{n-1} \\ &= (x - \alpha_n)p_n - \frac{\langle p_n, xp_{n-1}\rangle}{\langle p_{n-1}, p_{n-1}\rangle} p_{n-1}. \end{aligned}$$

Now we notice that xp_{n-1} is a monic polynomial of degree n so we can write this as $xp_{n-1} = p_n + q$. Thus

$$\langle p_n, xp_{n-1}\rangle = \langle p_n, p_n + q\rangle = \langle p_n, p_n\rangle.$$

Hence the coefficient of p_{n-1} is indeed the β we defined. □

2.4 Examples

2.5 Least-squares polynomial approximation

Theorem. If $\{p_n\}_{k=0}^n$ are orthogonal polynomials with respect to $\langle\,\cdot\,,\,\cdot\,\rangle$, then the choice of c_k such that

$$p = \sum_{k=0}^{n} c_k p_k$$

minimizes $\|f - p\|^2$ is given by

$$c_k = \frac{\langle f, p_k\rangle}{\|p_k\|^2},$$

and the formula for the error is

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

Proof. We consider a general polynomial

$$p = \sum_{k=0}^{n} c_k p_k.$$

We substitute this in to obtain

$$\langle f - p, f - p\rangle = \langle f, f\rangle - 2\sum_{k=0}^{n} c_k \langle f, p_k\rangle + \sum_{k=0}^{n} c_k^2 \|p_k\|^2.$$

Note that there are no cross terms between the different coefficients. We minimize this quadratic by setting the partial derivatives to zero:

$$0 = \frac{\partial}{\partial c_k}\langle f - p, f - p\rangle = -2\langle f, p_k\rangle + 2c_k\|p_k\|^2.$$

To check this is indeed a minimum, note that the Hessian matrix is simply $2I$, which is positive definite. So this is really a minimum. So we get the formula for the c_k's as claimed, and putting the formula for c_k gives the error formula. □

3 Approximation of linear functionals

3.1 Linear functionals

3.2 Gaussian quadrature

Proposition. There is no choice of ν weights and nodes such that the approximation of $\int_a^b w(x)f(x)\,\mathrm{d}x$ is exact for all $f \in P_{2\nu}[x]$.

Proof. Define

$$q(x) = \prod_{k=1}^{\nu}(x - c_k) \in P_\nu[x].$$

Then we know

$$\int_a^b w(x)q^2(x)\,\mathrm{d}x > 0,$$

since q^2 is always non-negative and has finitely many zeros. However,

$$\sum_{k=1}^{\nu} b_k q^2(c_n) = 0.$$

So this cannot be exact for q^2. □

Theorem (Ordinary quadrature)**.** For any distinct $\{c_k\}_{k=1}^{\nu} \subseteq [a, b]$, let $\{\ell_k\}_{k=1}^{\nu}$ be the Lagrange cardinal polynomials with respect to $\{c_k\}_{k=1}^{\nu}$. Then by choosing

$$b_k = \int_a^b w(x)\ell_k(x)\,\mathrm{d}x,$$

the approximation

$$L(f) = \int_a^b w(x)f(x)\,\mathrm{d}x \approx \sum_{k=1}^{\nu} b_k f(c_k)$$

is exact for $f \in P_{\nu-1}[x]$.

We call this method ordinary quadrature.

Theorem. For $\nu \geq 1$, the zeros of the orthogonal polynomial p_ν are real, distinct and lie in (a, b).

Proof. First we show there is at least one root. Notice that $p_0 = 1$. Thus for $\nu \geq 1$, by orthogonality, we know

$$\int_a^b w(x)p_\nu(x)p_1(x)\,\mathrm{d}x = \int_a^b w(x)p_\nu(x)\,\mathrm{d}x = 0.$$

So there is at least one sign change in (a, b). We have already got the result we need for $\nu = 1$, since we only need one zero in (a, b).

Now for $\nu > 1$, suppose $\{\xi_j\}_{j=1}^{m}$ are the places where the sign of p_ν changes in (a, b) (which is a subset of the roots of p_ν). We define

$$q(x) = \prod_{j=1}^{m}(x - \xi_j) \in P_m[x].$$

Since this changes sign at the same place as p_ν, we know qp_ν maintains the same sign in (a, b). Now if we had $m < \nu$, then orthogonality gives

$$\langle q, p_\nu\rangle = \int_a^b w(x)q(x)p_\nu(x)\,\mathrm{d}x = 0,$$

which is impossible, since qp_ν does not change sign. Hence we must have $m = \nu$. □

Theorem. In the ordinary quadrature, if we pick $\{c_k\}_{k=1}^\nu$ to be the roots of $p_\nu(x)$, then get we exactness for $f \in P_{2\nu-1}[x]$. In addition, $\{b_n\}_{k=1}^\nu$ are all positive.

Proof. Let $f \in P_{2\nu-1}[x]$. Then by polynomial division, we get

$$f = qp_\nu + r,$$

where q, r are polynomials of degree at most $\nu - 1$. We apply orthogonality to get

$$\int_a^b w(x)f(x)\,\mathrm{d}x = \int_a^b w(x)(q(x)p_\nu(x) + r(x))\,\mathrm{d}x = \int_a^b w(x)r(x)\,\mathrm{d}x.$$

Also, since each c_k is a root of p_ν, we get

$$\sum_{k=1}^\nu b_k f(c_k) = \sum_{k=1}^\nu b_k(q(c_k)p_\nu(c_k) + r(c_k)) = \sum_{k=1}^\nu b_k r(c_k).$$

But r has degree at most $\nu - 1$, and this formula is exact for polynomials in $P_{\nu-1}[x]$. Hence we know

$$\int_a^b w(x)f(x)\,\mathrm{d}x = \int_a^b w(x)r(x)\,\mathrm{d}x = \sum_{k=1}^\nu b_k r(c_k) = \sum_{k=1}^\nu b_k f(c_k).$$

To show the weights are positive, we simply pick as special f. Consider $f \in \{\ell_k^2\}_{k=1}^\nu \subseteq P_{2\nu-2}[x]$, for ℓ_k the Lagrange cardinal polynomials for $\{c_k\}_{k=1}^\nu$. Since the quadrature is exact for these, we get

$$0 < \int_a^b w(x)\ell_k^2(x)\,\mathrm{d}x = \sum_{j=1}^\nu b_j\ell_k^2(c_j) = \sum_{j=1}^\nu b_j\delta_{jk} = b_k.$$

Since this is true for all $k = 1, \cdots, \nu$, we get the desired result. □

4 Expressing errors in terms of derivatives

Theorem (Peano kernel theorem)**.** If λ annihilates polynomials of degree k or less, then

$$\lambda(f) = \frac{1}{k!} \int_a^b K(\theta) f^{(k+1)}(\theta) \, \mathrm{d}\theta$$

for all $f \in C^{k+1}[a, b]$, where

5 Ordinary differential equations

5.1 Introduction

5.2 One-step methods

Theorem (Convergence of Euler's method)**.**

(i) For all $t \in [0, T]$, we have

$$\lim_{\substack{h\to 0\\ nh\to t}} \mathbf{y}_n - \mathbf{y}(t) = 0.$$

(ii) Let λ be the Lipschitz constant of f. Then there exists a $c \geq 0$ such that

$$\|\mathbf{e}_n\| \leq ch\frac{e^{\lambda T} - 1}{\lambda}$$

for all $0 \leq n \leq [T/h]$, where $\mathbf{e}_n = \mathbf{y}_n - \mathbf{y}(t_n)$.

Proof. There are two parts to proving this. We first look at the *local truncation error*. This is the error we would get at each step assuming we got the previous steps right. More precisely, we write

$$\mathbf{y}(t_{n+1}) = \mathbf{y}(t_n) + h(\mathbf{f}, t_n, \mathbf{y}(t_n)) + \mathbf{R}_n,$$

and $\mathbf{R}_n$ is the local truncation error. For the Euler's method, it is easy to get $\mathbf{R}_n$, since $\mathbf{f}(t_n, \mathbf{y}(t_n)) = \mathbf{y}'(t_n)$, by definition. So this is just the Taylor series expansion of $\mathbf{y}$. We can write $\mathbf{R}_n$ as the integral remainder of the Taylor series,

$$\mathbf{R}_n = \int_{t_n}^{t_{n+1}} (t_{n+1} - \theta) y''(\theta)\, \mathrm{d}\theta.$$

By some careful analysis, we get

$$\|\mathbf{R}_n\|_\infty \leq ch^2,$$

where

$$c = \frac{1}{2}\|y''\|_\infty.$$

This is the easy part, and tends to go rather smoothly even for more complicated methods.

Once we have bounded the local truncation error, we patch them together to get the actual error. We can write

$$\begin{aligned}
\mathbf{e}_{n+1} &= \mathbf{y}_{n+1} - \mathbf{y}(t_{n+1})\\
&= \mathbf{y}_n + h\mathbf{f}(t_n, \mathbf{y}_n) - \Big(\mathbf{y}(t_n) + h\mathbf{f}(t_n, \mathbf{y}(t_n)) + \mathbf{R}_n\Big)\\
&= (\mathbf{y}_n - \mathbf{y}(t_n)) + h\Big(\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_n, \mathbf{y}(t_n))\Big) - \mathbf{R}_n
\end{aligned}$$

Taking the infinity norm, we get

$$\begin{aligned}
\|\mathbf{e}_{n+1}\|_\infty &\leq \|\mathbf{y}_n - \mathbf{y}(t_n)\|_\infty + h\|\mathbf{f}(t_n, \mathbf{y}_n) - \mathbf{f}(t_n, \mathbf{y}(t_n))\|_\infty + \|\mathbf{R}_n\|_\infty\\
&\leq \|\mathbf{e}_n\|_\infty + h\lambda\|\mathbf{e}_n\|_\infty + ch^2\\
&= (1 + \lambda h)\|\mathbf{e}_n\|_\infty + ch^2.
\end{aligned}$$

This is valid for all $n \geq 0$. We also know $\|\mathbf{e}_0\| = 0$. Doing some algebra, we get

$$\|\mathbf{e}_n\|_\infty \leq ch^2 \sum_{j=0}^{n-1} (1 + h\lambda)^j \leq \frac{ch}{\lambda}\left((1 + h\lambda)^n - 1\right).$$

Finally, we have

$$(1 + h\lambda) \leq e^{\lambda h},$$

since $1 + \lambda h$ is the first two terms of the Taylor series, and the other terms are positive. So

$$(1 + h\lambda)^n \leq e^{\lambda h n} \leq e^{\lambda T}.$$

So we obtain the bound

$$\|\mathbf{e}_n\|_\infty \leq ch\frac{e^{\lambda T} - 1}{\lambda}.$$

Then this tends to 0 as we take $h \to 0$. So the method converges. □

5.3 Multi-step methods

Theorem. An s-step method has order p ($p \geq 1$) if and only if

$$\sum_{\ell=0}^{s} \rho_\ell = 0$$

and

$$\sum_{\ell=0}^{s} \rho_\ell \ell^k = k \sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1}$$

for $k = 1, \cdots, p$, where $0^0 = 1$.

Proof. The local truncation error is

$$\sum_{\ell=0}^{s} \rho_\ell \mathbf{y}(t_{n+\ell}) - h \sum_{\ell=0}^{s} \sigma_\ell \mathbf{y}'(t_{n+\ell}).$$

We now expand the $\mathbf{y}$ and $\mathbf{y}'$ about t_n, and obtain

$$\left(\sum_{\ell=0}^{s} \rho_\ell\right)\mathbf{y}(t_n) + \sum_{k=1}^{\infty} \frac{h^k}{k!}\left(\sum_{\ell=0}^{s} \rho_\ell \ell^k - k\sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1}\right)\mathbf{y}^{(k)}(t_n).$$

This is $O(h^{p+1})$ under the given conditions. □

Theorem. A multi-step method has order p (with $p \geq 1$) if and only if

$$\rho(e^x) - x\sigma(e^x) = O(x^{p+1})$$

as $x \to 0$.

Proof. We expand

$$\rho(e^x) - x\sigma(e^x) = \sum_{\ell=0}^{s} \rho_\ell e^{\ell x} - x\sum_{\ell=0}^{s} \sigma_\ell e^{\ell x}.$$

We now expand the $e^{\ell x}$ in Taylor series about $x = 0$. This comes out as

$$\sum_{\ell=0}^{s} \rho_\ell + \sum_{k=1}^{\infty} \frac{1}{k!} \left(\sum_{\ell=0}^{s} \rho_\ell \ell^k - k \sum_{\ell=0}^{s} \sigma_\ell \ell^{k-1} \right) x^k.$$

So the result follows. □

Theorem (Dahlquist equivalence theorem)**.** A multi-step method is convergent if and only if

(i) The order p is at least 1; and

(ii) The root condition holds.

Lemma. An s-step backward differentiation method of order s is obtained by choosing

$$\rho(w) = \sigma_s \sum_{\ell=1}^{s} \frac{1}{\ell} w^{s-\ell} (w-1)^\ell,$$

with σ_s chosen such that $\rho_s = 1$, namely

$$\sigma_s = \left(\sum_{\ell=1}^{s} \frac{1}{\ell} \right)^{-1}.$$

Proof. We need to construct ρ so that

$$\rho(w) = \sigma_s w^s \log w + O(|w-1|^{s+1}).$$

This is easy, if we write

$$\begin{aligned} \log w &= -\log\left(\frac{1}{w}\right) \\ &= -\log\left(1 - \frac{w-1}{w}\right) \\ &= \sum_{\ell=1}^{\infty} \frac{1}{\ell} \left(\frac{w-1}{w}\right)^\ell. \end{aligned}$$

Multiplying by $\sigma_s w^s$ gives the desired result. □

5.4 Runge-Kutta methods

6 Stiff equations

6.1 Introduction

6.2 Linear stability

Theorem (Maximum principle)**.** Let g be analytic and non-constant in an open set $\Omega \subseteq \mathbb{C}$. Then $|g|$ has no maximum in Ω.

7 Implementation of ODE methods

7.1 Local error estimation

7.2 Solving for implicit methods

8 Numerical linear algebra

8.1 Triangular matrices

8.2 LU factorization

8.3 $A = LU$ for special A

Theorem. A sufficient condition for the existence for both the existence and uniqueness of $A = LU$ is that $\det(A_k) \neq 0$ for $k = 1, \cdots, n-1$.

Proof. Straightforward induction. □

Theorem. If $\det(A_k) \neq 0$ for all $k = 1, \cdots, n$, then $A \in R^{n\times n}$ has a unique factorization of the form

$$A = LD\hat{U},$$

where D is non-singular diagonal matrix, and both L and $\hat{U}$ are unit triangular.

Proof. From the previous theorem, $A = LU$ exists. Since A is non-singular, U is non-singular. So we can write this as

$$U = D\hat{U},$$

where D consists of the diagonals of U and $\hat{U} = D^{-1}U$ is unit. □

Theorem. Let $A \in \mathbb{R}^{n\times n}$ be non-singular and $\det(A_k) \neq 0$ for all $k = 1, \cdots, n$. Then there is a unique "symmetric" factorization

$$A = LDL^T,$$

with L unit lower triangular and D diagonal and non-singular.

Proof. From the previous theorem, we can factorize A uniquely as

$$A = LD\hat{U}.$$

We take the transpose to obtain

$$A = A^T = \hat{U}^T DL^T.$$

This is a factorization of the form "unit lower"-"diagonal"-"unit upper". By uniqueness, we must have $\hat{U} = L^T$. So done. □

Theorem. Let $A \in \mathbb{R}^{n\times n}$ be a positive-definite matrix. Then $\det(A_k) \neq 0$ for all $k = 1, \cdots, n$.

Proof. First consider $k = n$. To show A is non-singular, it suffices to show that $A\mathbf{x} = \mathbf{0}$ implies $\mathbf{x} = \mathbf{0}$. But we can multiply the equation by $\mathbf{x}^T$ to obtain $\mathbf{x}^T A\mathbf{x} = 0$. By positive-definiteness, we must have $\mathbf{x} = \mathbf{0}$. So done.

Now suppose $A_k\mathbf{y} = \mathbf{0}$ for $k < n$ and $\mathbf{y} = \mathbb{R}^k$. Then $\mathbf{y}^T A_k\mathbf{y} = 0$. We invent a new $\mathbf{x} \in \mathbb{R}^n$ by taking $\mathbf{y}$ and pad it with zeros. Then $\mathbf{x}^T A\mathbf{x} = 0$. By positive-definiteness, we know $\mathbf{x} = \mathbf{0}$. Then in particular $\mathbf{y} = \mathbf{0}$. □

Theorem. A symmetric matrix $A \in \mathbb{R}^{n\times n}$ is *positive-definite* iff we can factor it as

$$A = LDL^T,$$

where L is unit lower triangular, D is diagonal and $D_{kk} > 0$.

Proof. First suppose such a factorization exists, then

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T LDL^T\mathbf{x} = (L^T\mathbf{x})^T D(L^T\mathbf{x}).$$

We let $\mathbf{y} = L^T\mathbf{x}$. Note that $\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{x} = \mathbf{0}$, since L is invertible. So

$$\mathbf{x}^T A\mathbf{x} = \mathbf{y}^T D\mathbf{y} = \sum_{k=1}^{n} y_k^2 D_{kk} > 0$$

if $\mathbf{y} \neq 0$.

Now if A is positive definite, it has an LU factorization, and since A is symmetric, we can write it as

$$A = LDL^T,$$

where L is unit lower triangular and D is diagonal. Now we have to show $D_{kk} > 0$. We define $\mathbf{y}_k$ such that $L^T\mathbf{y}_k = \mathbf{e}_k$, which exist, since L is invertible. Then clearly $\mathbf{y}_k \neq \mathbf{0}$. Then we have

$$D_{kk} = \mathbf{e}_k^T D\mathbf{e}_k = \mathbf{y}_k^T LDL^T\mathbf{y}_k = \mathbf{y}_k^T A\mathbf{y}_k > 0.$$

So done. □

Proposition. If a band matrix A has band width r and an LU factorization $A = LU$, then L and U are both band matrices of width r.

Proof. Straightforward verification. □

9 Linear least squares

Theorem. A vector $\mathbf{x}^* \in \mathbb{R}^n$ minimizes $\|A\mathbf{x} - \mathbf{b}\|^2$ if and only if

$$A^T(A\mathbf{x}^* - \mathbf{b}) = 0.$$

Proof. A solution, by definition, minimizes

$$\begin{aligned} f(\mathbf{x}) &= \langle A\mathbf{x} - \mathbf{b}, A\mathbf{x} - \mathbf{b}\rangle \\ &= \mathbf{x}^T A A\mathbf{x} - 2\mathbf{x}^T A^T \mathbf{b} + \mathbf{b}^T \mathbf{b}. \end{aligned}$$

Then as a function of $\mathbf{x}$, the partial derivatives of this must vanish. We have

$$\nabla f(\mathbf{x}) = 2A^T(A\mathbf{x} - \mathbf{b}).$$

So a necessary condition is

$$A^T(A\mathbf{x} - \mathbf{b}).$$

Now suppose our $\mathbf{x}^*$ satisfies $A^T(A\mathbf{x}^* - \mathbf{b}) = 0$. Then for all $\mathbf{x} \in \mathbb{R}^n$, we write $\mathbf{x} = \mathbf{x}^* + \mathbf{y}$, and then we have

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|^2 &= \|A(\mathbf{x}^* + \mathbf{y}) - \mathbf{b}\| \\ &= \|A\mathbf{x}^* - \mathbf{b}\| + 2\mathbf{y}^T A^T(A\mathbf{x} - \mathbf{b}) + \|A\mathbf{y}\|^2 \\ &= \|A\mathbf{x}^* - \mathbf{b}\| + \|A\mathbf{y}\|^2 \\ &\geq \|A\mathbf{x}^* - \mathbf{b}\|. \end{aligned}$$

So $\mathbf{x}^*$ must minimize the Euclidean norm. □

Corollary. If $A \in \mathbb{R}^{m\times n}$ is a full-rank matrix, then there is a unique solution to the least squares problem.

Proof. We know all minimizers are solutions to

$$(A^T A)\mathbf{x} = A^T \mathbf{b}.$$

The matrix A being full rank means $\mathbf{y} \neq \mathbf{0} \in \mathbb{R}^n$ implies $A\mathbf{y} \neq \mathbf{0} \in \mathbb{R}^m$. Hence $A^T A \in \mathbb{R}^{n\times n}$ is positive definite (and in particular non-singular), since

$$\mathbf{x}^T A^T A\mathbf{x} = (A\mathbf{x})^T(A\mathbf{x}) = \|A\mathbf{x}\|^2 > 0$$

for $\mathbf{x} \neq \mathbf{0}$. So we can invert $A^T A$ and find a unique solution $\mathbf{x}$. □

Proposition. A matrix $A \in \mathbb{R}^{m\times n}$ can be transformed into upper-triangular form by applying n Householder reflections, namely

$$H_n \cdots H_1 A = R,$$

where each H_n introduces zero into column k and leaves the other zeroes alone.

Lemma. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, with $\mathbf{a} \neq \mathbf{b}$, but $\|\mathbf{a}\| = \|\mathbf{b}\|$. Then if we pick $\mathbf{u} = \mathbf{a} - \mathbf{b}$, then

$$H_{\mathbf{u}}\mathbf{a} = \mathbf{b}.$$

Proof. We just do it:

$$H_{\mathbf{u}}\mathbf{a} = \mathbf{a} - \frac{2(\|\mathbf{a}\| - \mathbf{a}^T\mathbf{b})}{\|\mathbf{a}\|^2 - 2\mathbf{a}^T\mathbf{b} + \|\mathbf{b}\|^2}(\mathbf{a} - \mathbf{b}) = \mathbf{a} - (\mathbf{a} - \mathbf{b}) = \mathbf{b},$$

where we used the fact that $\|\mathbf{a}\| = \|\mathbf{b}\|$. □

Lemma. If the first $k-1$ components of $\mathbf{u}$ are zero, then

(i) For every $\mathbf{x} \in \mathbb{R}^m$, $H_{\mathbf{u}}\mathbf{x}$ does not alter the first $k-1$ components of $\mathbf{x}$.

(ii) If the last $(m-k+1)$ components of $\mathbf{y} \in \mathbb{R}^m$ are zero, then $H_{\mathbf{u}}\mathbf{y} = \mathbf{y}$.

Lemma. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$, with

$$\begin{pmatrix} a_k \\ \vdots \\ a_m \end{pmatrix} \neq \begin{pmatrix} b_k \\ \vdots \\ b_m \end{pmatrix},$$

but

$$\sum_{j=k}^{m} a_j^2 = \sum_{j=k}^{m} b_j^2.$$

Suppose we pick

$$\mathbf{u} = (0, 0, \cdots, 0, a_k - b_k, \cdots, a_m - b_m)^T.$$

Then we have

$$H_{\mathbf{u}}\mathbf{a} = (a_1, \cdots, a_{k-1}b_k, \cdots, b_m).$$