Math 514

December 10, 2020

- Instructor: Chenxi Wu cwu367@wisc.edu
- Section 1: 9:55-10:45 am Section 2: 12:05-12:55 pm
- Office hours: 10:45am-noon Monday, Wednesday or by appointment

Recall that in the first half of the semester, we covered the following topics:

(i) Methods for root finding: Newton's method etc.

(ii) Numerical Linear Algebra: LU decomposition, QR algorithm etc.

The following are the new topics we will cover in the second half of the semester:

(i) Interpolation and approximation: how to get the formula of a function using discrete date.

(ii) Numerical integration: how to integrate a function knowing only its value on a discrete set.

(iii) Numerical solution for differential equations: numerical solution for ODE and PDE.

The first topic will be the foundation of the second and third topic.

Contents

1 Polynomial interpolation (Chapter 6)

Definition 1.1. *The* **polynomial interpolation** *of a function f at points $x_0, \ldots x_n$, is a polynomial p that shares some properties of f at those points, e.g. has the same value or same derivatives.*

We will focus on single variable functions, and discuss two kinds of polynomial interpolation problems:

(i) Lagrange interpolation: find a polynomial p of degree at most n, such that $p(x_i) = f(x_i)$.

(ii) Hermite interpolation: find a polynomial p of degree at most $2n+1$, such that $p(x_i) = f(x_i)$, $p'(x_i) = f'(x_i)$.

An application for Hermite interpolation is for approximating smooth curves where the direction of the curve at certain points are also specified.

1.1 Lagrange interpolation

1.1.1 Existence

Theorem 1.2. *(Lemma 6.1 in textbook) The Lagrange interpolation polynomial exists. In other words, for any $n+1$ distinct real numbers $x_0, \ldots, x_n$, and $n+1$ real numbers $y_0, \ldots, y_n$, there is a polynomial p of degree at most n such that $p(x_i) = y_i$.*

How do we find such a p?

Firstly, we observe that the map

$$T : p \mapsto [p(x_0), \ldots p(x_n)]^T \in \mathbb{R}^{n+1}$$

is linear. In other words, $(cp + dq)(x_i) = cp(x_i) + dq(x_i)$ for all i. Hence, finding p is like solving a system of non homogenous linear equations. Recall from linear algebra, let e_i be the standard basis vector of $\mathbb{R}^{n+1}$ corresponding to $y_i = 1$ and $y_j = 0$ for all $j \neq i$, then, if we can find some p_i of degree at most n such that $T(p_i) = e_i$, (i.e. $p_i(x_j) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$) then

$$T(\sum_i y_i p_i) = \sum_i y_i e_i = [y_0, \ldots y_n]^T \ .$$

Now we try and find the p_i: $p_i(x_j) = 0$ for all $j \neq i$, so $(x - x_j)$ must be a factor of p_i. So p_i must be something times

$$\prod_{j \neq i}(x - x_j) \ .$$

On the other hand, $p_i(x_i) = 1$, so the "something" should be

$$\frac{1}{\prod_{j \neq i}(x_i - x_j)} \ .$$

Now we have a proof of Theorem 1:

Proof. Let

$$p(x) = \sum_i \left(y_i \cdot \frac{\prod_{j \neq i}(x - x_j)}{\prod_{j \neq i}(x_i - x_j)} \right)$$

Then $p(x_k) = \sum_i \left(y_i \cdot \frac{\prod_{j \neq i}(x_k - x_j)}{\prod_{j \neq i}(x_i - x_j)} \right) y_i$. The k-th term is $y_k \times 1$ while all other terms are zero, hence the answer is y_k. □

1.1.2 Uniqueness

The problem of finding Lagrange interpolation polynomial is one with $n+1$ conditions and $n+1$ unknowns, so intuitively there should be a discrete set of solutions. Actually the solution can be shown to be unique:

Theorem 1.3. *(Theorem 6.1 in textbook) The Lagrange interpolation polynomial is unique. In other words, given $x_0, \ldots, x_n$ and $y_0, \ldots, y_n$ with x_i distinct, there is a single polynomial p of degree at most n such that $p(x_i) = y_i$*

Proof. The first step of the proof is to reduce the problem to the case where $y_i = 0$. Suppose p and q are two such polynomials, then $(p - q)(x_i) = 0$ for all i. So, to prove the theorem, we only need to show that if a polynomial $r = p - q$ of degree at most n vanishes at $n+1$ distinct points, then $r = 0$. This fact follows from the "fundamental theorem of algebra" and can be proved using long division. Here we provide two other alternative proofs:

(i) Approach I: use the mean value theorem in calculus. Firstly we show the following fact:

Lemma 1.4. *If $f \in C^{m-1}$ (f is $m-1$-th order differentiable with $m-1$-th derivative continuous), and $f = 0$ at m distinct points, then for any $0 \le k \le m-1$, $f^{(k)} = 0$ at at least $m-k$ points.*

Proof. By mean value theorem, between two consecutive zeros of f there must be a zero of f'. Hence f' vanishes at at least $m-1$ points. Now let f' take the role of f and continue the process, we get f'' vanishes at at least $m-2$ points, etc. $\square$

Suppose r vanishes at $x_0, \ldots, x_n$, and r is of degree $d > 0$. Then $r^{(d)}$ is a non zero constant. Apply Lemma 1.4 with $m = n+1$ and $k = d$, we see a contradiction. Hence $r = const$, which implies $r = 0$.

(ii) Approach II: Let $r = \sum_j a_j x^j$, then a_j are solutions of a system of linear equation $\sum_j a_j x_i^j = 0$. However from linear algebra,

$$\begin{vmatrix} 1 & 1 & \ldots & 1 \\ x_1 & x_2 & \ldots & x_{n+1} \\ \ldots & \ldots & \ldots & \ldots \\ x_1^n & x_2^n & \ldots & x_{n+1}^n \end{vmatrix} = \prod_{i<j}(x_j - x_i) \neq 0$$

Hence $a_j = 0$ for all j, which implies that $r = 0$.

$\square$

1.1.3 Error Estimate

We know that $f^{(n+1)} = 0$ iff f is a polynomial of degree at most n, so one may guess that if $f^{(n+1)}$ is small, f should be close to a polynomial of degree at most n, hence probably close to its Lagrange interpolation polynomial at $n + 1$ points. To make this more precise, we have the following theorem on error estimate of Lagrange interpolation:

Theorem 1.5. *(Theorem 6.2 in textbook) If $f \in C^{n+1}$, p is the Lagrange interpolation of f at $n+1$ distinct points $x_0, \ldots x_n$. Then for any x, there is some $s \in [\min\{x_i, x\}, \max\{x_i, x\}]$, such that*

$$f(x) - p(x) = \frac{f^{(n+1)}(s) \prod_i (x - x_i)}{(n+1)!}$$

Proof. When $x = x_i$ it's obvious. Now suppose x is distinct from all x_i. Consider the auxiliary function:

$$G(t) = f(t) - p(t) - (f(x) - p(x)) \cdot \frac{\prod_i(t - x_i)}{\prod_i(x - x_i)}$$

Then $G = 0$ at x_i and x, hence by Lemma 1.4 (let $m = n + 2$, $k = n + 1$), there must be some point $s \in [\min\{x_i, x\}, \max\{x_i, x\}]$ where

$$G^{(n+1)}(s) = f^{(n+1)}(s) - \frac{(f(x) - p(x))(n+1)!}{\prod_i(x - x_i)} = 0$$

Hence

$$f(x) - p(x) = \frac{f^{(n+1)}(s)\prod_i(x - x_i)}{(n+1)!}$$

$\square$

When the set $\{x_i\}$ becomes denser, $\prod_i(x-x_i)$ decreases, and $(n+1)!$ increases. However, when $n \to \infty$, the Lagrange interpolation polynomial may not converge to f even if f is smooth, if $f^{(n)}$ increases too fast.

Example 1.6. $f(x) = \cos(x)$, $x_i = 5i/n$, $i = 0, 1, 2, \ldots n$
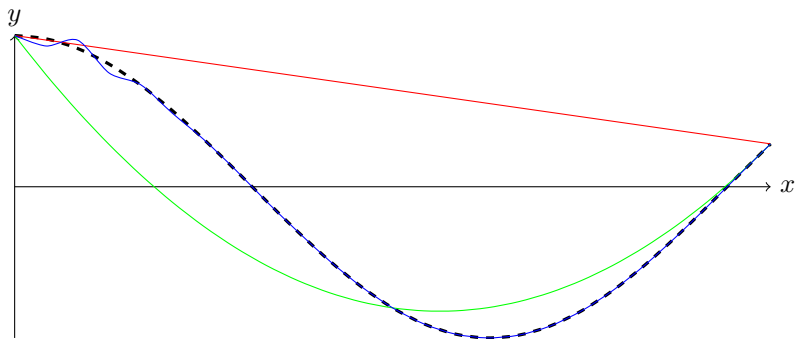
Figure 1: Black dashed line: $y = \cos(x)$. Red line: Lagrange interpolation with 2 points. Green line: Lagrange interpolation with 3 points. Blue line: Lagrange interpolation with 11 points.

Example 1.7. $f(x) = 1/(1 + 2(x - 2)^2)$, $x_i = 5i/n$, $i = 0, 1, 2, \ldots n$.
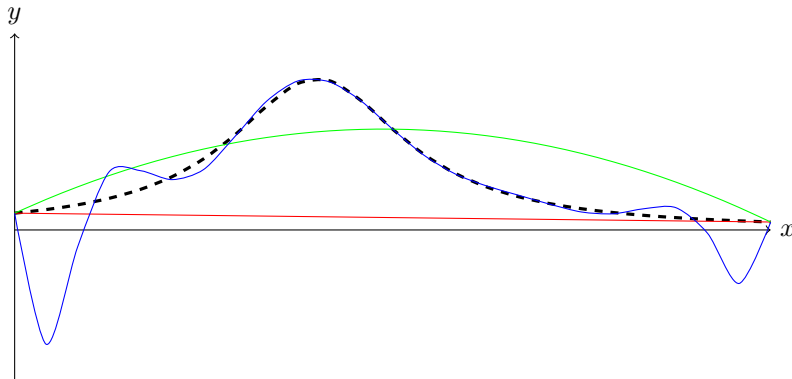
Figure 2: Black dashed line: $y = 1/(1 + 2(x - 2)^2)$. Red line: Lagrange interpolation with 2 points. Green line: Lagrange interpolation with 3 points. Blue line: Lagrange interpolation with 11 points.

The reason that the Lagrange interpolation polynomials in Example 1.6 converges but those in Example 1.7 don't, is that the higher order derivatives of cos is $\pm \sin$, $\pm \cos$ hence all bounded, while it is not true for the function in Example 1.7. As a practice, calculate the k-th derivative of $1/(1 + 2(x - 2)^2)$ at $x = 2$.

1.2 Hermite interpolation polynomial

1.2.1 Existence

Similar to the Lagrange case, we can construct the Hermite interpolation polynomial as follows:

Theorem 1.8. *(Existence part of Theorem 6.3 in textbook) There is a polynomial p of degree at most $2n + 1$, such that $p(x_i) = y_i$, $p'(x_i) = z_i$, $i = 0, \ldots n$, where x_i are distinct.*

Use the same strategy as the Lagrange case, possibly via a few trials and errors, one can find the formula of p as below:

Proof. Let

$$p(x) = \sum_i \left(z_i \cdot \frac{(x - x_i) \prod_{j \neq i}(x - x_j)^2}{\prod_{j \neq i}(x_i - x_j)^2} \right.$$

$$\left. + y_i \cdot \left(1 - (x - x_i) \sum_{j \neq i} \frac{2}{x_i - x_j} \right) \cdot \frac{\prod_{j \neq i}(x - x_j)^2}{\prod_{j \neq i}(x_i - x_j)^2} \right)$$

Then by calculation,

$$p(x_k) = \sum_i \left(z_i \cdot \frac{(x_k - x_i) \prod_{j \neq i}(x_k - x_j)^2}{\prod_{j \neq i}(x_i - x_j)^2} \right.$$

$$\left. + y_i \cdot \left(1 - (x_k - x_i) \sum_{j \neq i} \frac{2}{x_i - x_j} \right) \cdot \frac{\prod_{j \neq i}(x_k - x_j)^2}{\prod_{j \neq i}(x_i - x_j)^2} \right)$$

In the first sum, all terms have a factor $(x_k - x_k)$, so it must be zero. In the second sum, all but the k-th term is zero, and the k-th term is y_k. Similarly, by taking derivative and let $x = x_k$, we can show that $p'(x_k) = z_k$. $\square$

1.2.2 Uniqueness and Error Estimate

The mean value theorem argument (i.e. Lemma 1.4) can also be used to show the uniqueness and error estimate for Hermite interpolation polynomials:

Theorem 1.9. *(Uniqueness part of Theorem 6.3 in textbook) The Hermite interpolation polynomial is unique. In other words, there is a unique p of degree at most $2n + 1$ such that $p(x_i) = y_i$, $p'(x_i) = z_i$, $i = 0, \ldots n$, where x_i are distinct.*

Proof. Similar to the proof of Theorem 1.3, if we have two Hermite interpolation polynomials p and q, then $r = p - q$ satisfies $r(x_i) = r'(x_i) = 0$ and r has degree at most $2n + 1$. However, if r is non zero, it can not have $n + 1$ distincts roots x_i with multiplicity at least 2 each, hence $r = 0$.

We can also prove $r = 0$ using analysis like in Theorem 1.3. If r has degree at most $2n + 1$, $r' = r = 0$ at $n + 1$ points, then there must be n other points where $r' = 0$. Now suppose r has degree $d > 0$. Apply Lemma 1.4, let $m = 2n + 2$, $k = d$, then we get $r^{(d)}$ vanishes at $2n + 2 - d$ points, which contradicts with the fact that $r^{(d)}$ is a non zero constant. Hence $r = const$ which implies that $r = 0$. $\qquad\square$

Theorem 1.10. *(Theorem 6.4 in textbook)* *If $f \in C^{(2n+2)}$, there is $s \in [\min\{x_i, x\}, \max\{x_i, x\}]$, such that*

$$f(x) - p(x) = \frac{f^{(2n+2)}(s) \prod_i (x - x_i)^2}{(2n+2)!}$$

Proof. If $x = x_i$ then it is trivially true. Now assume x is not in $\{x_i\}$. Let

$$G(t) = f(t) - p(t) - \frac{(f(x) - p(x)) \prod_i (t - x_i)^2}{\prod_i (x - x_i)^2}$$

Then G vanishes at the $n+2$ points $x, x_0, \ldots x_n$, and G' vanishes at $n+1$ of them $x_0, \ldots, x_n$. By the same argument as above, G' vanishes at $n+1$ more points, hence it is zero at at least $2n+2$ points. Now use Lemma 1.4 on G' for $m = 2n+2$, $k = 2n+1$. $\qquad\square$

1.3 Applications

1.3.1 Numerical Differentiation

Suppose p is the Lagrange interpolation of f at $n+1$ points. By mean value theorem, $f'-p'$ is zero at n points $d_1, \ldots d_n$, so p' can be seen as the Lagrange interpolation polynomial with condition $p'(d_i) = f'(d_i)$ (see Theorem 6.5 in textbook). Now one can get an estimate for $f'(x) - p'(x)$ using Theorem 1.5.

Example 1.11. *For example, if we know the value of f at $x + ih$ for $i = -1, 0, 1$ as y_{-1}, y_0 and y_1, then the Lagrange interpolation polynomial is:*

$$p(x + t) = y_{-1}t(t - h)/(2h^2) - y_0(t + h)(t - h)/h^2 + y_1t(t + h)/(2h^2)$$

So

$$p'(0) = \frac{y_1 - y_{-1}}{2h} = \frac{f(x + h) - f(x - h)}{2h}$$

As $h \to 0$ this indeed converges to $f'(x)$.

However, this approach is generally unstable. If f is complex analytic one can use complex analysis to do it which is stable, which we will not cover in this class.

Numerical differentiation is useful in optimization or root finding via Newton's method.

1.3.2 Cubic Bézier curves

The **cubic Bézier curve** is a curve parametrized by cubic functions: $\gamma : [0,1] \to \mathbb{R}^2$, $\gamma(t) = (\gamma_1(t), \gamma_2(t))$, where γ_1 and γ_2 are both of degree at most 3, and $\gamma(0) = P_0$, $\gamma'(0) = 3(P_1 - P_0)$, $\gamma(1) = P_3$, $\gamma'(1) = 3(P_3 - P_2)$, where P_0, P_1, P_2 and P_3 are the four "control points".

To find the formula for cubic Bézier curve, we can apply the formula for Hermite interpolation polynomial for $n = 1$. Bézier curves has many applications in computer graphics and font design, and you might have already used it in applications that generate or edit vector graphics. Below is an example (drawn using LaTeX/TikZ):

P_2

P_0

P_3

P_1
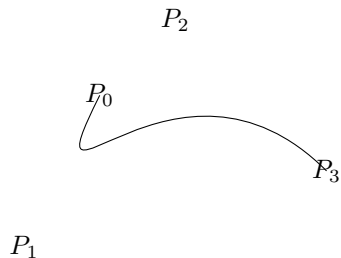
1.3.3 Linear and Hermite splines (Chapter 11)

From the Example 2 above we see that polynomial interpolation with high degree is not guaranteed to work well. Hence, in practice, we often try to keep the degree of the polynomial low, which means that we will need to use piecewise functions for interpolation. We will discuss two kinds of piecewise polynomial interpolation: **linear spline** and **Hermite cubic spline**. The textbook also covered the **natural cubic spline**.

Linear Spline

Definition 1.12. *Let f be a single variable function on $[a, b]$, $a = x_0 < x_1 \cdots < x_n = b$ $n + 1$ distinct points. The* **Linear Spline** s_L *with* **knots** *at x_i is defined as*

$$s_L(x) = \frac{x_i - x}{x_i - x_{i-1}} f(x_{i-1}) + \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i), \ \text{where } x_{i-1} \leq x \leq x_i$$

In other words, use the 2-point Lagrange interpolation for each interval $[x_{i-1}, x_i]$.

Theorem 1.13. *(Theorem 11.1 in textbook) Let $f \in C^2$, $h = \max\{x_i - x_{i-1}\}$, $M = \max |f''|$, then for any $x \in [a, b]$, $|f(x) - s_L(x)| \leq \frac{1}{8} h^2 M$.*

Proof. Suppose x is between x_{i-1} and x_i. Theorem 1.5 implies that

$$f(x) - s_L(x) = \frac{f''(c)(x - x_{i-1})(x - x_i)}{2!}$$

for some $c \in [x_{i-1}, x_i]$. From assumption, $|f''(c)| < M$ and

$$|(x - x_{i-1})(x - x_i)| \leq |(x_i - x_{i-1})/2|^2 \leq h^2/4 \ .$$

$\square$

The linear spline formula can be alternatively written as $s_L = \sum_i f(x_i)\phi_i$, where ϕ_i are the "hat functions", where, if $i = 1, \ldots, n - 1$,

$$\phi_i(x) = \begin{cases} (x - x_{i-1})/(x_i - x_{i-1}) & x \in [x_{i-1}, x_i] \\ (x - x_{i+1})/(x_i - x_{i+1}) & x \in [x_i, x_{i+1}] \\ 0 & otherwise \end{cases}$$

ϕ_0 and ϕ_n can be written down similarly. As a consequence, s_L lies in the span of ϕ_n.

Hermite cubic spline

Definition 1.14. *Let* $f \in C^1[a, b]$, $a = x_0 < x_1 \cdots < x_n = b$ $n + 1$ *distinct points. The* **Hermite Cubic Spline** s_H *with* **knots** *at* x_i *is defined as* $s_H(x) = p_i(x)$ *for* $x \in [x_{i-1}, x_i]$, *where* p_i *is the Hermite interpolation polynomial defined using* $\{x_{i-1}, x_i\}$.

Theorem 1.15. *(Theorem 11.4 in textbook) Let* $f \in C^2$, $h = \max\{x_i - x_{i-1}\}$, $M = \max |f^{(4)}|$, *then for any* $x \in [a, b]$, $|f(x) - s_H(x)| \leq \frac{1}{384} h^4 M$.

The proof is similar to Theorem 1.13. Note that $384 = 4!2^4$.

One can also find a set of basis functions for s_H.

1.4 Review

- Definition and formula of Lagrange/Hermite interpolation polynomials.

- Uniqueness.

- Error estimate.

2 Approximation Theory (Chapter 8, 9)

We can see that the Lagrange interpolation polynomial, Hermite interpolation polynomial, and the splines all lie in a vector space spanned by finitely many functions. In other words, all these algorithms can be seen as a way to **approximate a function using the linear combination of simpler functions**.

Recall that a set of functions form a vector space if it is closed under addition and scalar multiplication.

2.1 Approximation in normed vector space

Definition 2.1. *Let V be a vector space. A* **norm** *on V is a function: $\|\cdot\| : V \to \mathbb{R}_{\geq 0}$ such that:*

- $\|x\| = 0$ *iff* $x = 0$

- $\|x + y\| \leq \|x\| + \|y\|$

- $\|cx\| = |c|\|x\|$.

Example 2.2. *(i) $V = C([a,b])$ (continuous functions on $[a,b]$), $\|f\|_\infty = \max|f|$. This is called the L^∞ norm.*

(ii) $V = L^2([a,b])$, $\|f\|_2 = (\int_a^b |f(x)|^2 dx)^{1/2}$. This is called the L^2 norm.

(iii) Replace 2 with $p \geq 1$ we get L^p norm. If $p < 1$, the triangle inequality is no longer satisfied.

Definition 2.3. *Let $L = span\{x_1, \ldots x_m\}$ be a m-dimensional subspace of V, $x \in V$. The* **best approximation** *of x is the element $x' \in L$ that minimizes $\|x - x'\|$.*

Theorem 2.4. *(Theorem 8.2 in textbook) The best approximation always exists.*

The proof has two steps:

(i) $\| \cdot -x\|$ is continuous on L.

(ii) $\| \cdot -x\|$ goes to infinity at infinity.

Key idea: if a function is defined on a finite dimensional vector space, continuous, and goes to infinity at infinity, then it has a minimum.

Please ignore the proof below if you are not interested.

Proof. Let $x_1, \ldots x_m$ be a basis of L. Consider a function $F_x : \mathbb{R}^m \to \mathbb{R}$ defined as

$$F_x((t_1, \ldots, t_m)) = \|x - \sum_i t_i x_i\|$$

The first step of the proof is to show that F is continuous:

Lemma 2.5. *F_x is continuous.*

Proof. Suppose $t' \in \mathbb{R}^m$ satisfies that $|t'_i| < \epsilon$ for all i, then by triangle inequality,

$$|F_x(t + t') - F_x(t)| \leq |\sum_i t'_i x_i| \leq \epsilon \sum_i |x_i|$$

This implies that if t' is sufficiently small, $F_x(t + t')$ can be arbitrarily close to $F_x(t)$, hence F_x is continuous. $\qquad \square$

Now, let $D_R \subset \mathbb{R}^m = \{t : |t_i| \leq R \text{ for all } i\}$. It is a closed set, hence compact (recall the definition of compactness in your analysis class), hence a continuous function F_x takes minimum at some point x_R^* on D_R. We just need to show that if R is large enough, x_R^* is also the minimum of F_x.

Let $g > 0$ be the minimum of F_0 on the set D_1. Now we set $R_0 = (2|x| + 1)/g$. Then for any y outside D_{R_0}, then $F_x(y) = \|y - x\| \geq \|x\| + 1 > F_x(0) \geq F_x(x_{R_0}^*)$ $\qquad\square$

2.2 Stone-Weiersterass theorem

Theorem 2.6. *(Theorem 8.1 in textbook) For continuous function $f \in C([a,b])$, any $\epsilon > 0$, there is some polynomial p such that $\|f - p\|_\infty < \epsilon$.*

There are many proofs, some work for more general settings. An easy proof is first use linear spline to approximate f, then use polynomials to approximate the basis function (which is a linear combination of absolute values, which can be approximated by $(x^2 + \epsilon')^{1/2}$, which can be approximated using Taylor expansion).

2.3   Approximation in inner product space

Sometimes the norm on a vector space arises from a inner product (a symmetric, positive definite, bilinear form) $(\cdot, \cdot) : V \times V \to \mathbb{R}$, by $||x|| = \sqrt{(x, x)}$. If so, we call it an **inner product space**.

Example 2.7. *The L^2 norm on $L^2([a, b])$ arises from inner product $(f, g) = \int_a^b fg dx$. Let w be a non negative, continuous and integrable "weight function" on $[a, b]$, we can also defined the "weighted L^2 norm" which is from $(f, g)_w = \int_a^b wfg dx$.*

It's easy to see that the L_w^2 norm satisfies:

$$\|f\|_w \leq (\int_a^b w dx)^{1/2} \|f\|_\infty$$

Example 2.8. *On $C^1([a, b])$ we can define the $(1, 2)$ Sobolev norm $\|f\|_{1,2} = (\int_a^b |f(x)|^2 + |f'(x)|^2 dx)^{1/2}$. This norm also come from an inner product*

$$(f, g)_{1,2} = \int_a^b f(x)g(x) + f'(x)g'(x) dx$$

Let $L = span\{x_1, \ldots, x_m\}$, then we can use Gram-Schmidt process to get an orthonormal basis of L under $(\cdot, \cdot)$, called $\{e_1, \ldots, e_m\}$. Then we have:

Theorem 2.9. *The best approximation of $x \in V$ by an element of L is unique, and it is*

$$x^* = \sum_i (x, e_i) e_i$$

Proof. For any other $x' = \sum_i t_i e_i \in L$,

$$\|x' - x\|^2 = ((x' - x^*) + (x^* - x), (x' - x^*) + (x^* - x))$$

$$= \|x' - x^*\|^2 + \|x^* - x\|^2 + 2(\sum_i (t_i - (x, e_i)) e_i, \sum_i (x, e_i) e_i - x)$$

$$= \|x' - x^*\|^2 + \|x^* - x\|^2 + 2\sum_j (t_i - (x, e_j))(e_j, \sum_i (x, e_i) e_i - x)$$

$$= \|x' - x^*\|^2 + \|x^* - x\|^2 + 2\sum_j (t_i - (x, e_j))(\sum_i (x, e_j)(e_j, e_j) - (x, e_j))$$

$$= \|x' - x^*\|^2 + \|x^* - x\|^2 \geq \|x^* - x\|^2$$

And equality is reached only when $x' = x^*$. $\qquad\qquad\square$

When the inner product is the L_w^2 inner product, the integrals in the formula for best approximation will often be calculated numerically (cf. next Section).

The proof is the same as the finite dimensional case you have seen in linear algebra.

If x_i are only orthogonal and not orthonormal, the formula becomes

$$x^* = \sum_i \frac{(x, x_i)}{(x_i, x_i)} x_i$$

If x_i are not known to be orthogonal either, the formula becomes

$$x^* = \sum_i (\sum_j (x, x_j)(A^{-1})_{i,j}) x_i$$

Where

$$A_{i,j} = (x_i, x_j)$$

2.4   Orthogonal polynomials

Definition 2.10. *We call ϕ_j, $j = 0, 1, 2, \ldots$ a system of* **orthogonal polynomials** *with weight w, if*

  (i) ϕ_j *is of degree j.*

  (ii) ϕ_j *are orthogonal to each other in L_w^2 norm.*

Theorem 2.11. *If w is positive, continuous and integrable on (a, b) then a system of* **orthogonal polynomials** *with weight w exists.*

Proof. This is Gram-Schmidt applied to $\{1, x, x^2, x^3, \ldots\}$.

$$\phi_0 = 1$$

$$\phi_j = x^j - \sum_{i=0}^{j-1} \frac{\int_a^b wt^j \phi_i(t)dt}{\int_a^b w\phi_i^2(t)dt} \phi_i(x)$$

$\square$

From linear algebra, we know that the system of orthogonal polynomials is unique up to scaling, since ϕ_j is the basis vector of the orthogonal complement of $span\{1, x, \ldots, x^{j-1}\}$ in $span\{1, x, \ldots, x^j\}$.

Remark 2.12. *Stone-Weiersterass theorem implies that as degree increases, optimal approximation in $L_w^2([a, b])$ can become arbitrarily accurate. In other words, the orthogonal polynomials form an orthonormal basis of $L_w^2([a, b])$. (Which is NOT a basis in the sense of linear algebra. In algebra there is only finite sum.)*

Example 2.13. *Let $(a, b) = (-1, 1)$.*

- *If $w = 1$, the resulting orthogonal polynomials are called the **Legendre polynomials** L_j.*

- *If $w(x) = (1 - x^2)^{-1/2}$, the resulting orthogonal polynomials are called the **Chebyshev polynomials** T_j.*

Remark 2.14. *The Chebyshev polynomials have a particularly nice formula:*

$$T_j = \cos(j \cos^{-1} x) \ .$$

They are polynomials because

$$T_0 = 1, T_1 = x, T_2 = 2x^2 - 1$$

$$
\begin{aligned}
T_j = \cos(j \cos^{-1} x) &= x \cos((j-1) \cos^{-1} x) - \sin(\cos^{-1} x) \sin((j-1) \cos^{-1} x) \\
&= x \cos((j-1) \cos^{-1} x) - \sin^2(\cos^{-1}(x)) \cos((j-2) \cos^{-1} x) \\
&\quad - \sin(\cos^{-1} x) \sin((j-2) \cos^{-1} x) \cos(\cos^{-1} x) \\
&= x T_{j-1} - (1 - x^2) T_{j-2} - x(T_{j-3} - T_{j-1})/2 \ .
\end{aligned}
$$

They are othogonal, because $j \neq j'$,

$$\int_{-1}^{1} \cos(j \cos^{-1} x) \cos(j' \cos^{-1} x)(1 - x^2)^{-1/2} dx$$

$$= - \int_{-1}^{1} \cos(j \cos^{-1} x) \cos(j' \cos^{-1} x) d \cos^{-1}(x)$$

$$= \int_{0}^{\pi} \cos(jt) \cos(j't) dt = 0$$

Remark 2.15. *Furthermore, if $T_j = \cos(j \cos^{-1} x)$, $2^{-j} T_{j+1}$ is the degree $j + 1$ monic (leading coefficient being 1) polynomial with the smallest L^∞ norm. This tells us that the term $\prod_i (x - x_i)$ in Theorem 1.5 can be minimized (in L^∞) if x_i are chosen as the roots of Chebyshev polynomials, or, in other words, if $\prod_i (x - x_i) = 2^{-n} T_{n+1}$. This is proved in Chapter 8 of the textbook.*
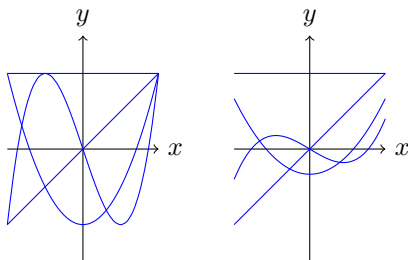
Figure 3: Chebyshev polynomials and Lagrange polynomials

- First Legendre polynomials (which I calculated using Gram-Schmidt, another alternative calculation can be found in the exercises, and also HW 4)

$$L_0 = 1, L_1 = x, L_2 = x^2 - \frac{1}{3}$$

$$L_3 = x^3 - \frac{3}{5}x$$

- First Chebyshev polynomials:

$$T_0 = 1, T_1 = x, T_2 = 2x^2 - 1, T_3 = 4x^3 - 3x$$

Theorem 2.16. *If the weight function w is positive, continuous and integrable on (a, b), then ϕ_j has j distinct real roots in (a, b).*

Proof. Suppose not, then ϕ_j switches sign fewer than j times in (a, b). Suppose $x_1, \ldots, x_k$ are the points in (a, b) where ϕ_j changes sign, then $(\phi_j, \prod_{i=1}^{k}(x - x_i))$ is non zero. However $\prod_{i=1}^{k}(x - x_i) \in span\{\phi_0, \ldots \phi_{j-1}\}$, hence a contradiction. $\qquad\square$

This Theorem will be used in the next section when we discuss Gauss's method for numerical integration.

2.5   Review

- Normed vector space, inner product space, L^∞, L^2 and L_w^2 norms.

- Existence of optimal approximation. Calculation of optimal approximation for inner product space.

- Concept of orthogonal polynomials.

Example 2.17. *Consider the function $y = e^x$ on $[-1, 1]$.*

- *Find the Lagrange interpolation polynomial, interpolating at 0, ± 1.*

- *Find the Hermite interpolation polynomial, interpolating at ± 1.*

- *Find the best approximation via a polynomial of degree at most 2, under the L^2 norm.*

Answer:

- Use formula $p_L = \sum_i y_i \prod_{j \neq i}(x - x_j) / \prod_{j \neq i}(x_i - x_j)$:

$$p_L(x) = e^{-1} \cdot \frac{x(x-1)}{-1 \cdot -2} + 1 \cdot \frac{(x+1)(x-1)}{1 \cdot -1} + e \cdot \frac{x(x+1)}{1 \cdot 2}$$

$$= (e^{-1}/2 + e/2 - 1)x^2 + (e/2 - e^{-1}/2)x + 1$$

- Use formula $p_H = \sum_i z_i(x - x_i) \prod_{j \neq i}(x - x_j)^2 / \prod_{j \neq i}(x_i - x_j)^2 + \sum_i y_i(1 - (x - x_i) \sum_{j \neq i}(2/(x_i - x_j))) \prod_{j \neq i}(x - x_j)^2 / \prod_{j \neq i}(x_i - x_j)^2$:

$$p_H(x) = e^{-1} \cdot \frac{(x+1)(x-1)^2}{(-1-1)^2} + e \cdot \frac{(x-1)(x+1)^2}{(1+1)^2}$$

$$+e^{-1} \cdot (1 + (x+1)) \cdot \frac{(x-1)^2}{(-1-1)^2} + e \cdot (1 - (x-1)) \cdot \frac{(x+1)^2}{(1+1)^2}$$

$$= (e^{-1}/2)x^3 + (e/4 - e^{-1}/4)x^2 + (e/2 - e^{-1})x + e/4 + 3e^{-1}/4$$

- Use formula $x^* = \sum_i ((x, x_i)/(x_i, x_i))x_i$:

$$p_2(x) = \frac{\int_{-1}^1 e^t dt}{\int_{-1}^1 1^2 dt} \cdot 1 + \frac{\int_{-1}^1 te^t dt}{\int_{-1}^1 t^2} \cdot x + \frac{\int_{-1}^1 (t^2 - 1/3)e^t dt}{\int_{-1}^1 (t^2 - 1/3)^2 dt} \cdot (x^2 - 1/3)$$

$$= \frac{(e - e^{-1})}{2} \cdot 1 + \frac{2e^{-1}}{2/3} \cdot x + \frac{2e/3 - 14e^{-1}/3}{8/45}(x^2 - 1/3)$$

$$= \frac{15e - 105e^{-1}}{4}x^2 + 3e^{-1}x + \frac{-3e + 33e^{-1}}{4}$$
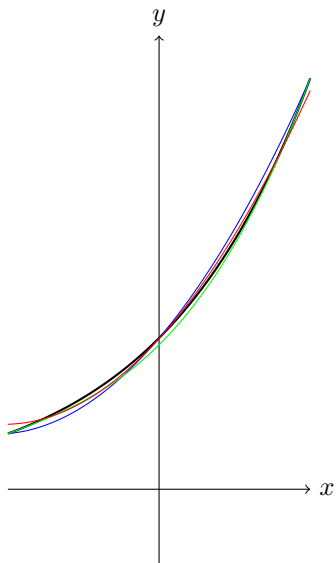
Figure 4: Black line: $y = e^x$. Blue line: Lagrange interpolation. Green line: Hermite interpolation. Red line: L^2 best approximation

3   Numerical Integration (Chapter 7, 10)

3.1   Quadrature rule

Question: Estimate $\int_a^b f(x)dx$.

Let $x_0 = a < x_1 < \cdots < x_n = b$ be $n+1$ distinct points in $[a,b]$, then we can use the Lagrange interpolation polynomial to estimate f, and hence

$$\int_a^b f(x)dx \approx \sum_k w_k f(x_k)$$

Where

$$w_k = \int_a^b \frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)}dx$$

To estimate the integration.

The points x_i are called **quadrature points**, and w_i called **quadrature weights**. The formula still works if some x_i is outside $[a,b]$.

3.2 Newton-Cotes method

Definition 3.1. *When $a = x_0 < x_1 < \cdots < x_n = b$ are $n + 1$ evenly spaced points, the formula above is called the* **Newton-Cotes formula**, *the evenly spaced x_i the* **Newton-Cotes quadrature**.

Example 3.2. *When $n = 1$, $w_0 = w_1 = \frac{b-a}{2}$, this is called the* **Trapezium rule** *(as it's like calculating the area of a collection of trapeziums). When $n = 2$, $w_0 = w_2 = \frac{b-a}{6}$, $w_1 = \frac{2(b-a)}{3}$. This is called* **Simpson's rule**.

Example 3.3. *$\int_0^1 \sin(x) dx$. Using Trapezium rule, the estimate is*

$$\frac{\sin(0) + \sin(1)}{2} = 0.4207$$

Using Simpson's rule, the estimate is

$$\sin(0)/6 + \sin(0.5) * 2/3 + \sin(1)/6 = 0.4599$$

The true value is $1 - \cos(1) = 0.4597$.

Theorem 3.4. *(Theorem 7.1 in the textbook) The error for the quadrature is bounded by*

$$\frac{\max|f^{(n+1)}|}{(n+1)!} \int_a^b \prod_i |x - x_i| dx$$

The proof is follows immediately from the error estimate of Lagrange interpolation. When x_i are Newton-Cotes quadrature, the error bound is $O(\max|f^{(n+1)}|(b-a)^{n+2})$, because when we scale the interval $[a, b]$ by C, the function being integrated is scaled by C^{n+1} while the range of the integration is also scaled by C.

For Newton-Cotes, when n is even (in order words when we have odd number of quadrature points), the error bound can be improved to $\max |f^{(n+2)}| \cdot O((b-a)^{n+3})$ provided $f \in C^{n+2}$:

Theorem 3.5. *Let n be an even number, $f \in C^{n+2}([a,b])$, $I_n(f)$ the Newton-Cotes formula using $n+1$ evenly spaced points on $[a,b]$, then there is some C_n (depending on n) such that*

$$|\int_a^b f(x)dx - I_n(f)| \leq C_n \max |f^{(n+2)}|(b-a)^{n+3}$$

Proof. The uniqueness of Lagrange interpolation implies that if f is a polynomial of degree at most n, $\int_a^b f(x)dx = I_n(f)$. Now consider the polynomial $g = \prod_i (x - x_i)$. Because x_i are evenly spaced, the graph of $\prod_i (x - x_i)$ is symmetric with respect to the point $(x_{n/2}, 0)$ where $x_{n/2} = (a+b)/2$. So $\int_a^b g dx = 0 = I_n(g)$. However, any polynomial of degree at most $n+1$ can be written in the form $cg + h$, where h is a polynomial of degree at most n. Hence, if f is any polynomial of degree at most $n+1$, $\int_a^b f(x)dx = I_n(f)$.

Now suppose $f \in C^{n+2}$. Let x_{n+1} be the midpoint of $[x_0, x_1]$, let p' be the Lagrange interpolation polynomial of f, then $f - p'$ vanishes at $x_0, \ldots, x_n, x_{n+1}$, hence the quadrature formula using $x_0, \ldots x_{n+1}$ is 0. Apply Theorem 3.4 we get

$$| \int_a^b (f - p')dx | \leq C_n \max |f^{n+2}|(b-a)^{n+3}$$

However, because $p'(x_i) = f(x_i)$ for $i = 0, 1, 2, \ldots, n$,

$$\int_a^b p'dx = I_n(p') = I_n(f)$$

Which proves the theorem. $\qquad\square$

We can also use mean value theorem to get finer bounds. As an example, when $n = 2$, we have

Theorem 3.6. *(Theorem 7.2 in the textbook) If $f \in C^4$, there is some $c \in [a, b]$ such that*

$$\int_a^b f(x)dx - (b-a) \cdot (f(a)/6 + 2f((a+b)/2)/3 + f(b)/6) = -\frac{f^{(4)}(c)(b-a)^5}{2880}$$

Proof. Let

$$G_1(t) = \int_{(a+b)/2-t}^{(a+b)/2+t} f(s)ds - 2t \cdot (f((a+b)/2 - t)/6$$

$$+ 2f((a+b)/2)/3 + f((a+b)/2 + t)/6)$$

$$G(t) = G_1(t) - (\frac{t}{(b-a)/2})^5 G_1((b-a)/2)$$

Then $G(0) = G((b-a)/2) = G'(0) = G''(0) = 0$. So there is $0 < c_1 < (b-a)/2$ such that $G'(c_1) = 0$, $0 < c_2 < c_1$ such that $G''(c_2) = 0$, $0 < c_3 < c_2$ such that $G'''(c_3) = 0$.

By calculation, $G_1'''(c_3) = \frac{c_3}{3} \cdot (f'''((a+b)/2 - c_3) - f'''((a+b)/2 + c_3))$, so

$$\frac{c_3}{3} \cdot (f'''((a+b)/2 - c_3) - f'''((a+b)/2 + c_3)) - \frac{1920}{(b-a)^5} c_3^2 G_1((b-a)/2) = 0$$

Hence

$$\frac{\cdot (f'''((a+b)/2 + c_3) - f'''((a+b)/2 - c_3))}{2c_3} = -\frac{2880}{(b-a)^5} G_1((b-a)/2)$$

Now apply mean value theorem for f''' on $[(b+a)/2 - c_3, (b+a)/2 + c_3]$, we get the c. $\square$

Alternative proof. Following the same argument as Theorem 3.5, we know that if f is a polynomial of degree 3, $\int_a^b f dx$ equals the result of Simpson's rule.

Now let p be the Lagrange interpolation of f at four distinct points a, b, $\frac{a+b}{2}$ and d. Then

$$\int_a^b p dx = (b-a)(\frac{p(a)}{6} + \frac{2p((a+b)/2)}{3} + \frac{p(b)}{6})$$

$$= (b-a)(\frac{f(a)}{6} + \frac{2f((a+b)/2)}{3} + \frac{f(b)}{6})$$

However, by error bound of Lagrange polynomials,

$$f(x) - p(x) = \frac{f^{(4)}(c)(x-a)(x-b)(x-(a+b)/2)(x-d)}{4!}$$

Now push d towards c, which makes $(x-a)(x-b)(x-(a+b)/2)(x-d)$ non positive on $[a, b]$. Now integrate for $x \in [a, b]$ and use integration mean value theorem, we get the theorem. $\square$

3.3 Composite Method

For the same reason as in Example 1.7, when $n \to \infty$ the error can not be guaranteed to decay to 0. So we often evenly decompose the interval $[a, b]$ into subintervals then carry out low order Newton-Cotes.

Let $a = x_0 < \cdots < x_n = b$ be $n+1$ evenly spaced points on $[a, b]$. If $n = dm$, we can cut $[a, b]$ into m subintervals each with $d+1$ quadrature points, and apply Newton-Cotes on each.

For example, if $d = 1$, we cut $[a, b]$ into n subintervals and apply trapezium rule on each we get

$$\frac{b-a}{n}(f(x_0)/2 + \sum_{i=1}^{n-1} f(x_i) + f(x_n)/2)$$

If $d = 2$, we cut $[a, b]$ into $n/2$ subintervals, and apply simpson's rule on each, we get

$$\frac{b-a}{3n}(f(x_0) + 4\sum_{i=1}^{n/2} f(x_{2i-1}) + 2\sum_{i=1}^{n/2-1} f(x_{2i}) + f(x_n))$$

When f is smooth with bounded higher order derivatives, the error estimate for each subinterval is $O(n^{-3})$ and $O(n^{-5})$ using the trapezium and Simpson's rule. Hence, the error composite trapezium and composite Simpson's rules decay like $O(n^{-2})$ and $O(n^{-4})$ respectively.

Example 3.7. $\int_0^1 \sin(x)dx$. *For this case, the error for composite trapezium & Simpson's rule can be calculated explicitly.*

- Composite trapezium rule with $n + 1$ points:

$$I_n = \frac{\sum_{i=1}^{n-1} \sin(\frac{i}{n}) + sin(1)/2}{n}$$

$$= \frac{\sum_{i=1}^{n-1}(\cos(\frac{i-1/2}{n}) - \cos(\frac{i+1/2}{n})) + \sin(\frac{1}{2n})\sin(1)}{2n\sin(\frac{1}{2n})}$$

$$= \frac{\cos(\frac{1}{2n}) - \cos(1)\cos(\frac{1}{2n})}{2n\sin(\frac{1}{2n})} = (1 - \cos(1)) \cdot \frac{\cos(\frac{1}{2n})}{2n\sin(\frac{1}{2n})}$$

And it is easy to see that

$$\frac{\cos(\frac{1}{2n})}{2n\sin(\frac{1}{2n})} = 1 - \frac{1}{3}(2n)^{-2} + \dots$$

So the error decay at $O(n^{-2})$.

- Composite Simpson's rule with $2m + 1$ points:

$$I_m = \frac{4 \sum_{i=1}^{m} \sin(\frac{2i-1}{2m}) + 2 \sum_{i=1}^{m-1} \sin(\frac{i}{m}) + sin(1)}{6m}$$

$$= \frac{2(1 - \cos(1)) + \cos(\frac{1}{2m}) - \cos(\frac{2m-1}{2m}) + \sin(1)\sin(\frac{1}{2m})}{6m \sin(\frac{1}{2m})}$$

$$= (1 - \cos(1)) \cdot \frac{2 + \cos(\frac{1}{2m})}{6m \sin(\frac{1}{2m})}$$

$$= (1 - \cos(1)) \cdot \frac{3 - (2m)^{-2}/2 + (2m)^{-4}/24 + O((2m)^{-6})}{3 - (2m)^{-2}/2 + (2m)^{-4}/40 + O((2m)^{-6})}$$

$$= (1 - \cos(1)) + O(m^{-4})$$

Example 3.8. $\int_{-0.5}^{0.5} \sqrt{1 - x^2} dx$. *The right answer should be*

$$\sqrt{3}/4 + \pi/6 = 0.9566114774905181$$

(i) Trapezium rule:

$$(\sqrt{3/4} + \sqrt{3/4})/2 = 0.8660254037844386$$

(ii) Simpson's rule:

$$\sqrt{3/4}/6 + 1 \times 2/3 + \sqrt{3/4}/6 = 0.9553418012614795$$

Algorithm 1: Composite Trapezium rule

---

1 $r \leftarrow f(a) + f(b)$;
2 **for** $i = 1, \ldots n - 1$ **do**
3 $\quad \big|\quad r \leftarrow r + 2 \times f(\frac{(n-i)a+ib}{n})$;
4 **end**
5 The answer is $\frac{(b-a)r}{2n}$;

Algorithm 2: Composite Simpson's rule

---

1 $r \leftarrow f(a) + f(b)$;
2 **for** $i = 1, \ldots n - 1$ **do**
3 $\quad$ **if** i *is odd* **then**
4 $\quad\quad \big|\quad r \leftarrow r + 4 \times f(\frac{(n-i)a+ib}{n})$;
5 $\quad$ **else**
6 $\quad\quad \big|\quad r \leftarrow r + 2 \times f(\frac{(n-i)a+ib}{n})$;
7 $\quad$ **end**
8 **end**
9 The answer is $\frac{(b-a)r}{3n}$;

```python
from math import *
f=lambda x : (1-x*x)**0.5
def composite_trapezium(n, a, b, f):
    r=0
    r+=0.5*(f(a)+f(b))
    for i in range(1, n):
        r+=f(((n-i)*a+i*b)/n)
    return r*(b-a)/n
def composite_simpsons(n, a, b, f):
    r=0
    r+=f(a)+f(b)
    for i in range(1, n, 2):
        r+=4*f(((n-i)*a+i*b)/n)
    for i in range(2, n, 2):
        r+=2*f(((n-i)*a+i*b)/n)
    return r*(b-a)/3/n
```
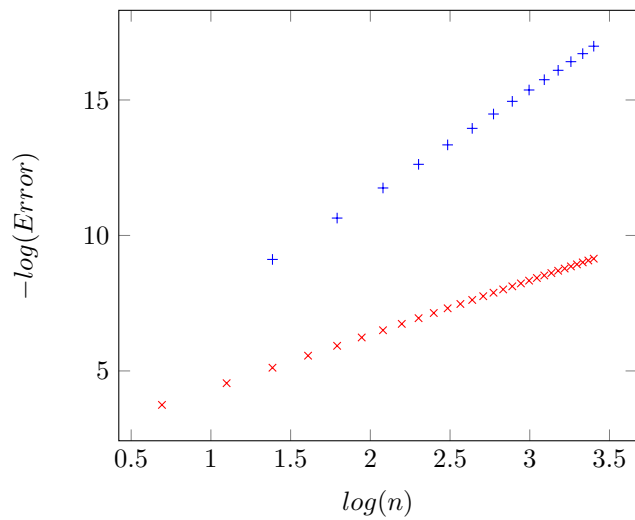
If we do $[a,b] = [-1,1]$, with the same function as above, we get:
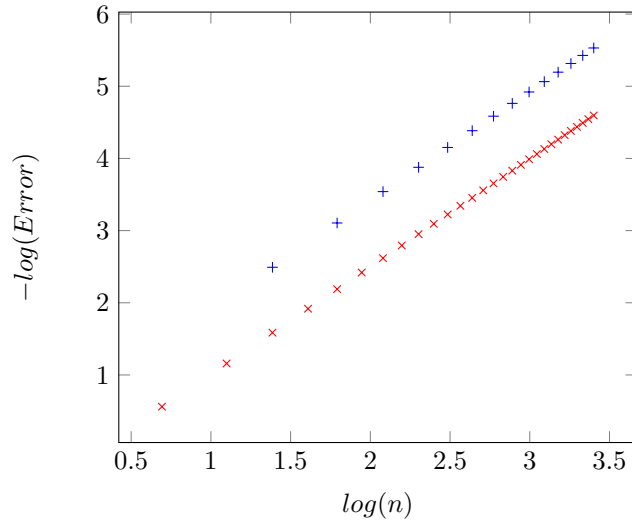
Because the derivatives go to infinity at x close to ± 1, both methods only have $O(n^{-1})$ error bound.

3.4　Friday Review And Examples

- Interpolation

  - Lagrange interpolation
  - Hermite interpolation
  - Uniqueness and error estimate

- Approximation theory

  - Normed vector space and inner product space
  - Gram-Schmidt
  - Orthogonal projection

- Numerical integration

  - Quadrature rule: $I = \sum_{i=0}^{n} w_i f(x_i)$
  - Newton-Cotes quadrature, $n = 1, 2$
  - Error estimate via error estimate of Lagrange interpolation
  - Composite methods

Example: $f(x) = \sin(x)$, $x \in [0, \pi]$, $x_0 = 0, x_1 = \pi/2, x_2 = \pi$.

- The Lagrange interpolation polynomial is

$$p(x) = 0 \cdot \frac{(x - \pi/2)(x - \pi)}{\pi^2/2} + 1 \cdot \frac{x(x - \pi)}{-\pi^2/4} + 0 \cdot \frac{(x - \pi/2)x}{\pi^2/2}$$

- The numerical integration using Simpson's rule is

$$I(f) = \int_0^\pi p(x)dx = 0 \cdot \frac{\pi}{6} + 1 \cdot \frac{2\pi}{3} + 0 \cdot \frac{\pi}{6} = \frac{2\pi}{3}$$

- The error estimate for Lagrange interpolation:

$$|f(x) - p(x)| = \left| \frac{f'''(s)x(x - \pi/2)(x - \pi)}{3!} \right| \leq \frac{1}{6}|x(x - \pi/2)(x - \pi)|$$

- Integrate the error estimate above, we get

$$|\int_0^\pi f(x)dx - I(f)| \leq \frac{1}{6}\int_0^\pi |x(x - \pi/2)(x - \pi)| = \frac{1}{6} \cdot \pi^4 \cdot \frac{1}{32} = \frac{\pi^4}{192} \approx 0.507$$

- We can find an alternative error estimate via the following procedure

  (i) We see that if g is a polynomial of degree 2, g equals its Lagrange interpolation at x_0, x_1, x_2, hence $I(g) = \int_0^\pi gdx$.

  (ii) Furthermore, if $f_3 = x(x - \pi/2)(x - \pi)$, $I(f_3) = 0 = \int_0^\pi f_3 dx$.

  (iii) Hence, if g_3 is a polynomial of degree 3, then $g_3 = af_3 + g$ for some $a \in \mathbb{R}$, and g is a polynomial of degree 2. Hence

$$I(g_3) = aI(f_3) + I(g) = \int_0^\pi g(x)dx = \int_0^\pi (af_3 + g)dx = \int_0^\pi g_3 dx$$

  (iv) Now Let p be the Lagrange interpolation polynomial of f at $x_0, x_1, x_2, x_3 = c$, then $\int_0^\pi pdx = I(p) = I(f)$.

(v) The error estimate for Lagrange interpolation gives us

$$|f(x) - p(x)| = \left| \frac{f''''(s)x(x-c)(x-\pi/2)(x-\pi)}{4!} \right|$$

$$\leq \frac{1}{24}|x(x-c)(x-\pi/2)(x-\pi)|$$

(vi) Integrate the error estimate, let $c \to \pi/2$, we get

$$|\int_0^\pi f(x)dx - I(f)| \leq \frac{1}{24}\int_0^\pi |x(x-\pi/2)^2(x-\pi)|dx = \frac{\pi^5}{2880} \approx 0.106$$

This is the bound in Theorem 7.2 in the textbook.

- The actual error is $2\pi/3 - 2 = 0.094$.

Exercises 1: Let $f(x) = x^3(1-x)$, $x \in [0,1]$. $x_0 = 0$, $x_1 = c$, $x_2 = 1$, $c \in (0,1)$.

- Find the Lagrange interpolation of f at x_0, x_1, x_2.

- Integrate the Lagrange interpolation on $[0,1]$ to get an estimate of $\int_0^1 f(x)dx$.

- Find c such that the estimate you found above is optimal.

Exercise 2: Let $[a,b] = [0,1]$, $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$.

- Write down the formula for Simpson's rule and the composite trapezium rule using x_i.

- Find a continuous function on $[a,b]$ where the Simpson's rule works better than the composite trapezium rule, and vice versa.

Answer:

Exercises 1:

- $p = c^3(1-c) \cdot \frac{x(x-1))}{c(c-1)} = -c^2 x(x-1)$.

- $I = c^2/6$.

- $\int_0^1 f(x)dx = 1/20$, so c should be $\sqrt{0.3}$.

Exercise 2:

- Simpson's rule: $f(0)/6 + 2f(0.5)/3 + f(1)/6$. Composite trapezium rule: $f(0)/4 + f(0.5)/2 + f(1)/4$.

- $f(x) = x^2$. $f(x) = |x - 0.5|$.

Correction for the lecture on 11/6: We can also estimate the integral by decomposing the interval into n subintervals with the same length, and calculate the sum of the areas of rectangles with those subintervals as base. If the height is taken as the value of the function at end points (i.e. $I = \frac{b-a}{n} \sum_{i=0}^{n-1} f(x_i)$), or use the maximum or minimum on the interval as in the definition of Riemann integration, then the error grows as $O(1/n)$ if f is smooth, where n is the number of subintervals, which is worse than composite trapezium rule. However, if the height is taken as the value of the function at midpoints (i.e. $I = \frac{b-a}{n} \sum_{i=0}^{n-1} f(\frac{x_i + x_{i+1}}{2})$) the error grows as $O(1/n^2)$ if f is smooth (so it's about as accurate as the trapezium rule).

3.5 Gauss quadrature (Also see Sections 10.2-10.4 in the textbook)

Motivation:

(i) From the error bound of composite rules, we know that if the quadrature has error bound $O(|b-a|^k)$, the composite rule will have an error bound of $O(n^{-k+1})$.

(ii) So suppose f is sufficiently smooth, it might be good to try and make the number k as large as possible.

(iii) From the proof of Theorem 3.5, we see that if a quadrature rule gives accurate answer to polynomials of degree d, then the error bound is $O(|b-a|^{d+2})$. For example, for Simpson's rule $d = 3$.

(iv) So, it may be good to **strategically choose the quadrature points** such that the quadrature rule works for polynomials of high degrees.

Recall from the definition of Legendre polynomial, if L_j are the Legendre polynomials, then

$$L_j \perp span\{L_0, L_1, \ldots L_{j-1}\}$$

Under $L^2([-1,1])$ norm.

Theorem 3.9. *Let $I(f)$ be the result of quadrature rule on $[-1,1]$, using quadrature points x_i, $i = 0, \ldots, n$, which the roots of L_{n+1}. For any polynomial g of degree no more than $2n + 1$, $I(g) = \int_{-1}^{1} g \, dx$.*

Proof. It's easy to see that

$$span\{1, x, \ldots x^{2n+1}\} = span\{L_0, \ldots L_{n+1}, L_{n+1}L_1, \ldots, L_{n+1}L_n\}$$

(A way to see it is by first recognizing that $span\{1, x, \ldots x^{2n+1}\}$ has dimension $2n + 2$, and show, from definition, that

$$\{L_0, \ldots L_{n+1}, L_{n+1}L_1, \ldots, L_{n+1}L_n\}$$

is a linearly independent set of $2n + 2$ elements (because they all have different degrees) hence must be a basis.)

Because both I and $\int_{-1}^{1}$ are linear on the vector space consisting of polynomials of degree no more than $2n + 1$, and two linear transformations are the same if and only if they are identical on the basis vectors, we only need to prove it when g is L_j, $0 \leq j \leq n$, as well as when g is $L_j L_{n+1}$, $0 \leq j \leq n$.

- Case 1: $g = L_j$ for some $j \leq n$. In this case, p is of degree no more than n, hence p is identical to its Lagrange interpolation at $n + 1$ points. Hence $I(g) = \int_{-1}^{1} g \, dx$.

- Case 2: $g = L_j L_{n+1}$, $j \leq n$. Note that L_{n+1} is proportional to $L_{n+1} L_0$ as L_0 is of degree 0 hence a constant. Because g has a factor L_{n+1}, $g(x_i) = 0$ for all i, hence $I(g) = 0$. On the other hand, because L_{n+1} and L_j are orthogonal on $L^2([-1, 1])$, $\int_{-1}^{1} g \, dx = 0$

$\square$

Definition 3.10. *The Gauss-Legendre quadrature points $x_0, \ldots, x_n$ on $[a, b]$ is defined as $x_j = \frac{a+b}{2} + c_j \frac{b-a}{2}$, where c_j is the $j+1$-th root of the $n+1$-th Legendre polynomial. In other words, x_j is the $j+1$-th root of the weight-1 orthogonal polynomial on $[a, b]$ with index $n+1$, ψ_{n+1}.*

Now we can use Theorem 3.9 to prove an error bound for Gauss-Legendre quadrature:

Theorem 3.11. *(Theorem 10.1 in textbook) Let $f \in C^{2n+2}$, then there is some $c \in [a, b]$ such that*

$$\int_a^b f dx - I(f) = \frac{f^{(2n+2)}(c)}{(2n+2)!} \int_a^b \prod_i (x - x_i)^2 dx$$

Proof. Let H be the Hermite interpolation polynomial of f at x_i, then Theorem 3.9 implies that $\int_a^b H dx = I(H) = I(f)$. Hence the result follows from the error bound of Hermite interpolation (Theorem 1.10).

More precisely, if M and m are the upper and lower bound of $f^{(2n+2)}$ on $[a, b]$ respectively, we have

$$\frac{m \prod_i (x - x_i)^2}{(2n+1)!} \leq f(x) - H(x) \leq \frac{M \prod_i (x - x_i)^2}{(2n+1)!}$$

Hence

$$\frac{m}{(2n+2)!} \int_a^b \prod_i (x - x_i)^2 dx \leq \int_a^b f dx - I(f) \leq \frac{M}{(2n+2)!} \int_a^b \prod_i (x - x_i)^2$$

And the existence of c follows from intermediate value theorem in analysis. $\qquad\square$

Remark 3.12. *As $|b - a| \to 0$ the error decay at $|b - a|^{2n+3}$, which is better than the $|b - a|^{n+2}$ or $|b - a|^{n+3}$ in Newton-Cotes.*

Example 3.13. *Use $n = 1$ Gauss-Legendre quadrature to estimate $\int_0^\pi \sin(x)dx$.*

Firstly find the quadrature points. The first method is by using the formula of Legendre polynomials in HW4 problem 4, which is $(1 - x^2)^{2''} = 12x^2 - 4$, so roots are $\pm\frac{1}{\sqrt{3}}$, so $x_0 = \pi/2(1 - 1/\sqrt{3}) = 0.6639$, $x_1 = \pi/2(1 + 1/\sqrt{3}) = 2.4777$.

The second method is by finding a monic quadratic polynomial orthogonal to both 1 and x under $L^2([0, \pi])$. In other words, we need a and b such that

$$\int_0^\pi (x^2 + ax + b)dx = \frac{\pi^3}{3} + \frac{a\pi^2}{2} + b\pi = 0$$

$$\int_0^\pi (x^2 + ax + b)x\,dx = \frac{\pi^4}{4} + \frac{a\pi^3}{3} + \frac{b\pi^2}{2} = 0$$

Solve for a and b then find the roots.

The quadrature weights are

$$w_0 = \int_0^\pi \frac{x - x_1}{x_0 - x_1} dx = \frac{x_1 \pi - \pi^2/2}{x_1 - x_0} = \pi/2 = 1.5708$$

$$w_1 = \int_0^\pi \frac{x - x_0}{x_1 - x_0} dx = \frac{\pi^2/2 - x_0 \pi}{x_1 - x_0} = \pi/2 = 1.5708$$

So

$$I_1(\sin) = w_0 \sin(x_0) + w_1 \sin(x_1) = 1.9358$$

Error is 0.064.

Another key property of the Gauss-Legendre quadrature is that all its weights are positive, because we have

Theorem 3.14. *The weight for Gauss-Legendre quadrature is*

$$w_k = \int_a^b \frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} dx = \int_a^b \left(\frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} \right)^2 dx$$

Proof. Suppose $\psi_{n+1} = c \prod_i (x - x_i)$ is the weight 1 orthogonal polynomial on $[a, b]$ with index $n + 1$, then

$$\int_a^b \frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} - \left(\frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} \right)^2 dx$$

$$= \frac{1}{(\prod_{j \neq k}(x_k - x_j))^2} \int_a^b \prod_{j \neq k}(x - x_j)(\prod_{j \neq k}(x_k - x_j) - \prod_{j \neq k}(x - x_j))dx$$

Let $h(x) = \prod_{j \neq k}(x_k - x_j) - \prod_{j \neq k}(x - x_j)$, then $\deg(h) = n$, and $h(x_k) = 0$, hence $h = (x - x_k)h_1$ where $\deg(h_1) = n - 1$. Now we have

$$\int_a^b \prod_{j \neq k}(x - x_j)(\prod_{j \neq k}(x_k - x_j) - \prod_{j \neq k}(x - x_j))dx = \int_a^b (\psi_{n+1} \cdot \frac{h_1}{c})dx = 0$$

$\square$

Remark 3.15. *For comparison, when $n = 8$, the Newton-Cotes quadrature weights are not all positive.*

As a consequence, we have:

Theorem 3.16. *(Theorem 10.2 in textbook) Given any continuous function f, the Gauss-Legendre quadrature result $I_n(f)$ converges to $\int_a^b f dx$ as $n \to \infty$.*

Proof. By Weierstrass approximation theorem there is some p such that $|f - p| < \epsilon$ on $[a, b]$, hence for any $n > \deg(p)$,

$$|\int_a^b f dx - I_n(f)| \leq |\int_a^b f dx - \int_a^b p dx| + |\int_a^b p dx - I_n(p)| + |I_n(p) - I_n(f)|$$

$$\leq \epsilon(b - a) + 0 + \epsilon(b - a)$$

The last bit is due to

$$|I_n(p) - I_n(f)| = |\sum_i w_i(p(x_i) - f(x_i))| \leq \epsilon \sum_i |w_i| = \epsilon \sum_i w_i$$

And $I_n(1) = \int_a^b 1 dx = b - a$ so $\sum_i w_i = b - a$. Now let $\epsilon \to 0$ we get the convergence. $\quad\square$

If w_i are not all positive, one can not remove the absolute value, hence this may not be true for other quadrature rules.

Remark 3.17. *The Gauss-Legendre quadrature can be generalized to deal with the problem of estimating $\int_a^b fwdx$ where w is a given weight function not depend on f, by replacing Legendre polynomials with other orthogonal polynomials, which are called "Gauss quadrature".*

Definition 3.18. *The Gauss quadrature points with weight w, $x_0, \ldots, x_n$ on $[a, b]$ is defined as x_j is the $j + 1$-th root of the weight-w orthogonal polynomial on $[a, b]$ with index $n + 1$, ψ_{n+1}. The quadrature weights are chosen as $w_k = \int_a^b w \frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} dx$.*

For example, if $[a, b] = [-1, 1]$, and w is $(1 - x^2)^{-1/2}$ we call it Chebyshev-Gauss quadrature.

Theorem 3.19.   *(i)* $w_k = \int_a^b w \left(\frac{\prod_{j \neq k}(x - x_j)}{\prod_{j \neq k}(x_k - x_j)} \right)^2 dx$ *hence is always positive.*

(ii) If f is a polynomial of degree no more than $2n + 1$, then $\int_a^b wfdx = I_n(f)$, where I is the Gauss quadrature rule with $n + 1$ quadrature points.

(iii) If f is continuous on $[a, b]$, $\lim_{n \to \infty} I_n(f) = \int_a^b wfdx$ (Theorem 10.2 in textbook).

(iv) If $f \in C^{2n+2}$, then there is some $c \in [a, b]$, where $\int_a^b wfdx - I(f) = \frac{f^{(2n+2)}(c)}{(2n+2)!} \int_a^b w \prod_i (x - x_i)^2 dx$. (Theorem 10.1 in textbook)

The proof is very similar to the Gauss-Legendre case.

3.6 Friday Review and Examples

- Definition of Gauss Legendre quadrature.

- Error bound.

- Convergence.

- General Gauss quadrature.

Key ideas:

- Error bound for Lagrange and Hermite interpolation.

- Quadrature and integration as linear transformation.

- Inner product and orthogonal polynomials.

Example 3.20. $I = \int_0^\pi \frac{1}{\sqrt{\sin x}} dx$. *The correct answer is 5.2441151*

- By change of variable,

$$I = \frac{\pi}{2} \int_{-1}^{1} \frac{1}{\sqrt{\cos(\pi x/2)}} dx$$

- Let $w = (1-x^2)^{-1/2}$, $f = \frac{(1-x^2)^{1/2}}{\sqrt{\cos(\pi x/2)}}$. Then $\int_{-1}^{1} \frac{1}{\sqrt{\cos(\pi x/2)}} dx = \int fw dx$.

- Now do Chebyshev-Gauss formula for f. Suppose $n = 2$, then the quadrature points, which are the roots of $T_3 = \cos(3\cos^{-1} x)$, are $0, \pm\frac{\sqrt{3}}{2}$.

- The quadrature weights are

$$w_1 = \int_{-1}^{1} \frac{(x^2 - 3/4)/(-3/4)}{\sqrt{1-x^2}} dx = \pi/3$$

$$w_0 = w_2 = \pi/3$$

(One can check that for Chebyshev-Gauss, the weights are always $\frac{\pi}{n+1}$.)

- The Chebyshev-Gauss for $n = 2$ is

$$I_2 = \sum_{i=0}^{2} w_i f(x_i) = 3.3384$$

- The final answer is 5.2439, difference is 0.00018.

Exercises:

(i) (a) Write down the Lagrange interpolation polynomial p of a function f, with interpolation points -1, 0 and 1.

   (b) Calculate $\int_0^1 p(x)dx$, write it into the form of $w_0 f(-1) + w_1 f(0) + w_2 f(1)$.

   (c) Show that the formula $I(f) = w_0 f(-1) + w_1 f(0) + w_2 f(1)$ will give accurate answer if f is a polynomial of degree 2.

   (d) Use the error estimate for Lagrange interpolation polynomial, find an upper bound for $\int_0^1 f(x)dx - I(f)$.

(ii) (a) Let $w = e^{-x}$, $L_w^2([0, \infty))$ be the space of functions f such that $\int_0^\infty w f^2 dx$ exists. Show that $1, x, x^2 \in L_w^2([0, \infty))$.

   (b) Use the inner product $(f, g) = \int_0^\infty w f g dx$, carry out Gram-Schmidt process for $\{1, x, x^2\}$ (the results are called **Laguerre polynomials**).

   (c) Find the roots of the third polynomial you get from the above step, call them x_0 and x_1.

   (d) Find w_0 and w_1 such that $\int_0^\infty w g dx = w_0 g(x_0) + w_1 g(x_1)$ for any polynomial g of degree 1. This is called the Gauss-Laguerre quadrature.

74

Answer:

(i) (a) $p(x) = f(-1)(x(x-1)/2) + f(0)(1-x^2) + f(1)(x(x+1)/2)$

(b) $-f(-1)/12 + 2f(0)/3 + 5w_2 f(1)/12$.

(c) Because in this case, due to uniqueness of Lagrange interpolation, $f = p$.

(d) The error bound for Lagrange interpolation polynomial is $|f(x)-p(x)| \leq \max |f'''||x^3 - x|/3!$, so the error bound for I is $\max |f'''|/24$ Here maximum is taken on $[-1, 1]$.

(ii) Note that $\int_0^\infty x^n e^{-x} dx = n!$

(a) Because the weighted L^2 norms of them are $1, 2, 24$ respectively.

(b) $\{1, x-1, x^2 - 4x + 2\}$

(c) $2 \pm \sqrt{2}$.

(d)
$$\frac{1+\sqrt{2}}{2\sqrt{2}} g(2-\sqrt{2}) + \frac{\sqrt{2}-1}{2\sqrt{2}} g(2+\sqrt{2})$$

3.7 Composite Gauss quadrature (Section 10.5)

One can combine composite method and Gauss quadrature: divide the interval into m subintervals of equal length, then apply Gauss quadrature on each. For example, if we apply Gauss-Legendre quadrature with $k+1$ quadrature points to each subinterval, when the function is in C^{2k+2}, the error decay at a speed of $O(m^{-2k-2})$.

Example 3.21. $\int_{-0.5}^{0.5} \sqrt{1 - x^2} dx$, *using composite Gauss quadrature with* $k = 1$

Roots of degree 2 Legendre polynomials are $\pm 1/\sqrt{3}$.

```
def composite_gauss(n, a, b, f):
    r=0
    m=int((n+1)/2)
    for i in range(m):
        l0=(i*b+(m-i)*a)/m
        l1=((i+1)*b+(m-i-1)*a)/m
        x0=(l0+l1)/2-(l1-l0)/2/(3**0.5)
        x1=(l0+l1)/2+(l1-l0)/2/(3**0.5)
        r+=f(x0)+f(x1)
    return r*(b-a)/(n+1)
```
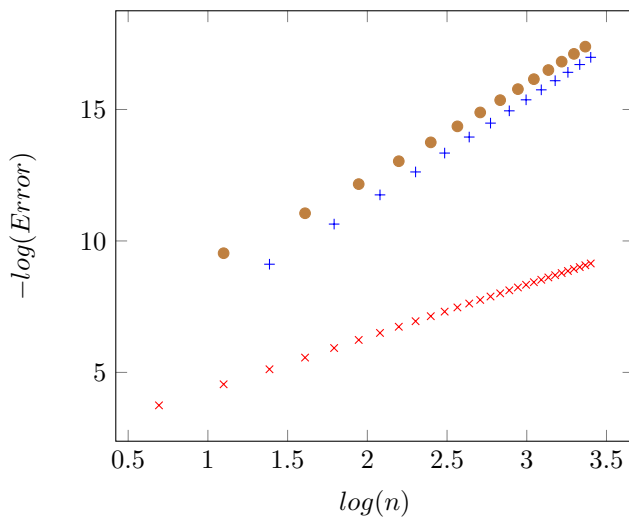
3.8   Other topics in numerical integration

3.8.1   Modified Gauss Quadratures (Section 10.6), won't be in final exam

Sometimes some quadrature points are pre-determined while others can be chosen strategically, in which case we can choose them as roots of orthogonal polynomials. If the pre-determined point is one of the end point we call this strategy **Radau quadrature**, if it is both end points we call it **Lobatto quadrature**.

If there are $n + 1$ quadrature points, k of them are predetermined and the rest roots of Legendre polynomial of degree $n + 1 - k$, the error is bounded by $O((b - a)^{2n+3-k})$. When $n = k = 2$ we get Simpson's method.

3.8.2 Richardson Extrapolation (Section 7.6, 7.7), won't be in final exam

A common trick in numerical analysis is **extrapolation**: if a sequence a_n can be calculated whose limit as $n \to \infty$ is some a, and the speed of convergence is known, we can use linear combination of successive terms to speed up the convergence. For example, for composite trapezium rule, Euler-Maclaurin formula says that, when f is "good enough",

$$I_n - I = \sum_{k=1}^{\infty} c_k n^{-2k}$$

So $\frac{4I_{2n} - I_n}{3}$ (which is identical to composite Simpson's rule) has $O(n^{-4})$ convergence, and one can do it repeatedly to get higher convergence speed.

3.9 Review

- Quadrature rule

- Newton-Cotes quadrature

- Gauss quadrature

- Composite method

4 Numerical ODE: IVP (Chapter 12)

Initial value problem of ODE:

$$y' = f(t, y), y(0) = y_0$$

Here y can be a real valued function or $\mathbb{R}^n$-valued function. For now we focus on the case where y is real valued.

We always assume f is continuous and Lipschitz on the second parameter ($|f(t, y) - f(t, z)| \leq L|y - z|$ for all pair y, z, L is called the Lipschitz constant). So by Picard's Theorem, IVP has a unique solution.

Numerical integration can be seen as a special case of IVP of numerical ODE, with f independent from y.

Strategy: Find some small positive number h as the "step size", estimate the value of y at nh for all n.

4.1 Euler's Method (12.2, 12.3)

h is a small number, z an approximation of y evaluated using:

$$z(0) = y_0$$

$$z((n+1)h) = z(nh) + hf(nh, z(nh))$$

- The motivation is using PL function to approximate y.

- **Truncated error** T_n is error introduced at step n, divided by step size. More precisely, suppose the method is $z((n+1)h) = G(z(nh))$, then

$$T_n = \frac{G(y(nh)) - y((n+1)h)}{h}$$

.

- A method has **order of accuracy** p if when f is sufficiently smooth, $T = O(h^p)$.

- **Global error at time** t is the difference between the estimated $z(t)$ and the true value of $y(t)$.

- A method is called **consistent** if truncated error goes to 0 as $h \to 0$. **Convergent** if global error goes to 0 as $h \to 0$.

- Under certain assumptions, global error at given time t is controlled by the bound on truncated error (Theorem 12.2 and 12.5 in textbook).

Example 4.1. *Euler's method applied to* $y' = y$, $y(0) = 1$

True solution: $y(t) = e^t$.

Approximated solution using Euler's method with step-size h:

$$z(0) = 1, z(h) = 1 + hz(0) = 1 + h, z(2h) = z(h) + hz(h) = 1 + h + h(1+h) = (1+h)^2, \ldots, z(nh) = (1+h)^n$$

So the global error at time $t = nh$ is $e^t - (1+h)^n = e^t - (1+h)^{t/h}$, which converges to 0 as $h \to 0$.

To understand the error created at each step, consider the sequence $z_k(nh)$, such that
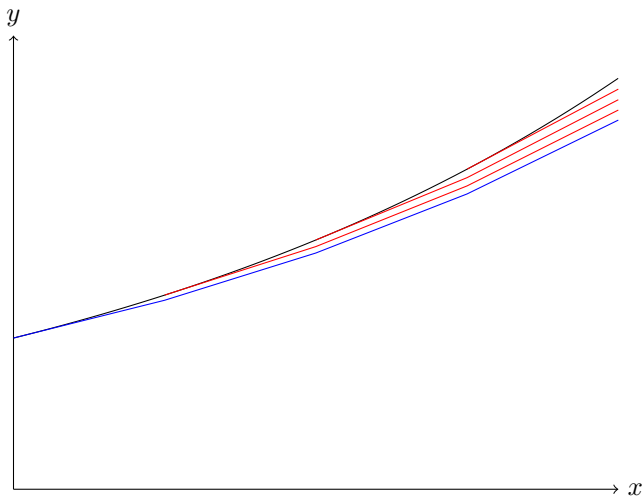
$$z_k(kh) = y(kh) = e^{kh}$$

$$z_k((n+1)h) = z_k(nh) + hz_k(nh)$$

So

$$z_k(nh) = e^{kh}(1+h)^{n-k}$$

$$T_n = \frac{e^{(n+1)h} - e^{nh}(1+h)}{h} = O(h)$$

So the method is consistent with order of accuracy 1.

To study global error, we need to understand:

- What are the truncated errors?

- How do error introduced in prior steps grow with time?

- By definition of derivative, if $t = nh$,

$$y(t + h) = y(t) + hf(t, y(t)) + o(h)$$

So $T_n = \frac{o(h)}{h}$, Euler's method is **consistent**.

- If f is sufficiently smooth, so is y, so

$$y(t + h) = y(t) + hf(t, y(t)) + \frac{h^2}{2}y''(t) + o(h^2)$$

So $T_n = O(h)$, Euler's method is of order of accuracy 1.

- Furthermore, if we write down the remainder of Taylor's series, we have

$$y(t + h) = y(t) + hy'(t) + \frac{h^2}{2}y''(s)$$

$$= y(t) + hf(t, y(t)) + \frac{h^2}{2}(f_1(s, y(s)) + f_2(s, y(s))f(s, y(s)))$$

for some $s \in [t, t+h]$, f_1 and f_2 are the partial derivative in first and second parameter. So, if f and its partial derivatives are all bounded, T is uniformly bounded by some Ch.

Let's now analyze how error grows in Euler's method:

Lemma 4.2. *If f is L-Lip. on second paramater, w_1, w_2 two functions defined on $mh, (m+1)h, \ldots, nh$, such that*

$$w_i((k+1)h) = w_i(kh) + hf(kh, w_i(kh))$$

Then

$$|w_1(nh) - w_2(nh)| \leq e^{L(n-m)h}|w_1(mh) - w_2(mh)|$$

Proof.

$$LHS \leq (1 + hL)|w_1((n-1)h) - w_2((n-1)h)| \leq$$
$$\cdots \leq (1 + hL)^{n-m}|w_1(mh) - w_2(mh)|$$

And

$$(1 + hL)^{n-m} \leq e^{L(n-m)h}$$

$\square$

Now let's estimate global error at time $t = nh$. Let $z_k(nh)$, $n \geq k$, be

$$z_k(kh) = y(kh)$$

$$z_k((n+1)h) = z_k(nh) + hf(nh, z_k(nh))$$

$$|y(nh) - z(nh)| = |z_n(nh) - z_0(nh)| \leq \sum_{k=0}^{n-1} |z_k(nh) - z_{k+1}(nh)|$$

(Triangle inequality)

$$\leq \sum_{k=0}^{n-1} e^{Lh(n-k-1)} |z_k((k+1)h) - z_{k+1}((k+1)h)|$$

(The Lemma from previous page)

$$\leq \sum_{k=0}^{n-1} Ch^2 e^{L(n-k-1)h} \leq \frac{Ch^2(e^{Lnh} - 1)}{e^{Lh} - 1}$$

(Bound on truncated error)
So

Theorem 4.3. *If*

- *f is smooth, f and its partial derivatives are bounded.*

- *nh is fixed and $n \to \infty$*

then the global error of Euler's method at time nh converges to zero at $O(h)$. In other words, Euler's method is **convergent**.

4.2   Friday Review and Examples

- Composite Gauss Quadrature

- Euler's method.

- Truncated error and global error.

- Consistency and order of accuracy.

- Convergence.

Example 4.4. *Midpoint rule*

(i) The weight 1 orthogonal polynomial on $[a, b]$ of degree 1. It is $x - \frac{b+a}{2}$. The root of it is $\frac{b+a}{2}$.

(ii) Hence, the Gauss-Legendre quadrature with $n = 0$ is $x_0 = \frac{b+a}{2}$. Quadrature weight is $b - a$.

(iii) If f is differentiable, the error bound is

$$|\int_a^b f dx - (b-a)f(\frac{a+b}{2})| \leq \frac{\max |f''| \int_a^b (x - \frac{a+b}{2})^2 dx}{2!} = \frac{\max |f''|(b-a)^3}{24}$$

(iv) Now decompose f into m subintervals, each applying the Gauss-Legendre quadrature, we get

$$I = \frac{b-a}{m} \sum_{i=0}^{m-1} f(a + \frac{(i+1/2)(b-a)}{m})$$

And the error bound is $\frac{\max |f''|(b-a)^3}{24m^2}$.

Example 4.5. $y' = \sin(y)$, $y(0) = 1$

  (i) True solution: $y(t) = 2\tan^{-1}(\tan(1/2)e^x)$

 (ii) Euler's method is:

$$z(0) = 1, z(nh) = z((n-1)h) + h\sin(z((n-1)h))$$

(iii) Truncated error for Euler's method at time t is

$$|\frac{y''(s)h}{2!}| = |\frac{y'\cos(y)h}{2!}|$$

$$= |\frac{\sin(y)\cos(y)h}{2!}| \leq Ch$$

   What is the number C?

 (iv) $\sin(y)$ is L-Lipschitz. In other words, $|\sin(a) - \sin(b)| \leq L|a - b|$ for all a, b. What's a valid L?

  (v) Now the argument from last lecture shows that the global error bound at time $t = nh$ is $\frac{Ch^2(e^{Lt}-1)}{e^{Lh}-1}$.

 (vi) Let $t = 1$, $h = 0.1$, Euler's method get $z(1) = 1.95109$. True answer is $y(1) = 1.95629$. Error bound as calculated from above, using $C = 1/4, L = 1$, is 0.04084.

89

Example 4.6. *Consider IVP $y' = f(y)$, $y(0) = y_0$. Suppose f is **real analytic**, i.e. the Taylor series at every point converges at a neighborhood of the point. Suppose h is a small positive number. The theory of ODE tells us that y is also real analytic.*

Suppose $f(x) = \sum_i a_i (x - y_0)^i$.

(i) Write down the Taylor expansion of y at $t = 0$, up to t^2 term.

(ii) Write down the result of $z(h) = z(0) + hf(z(0))$, as a power series of h.

(iii) Write down the result of $z(2h) = z(h) + hf(z(h))$, as a power series of h.

(iv) Find a linear combination of y_0, $z(h)$, $z(2h)$ which is close to $y(2h)$ up to the h^2 term. This is called the 2nd order Runge-Kutta method (rk2).

Answer:

(i) $y(t) = y_0 + a_0 t + \frac{a_0 a_1}{2} t^2 + O(t^3)$

(ii) $z(h) = y_0 + a_0 h$

(iii) $z(2h) = y_0 + 2a_0 h + a_0 a_1 h^2 + O(h^3)$

(iv) $y(2h) = y_0 - 2z(h) + 2z(2h) + O(h^3)$

90

4.3   Ways to get higher order methods

4.3.1   Method based on Lagrange Interpolation (12.4, 12.6)

Firstly some methods based on Lagrange interpolation:

Explicit Methods   Consider the function $g(t) = f(t, y(t))$. One can then use $g(t - h), \ldots, g(t - kh)$ as quadrature point to estimate $\int_{t-dh}^{t} g(s)ds$, then use $y(t) = y(t - dh) + \int_{t-dh}^{t} g(s)ds$.

Implicit Methods   One can also use $g(t), g(t - h), \ldots, g(t - kh)$ as quadrature point to estimate $\int_{t-dh}^{t} g(s)ds$, then use $y(t) = y(t - dh) + \int_{t-dh}^{t} g(s)ds$. This way $y(t)$ appears on the right hand side and the estimate of $y(t)$ requires solving an equation, hence is called **implicit**.

Example 4.7. • *Explicit method for $d = 1$, $k = 1$ is Euler's Method.*

• *Implicit method for $d = 1$, $k = 1$ is the* **trapezium method** *as the numerical integration is via trapezium rule:*

$$z(t) = z(t - h) + \frac{h}{2}(f(t, z(t)) + f(t - h, z(t - h)))$$

We can show that it is consistent with order of accuracy 2.

• *Implicit method for $k = d = 2$ will be*

$$z(t) = z(t - 2h) + \frac{h}{3}(f(t, z(t)) + 4f(t - h, z(t - h)) + f(t - 2h, z(t - 2h)))$$

When $f(t, y) = f(t)$ this reduces to Simpson's rule. This is called the 2-step **Milne's method**.

• *When $d = 1$, the explicit methods are called k-th step* **Adams-Bashforth** *methods, while the implicit methods are called* **Adams-Moulton** *methods.*

Example 4.8. *Let's deduce the Adams-Bashforth method for $k = 4$.*

(i) The Lagrange interpolation of $f(s, z(s))$ at $t - h$, $t - 2h$, $t - 3h$, $t - 4h$ is

$$p(s) = f(t - h, z(t - h)) \frac{(s - (t - 2h))(s - (t - 3h))(s - (t - 4h))}{6h^3}$$

$$+ f(t - 2h, z(t - 2h)) \frac{(s - (t - h))(s - (t - 3h))(s - (t - 4h))}{-2h^3}$$

$$+ f(t - 3h, z(t - 3h)) \frac{(s - (t - h))(s - (t - 2h))(s - (t - 4h))}{2h^2}$$

$$+ f(t - 4h, z(t - 4h)) \frac{(s - (t - h))(s - (t - 2h))(s - (t - 3h))}{-6h^3}$$

(ii)

$$z(t) = z(t - h) + \int_{t-h}^{t} p(s) ds$$

$$= z(t - h) + h(\frac{55}{24} f(t - h, z(t - h)) - \frac{59}{24} f(t - 2h, z(t - 2h))$$

$$+ \frac{37}{24} f(t - 3h, z(t - 3h)) - \frac{3}{8} f(t - 4h, z(t - 4h)))$$

93

An alternative to implicit method is the **predictor-corrector method**: for example, instead of Trapzsium method, we first estimate $z(t)$ using Euler's method, then "correct" it using the trapezium rule formula, and get:

$$z_p(t) = z(t - h) + hf(t - h, z(t - h))$$

$$z(t) = z(t - h) + \frac{h}{2}(f(t, z_p(t)) + f(t - h, z(t - h)))$$

This is called **Heun's method**.

Example 4.9. $y' = y$, $y(0) = 1$

- *Euler's method:*

$$z(nh + h) = z(nh) + hz(nh)$$

$$z(nh) = (1 + h)^n$$

- *Trapezium rule method:*

$$z(nh + h) = z(nh) + \frac{h}{2}(z(nh + h) + z(nh))$$

$$z(nh) = \frac{(1 + h/2)^n}{(1 - h/2)^n}$$

Truncated error is

$$O(\frac{e^h - \frac{1+h/2}{1-h/2}}{h}) = O(h^2)$$

- *Heun's rule is*

$$z(nh + h) = z(nh) + \frac{h}{2}(2z(nh) + hz(nh))$$

$$z(nh) = (1 + h + h^2/2)^n$$

Truncated error is also $O(h^2)$.

4.3.2 Theory of General Linear Multistep Methods (12.7-12.9)

General Linear k-step Method:

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

If $\beta_k = 0$ it is explicit, otherwise it is implicit. To start the k-step method, we need k initial values $z(0), \ldots, z((k-1)h)$, then solve the equation to get $z(kh), z((k+1)h, \ldots$

- We want a linear k-step method to be **zero-stable**. In other words, if the equation is $y' = 0$, then the $z(nh)$ does not go to infinity as $n \to \infty$. From the theory in linear difference equations in linear algebra, we know that this is equivalent to the **first characteristic polynomial** $\rho(z) = \sum_{j=0}^{k} \alpha_j z^j$ having all roots inside the closed unit disc and at most only single roots on the unit circle.

- All the method obtained via Lagrange interpolation, including Adams-Moulton and Adams-Bashforth, are zero-stable.

- Why do we need zero-stability? Suppose we let $h \to 0$ and $t = hn$ remain unchanged, then $n \to \infty$. If there is no zero stability, error at previous steps will grow indefinitely.

Example 4.10. $y' = y$, $y(0) = 1$. *Use multistep method* $z((n+2)h) - 3z((n+1)h) + 2z(nh) = -hf(nh, z(nh))$

(i) Start with $z(0) = 1$, $z(h) = e^h$.

(ii) One can verify by doing Taylor expansion of z, that the method is consistent.

(iii) Similar to Example 4.1, define $z_k(nh)$ be $z_k(kh) = y(kh)$, $z_k((k+1)h) = y((k+1)h)$, and for all $n > k+1$, $z_k(nh)$ are calculated using the multistep method.

(iv) Then the error created at time $(k+2)h$ is

$$z_k((k+2)h) - y((k+2)h) = z_k((k+2)h) - z_{k+1}((k+2)h) = O(h^2)$$

(v) Now let's see how this error grow with time: Let $w_m = z_k((k+1+m)h) - z_{k+1}((k+1+m)h)$. Then

$$w_0 = 0$$

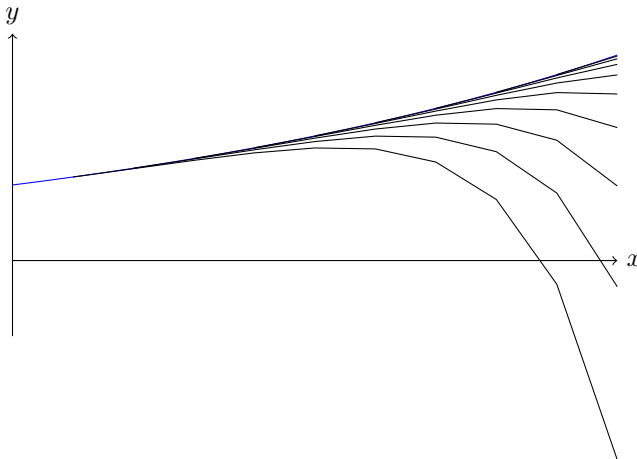$$w_1 = O(h^2)$$

$$w_{m+2} - 3w_{m+1} + 2w_m = O(h)$$

(vi) So if the linear difference equation

$$w_{m+2} - 3w_{m+1} + 2w_m = 0$$

Has solution that grows to infinity, then the error created at time $(k+2)h$, after propagating to $t = nh$, will be w_{n-k-1} which goes to ∞ as $h \to 0$.

This is when $t = 1$, $h = 0.1$ (See https://github.com/wuchenxi/Math-514/blob/main/zero_stability.py):

- To get consistency, let $\sigma(z) = \sum_{j=0}^{k} \beta_j z^j$ be the **second characteristic polynomial**. When f is smooth, $s << 1$, $f(nh + s, y(nh + s)) = f(nh, y(nh)) + O(s)$, $y(nh + s) = y(nh) + sf(nh, y(nh)) + O(s^2)$. To make method consistent, we want to make sure that if we put $y((n + j)h)$ in place of $z((n + j)h)$, the left hand side and right hand side are off by $o(h)$, hence

$$\sum_i \alpha_i + \sum_i i\alpha_i f(nh, y(nh))h = h \sum_i \beta_i f(nh, y(nh)) + O(h^2)$$

So $\rho(1) = \sum_i \alpha_i = 0$, $\rho'(1) = \sum_i i\alpha_i = \sum_i \beta_i = \sigma(1)$.

- All the methods obtained via Lagrange interpolation are consistent, by looking at $f = 1$.

- To get order of accuracy, carry out the same argument as above but do higher order power expansion for $y(nh + s)$ and $f(nh + s, y(nh + s))$.

For global convergence of linear multistep methods we have the Dahlquist's theorems:

- If a linear k-step method has zero stability, then it is consistent iff it is convergent, and the truncated error and global error has the same order as $h \to 0$.

- (First Dahlquist barrier) If a linear k-step method is 0-stable then the order of accuracy is no more than $k + 1$ if k is odd (e.g. Adams-Moulton, by Theorem 3.4), $k + 2$ if k is even (e.g. $d = k$ implicit, by Theorem 3.5), and k if it has to be explicit (e.g. Adams-Bashforth, by Theorem 3.4).

4.3.3 Runge-Kutta methods (12.5)

To get initial conditions for linear multistep methods, we need one step methods with higher orders of accuracy. To accomplish that we need to have more evaluations of f per step.

Recall Heun's method:

$$z_p((n+1)h) = z(nh) + hf(nh, z(nh))$$

$$z((n+1)h) = z(nh) + \frac{h}{2}(f(nh, z(nh)) + (f((n+1)h, z_p((n+1)h))))$$

We can rewrite it as

$$k_1 = f(nh, z(nh))$$

$$k_2 = f(nh + h, z(nh) + hk_1)$$

$$z((n+1)h) = z(nh) + h(\frac{k_1}{2} + \frac{k_2}{2})$$

General Runge-Kutta:

$$k_1 = f(nh, z(nh))$$

$$k_j = f(nh + \alpha_j h, z(nh) + \sum_{i<j} \beta_{ij} h k_i)$$

$$z((n+1)h) = z(nh) + h \sum_j c_j k_j$$

Some popular choice of parameters:

(i) Heun's method, or improved Euler method: $\alpha_2 = \beta_{12} = 1$, $c_1 = c_2 = 1/2$

(ii) RK2, or modified Euler's method: $\alpha_2 = \beta_{12} = 1/2$, $c_1 = 0$, $c_2 = 1$.

(iii) RK4: $\alpha_2 = \alpha_3 = \beta_{12} = \beta_{23} = 1/2$, $\alpha_4 = \beta_{34} = 1$, $c_1 = c_4 = 1/6$, $c_2 = c_3 = 1/3$.

How to check consistency and order of accuracy: Taylor series expansion for y and f.

Example 4.11. *Show that Heun's method is consistent and has order of accuracy* 2.

(i)
$$f(t + s, y(t) + r) = f(t, y(t)) + sf_1(t, y(t)) + rf_2(t, y(t)) + O(r^2 + s^2)$$

(ii) Now calculate the Taylor series of $y(t + s)$:

$$y(t + s) = y(t) + y'(t)s + \frac{y''(t)}{2}s^2 + O(s^3)$$

$$= y(t) + f(t, y(t))s + \frac{f_1 + ff_2}{2}s^2 + O(s^3)$$

(iii) Now do Heun's method:
$$k_1 = f(t, y(t))$$
$$k_2 = f(t + h, y(t) + hk_1) = f + h(f_1 + ff_2) + O(h^2)$$

So
$$z(t + h) = y(t) + \frac{h}{2}(k_1 + k_2) = y(t + h) + O(h^3)$$

(iv) The method is consistent and has order of accuracy 2.

Similarly, rk2 can be shown to have order of accuracy 2 and rk4 order of accuracy 4.

Global error bound can then be obtained in a way similar to Euler's method.

When using linear multistep methods of order of accuracy n, we can calculate the initial values using one-step methods of order $n - 1$. This way the error created in the initial steps will be $O(h^n)$ which is comparable with global error of an n-th order method.

103

Example 4.12. $y' = \sin(y)$, $y(0) = 1$. $h = 0.25$, $t = 4h = 1$.

True answer is $y(1) = 1.9562950$.

(i) Euler's Method:
$$z(0.25) = y(0) + h\sin(y(0)) = 1.2103677$$
$$z(0.5) = y(0.25) + h\sin(y(0.25)) = 1.4443042$$
$$z(0.75) = 1.6923068, z(1) = 1.9404635$$

(ii) Rouge-Kutta 2nd Order
$$z(0.25) = y(0) + h\sin(y(0) + h\sin(y(0))/2) = 1.2233867$$
$$z(0.5) = 1.4668103, z(0.75) = 1.7167586, z(1) = 1.9577257$$

(iii) Improved Euler's Method+Adams-Bashforth 3rd order

    (a) Improved Euler's Method for $z(0.25)$ and $z(0.5)$:

$$z(0.25) = y(0) + \frac{h}{2}(\sin(y(0)) + \sin(y(0) + h\sin(y_0))) = 1.2221521$$

$$z(0.5) = 1.4638248$$

    (b) Adams-Bashforth 3rd order:

$$z(0.75) = z(0.5) + h(\frac{23}{12}\sin(z(0.5)) - \frac{4}{3}\sin(z(0.25)) + \frac{5}{12}\sin(y(0))) = 1.7146269$$

$$z(1) = 1.9553174$$

(iv) Rouge-Kutta 4th Order:

(a) Calculate $z(0.25)$:

$$k_1 = \sin(y(0)) = 0.8414710, k_2 = \sin(y(0) + \frac{h}{2}k_1) = 0.8935468$$

$$k_3 = \sin(y(0) + \frac{h}{2}k_2) = 0.8964504, k_4 = \sin(y(0) + hk_3) = 0.9405047$$

$$z(0.25) = y(0) + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 1.2234154$$

(b) Continue with the calculation, we get:

$$z(0.5) = 1.4663981, z(0.75) = 1.7156965, z(1) = 1.9562859$$

Let $h = 1/n$, $n = 4, 5, \ldots, 20$, the behavior of the global error at $t = 1$ is (see https://github.com/wuchenxi/Math-514/blob/main/ivp.py):
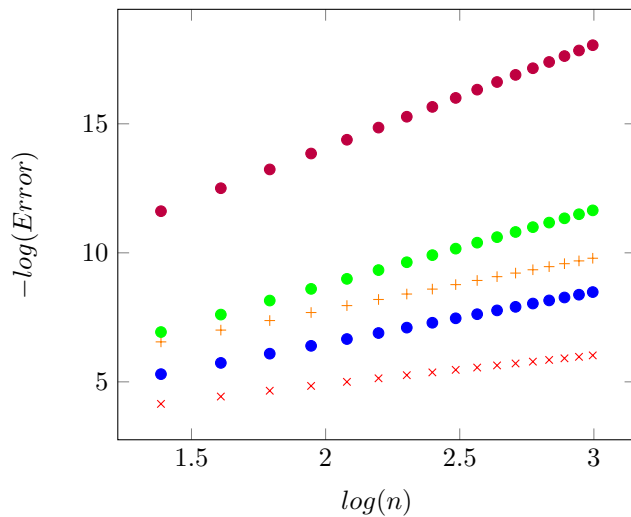
The dots from low to high are euler, heun, rk2, ab3, rk4.

4.4 Stiffness and Absolute Stability

All the methods we discussed so far can be easily generalized to systems of equations. Just see y as a vector-valued function.

Example 4.13.

$$y' = \begin{pmatrix} -1 & 0 \\ 1 & -100 \end{pmatrix} y$$

$$y(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Do time step $h = 0.1$, $t = 10h = 1$, using Euler's, improved Euler's and trapezium rule methods:

$$y(1) = \begin{pmatrix} 1/e \\ e^{-100}(e^{99}/99 + 98/99) \end{pmatrix}$$

Let $A = \begin{pmatrix} -1 & 0 \\ 1 & -100 \end{pmatrix}$.

- Euler's method:

$$z(1) = (I_2 + 0.1A)^{10} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.3486784401 \\ 3451564356.5489765499 \end{pmatrix}$$

108

- Heun's method:

$$z(1) = (I_2 + 0.1A + 0.005A^2)^{10} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.368540984834 \\ 1.32870768929 \times 10^{16} \end{pmatrix}$$

- Trapezium rule method:

$$z(1) = ((I_2 + 0.05A)/(I_2 - 0.05A))^{10} \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$= \begin{pmatrix} 0.367572542383 \\ 0.02087921691 \end{pmatrix}$$

- A system of equations is called stiff, if, after linearization into $y' = Ay$, A has eigenvalues with negative real parts, and the ratio between real parts of eigenvalues can be large.

- Stiffness means there are behavior in different time scale. A numerical method need to take small step size to accommodate for the faster behavior, but also need to calculate till a large t to see the slow behavior, resulting in huge amount of computation.

- There are other cases of stiffness which are beyond the scope of this course.

To deal with stiff equations efficiently, we need numerical methods which does not require a small time scale to get good answers. Usually we use test equation $y' = \lambda y$, $y(0) = 1$, where $Re(\lambda) < 0$, and the set of values $h\lambda$ that makes $\lim_{n\to\infty} z(nh) = 0$, are called **the region of absolute stability**.

Example 4.14. • *For Euler's method,*

$$z((n+1)h) = z(nh) + h\lambda(z(nh)) = (1 + h\lambda)z(nh)$$

So the solution goes to 0 if $|1 + h\lambda| < 1$, the region of absolute stability is the circle of radius 1 centered at -1.

• *For Trapezium Rule Method,*

$$z((n+1)h) = z(nh) + \frac{h}{2}(\lambda(z(nh)) + \lambda(z((n+1)h)))$$

So

$$z((n+1)h) = \frac{1 + h\lambda/2}{1 - h\lambda/2}z(nh)$$

The region of absolute stability is the left half plane.

A method whose region of absolute stability is the whole left half plane is called **A-stable**.

4.5   Review

$$y' = f(t, y), y(0) = y_0$$

Key idea: estimate $y(h), y(2h), y(3h), \ldots$ successively.

Methods:

(i) Euler's method

(ii) First Generalization of Euler's Method: Method based on quadrature rule (Adams-Bashforth, Adams-Moulton), general Linear Multistep methods

(iii) Second Generalization of Euler's Method: Rouge-Kutta family

Concepts:

(i) Local

    (a) Truncated error

    (b) Consistency

    (c) Order of accuracy

(ii) Global

    (a) Zero stability

    (b) Convergence

(iii) Efficiency issue

    (a) Region of absolute stability.

    (b) A-stability.

How to analyze numerical methods:

(i) Locally: Taylor series expansion.

(ii) Globally: separate the error created at each step, as in the argument for Euler's method. We can summarize it as below:

Theorem 4.15. *(Theorem 12.2 in textbook) If* $z((n+1)h) = z(nh) + h\Phi(nh, z(nh))$, $z(0) = y(0)$, $|y((n+1)h) - h\Phi(y(nh)) - y(nh)| \leq Th$, and Φ is L-Lispchitz with respect to the second parameter. Then

$$|z(nh) - y(nh)| \leq T \cdot \frac{e^{nhL} - 1}{L}$$

Proof. Let $z_k(kh) = y(kh)$, $z_k((n+1)h) = z_k(nh) + h\Phi(nh, z_k(nh))$, then

$$|z_k(nh) - z_{k+1}(nh)| \leq (1+hL)^{n-k-1}|z_k((k+1)h) - z_{k+1}((k+1)h)| \leq Th(1+hL)^{n-k-1}$$

$$|z(nh) - y(nh)| = |z_0(nh) - z_n(nh)| \leq T \cdot \frac{e^{nhL} - 1}{L}$$

$\square$

Example 4.16. *Consistency, order of accuracy, and convergence of rk2.*

$$y' = f(t, y), y(0) = y_0$$

$$z(t+h) = z(t) + hf(t + \frac{h}{2}, z(t) + \frac{h}{2}f(t, z(t)))$$

(i) Consistency and order of accuracy: Suppose $z(t) = y(t)$, then

$$y(t+h) = y(t) + hy'(t) + \frac{h^2}{2}y''(t) + \frac{h^3}{6}y'''(t) + \dots$$

$$= y(t) + f(t, y(t))h + (f_1(t, y(t)) + f(t, y(t))f_2(t, y_t))\frac{h^2}{2} + (f_{11} + ff_{12} + (f_1 + ff_2)f_2 + f(f_{12} + ff_{22}))\frac{h^3}{6} + \dots$$

$$z(t+h) = y(t) + hf(t + \frac{h}{2}, y(t) + \frac{h}{2}f(t, y(t)))$$

$$= y(t) + f(t, y(t))h + (f_1 + ff_2)\frac{h^2}{2} + (f_{11} + 2f_{12}f + f_{22}f^2)\frac{h^3}{8} + \dots$$

So

$$|y(t+h) - z(t+h)| = O(h^3)$$

It is consistent with order of accuracy 2.

(ii) If f, f_1, f_2, f_{11}, f_{12}, f_{22} are all bounded, then there is uniform constant K such that

$$|y(t+h) - z(t+h)| \leq Kh^3$$

(iii) $\Phi(t,y) = f(t + \frac{h}{2}, z(t) + \frac{h}{2}f(t,z(t)))$, hence it is $L + hL^2/2$-Lipschitz.

(iv) Hence in this case, by Theorem 4.15,

$$|y(t) - z(t)| \leq Kh^2 \frac{e^{(L+hL^2/2)t} - 1}{L + hL^2/2} = O(h^2)$$

Exercise: Consider $y' = y\cos(t)$, $y(0) = 1$.

(i) Find A, B such that the linear multistep method

$$z(t + 2h) = z(t) + Ahf(t, z(t)) + Bhf(t + h, z(t + h))$$

is consistent. (Answer: $A + B = 2$)

(ii) Find A, B such that the linear multistep method has order of accuracy 2. (Answer:

$$y(t + 2h) = y(t) + f \cdot 2h + (f_1 + ff_2)(2h)^2/2 + \ldots$$

$$z(t + 2h) = y(t) + Ahf + Bh(f + h(f_1 + ff_2)) + \ldots$$

Hence$A = 0$, $B = 2$)

(iii) Estimate $y(h)$ using Euler's method, then $y(2h)$ using this linear multistep method. (Answer: $z(h) = 1 + h$, $z(2h) = 1 + Ah + Bh(1+h)\cos(h)$.)

(iv) Find the region of absolute stability of this linear multistep method, using the fact (from linear algebra) that iterative relation

$$a_{n+2} = ca_{n+1} + da_n$$

has all solutions converging to 0 iff all roots of $z^2 - (cz + d)$ has absolute value less than 1. (Answer: apply the algorithm to $y' = \lambda y$, we get

$$z((n+2)h) = z(nh) + Ah\lambda z(nh) + Bh\lambda z((n+1)h)$$

so $h\lambda$ is in region of absolute stability iff $z^2 - Bh\lambda z - (1 + Ah\lambda) = 0$ has all roots inside the open unit circle, i.e.

$$|\frac{Bh\lambda \pm \sqrt{B^2h^2\lambda^2 + 4 + 4Ah\lambda}}{2}| < 1$$

5   Boundary Value Problems

This section will not be in final exam.

Example: $y'' = -y$, $y(0) = 0$, $y(1) = 1$. True answer is $y(x) = \frac{\sin(x)}{\sin(1)}$.

We discretize the problem by estimating the solution on a **uniform mesh**: $x_i = i/n$, $i = 0, 1, \ldots n$. Denote z as the estimation.

5.1   Finite Difference (Chap. 13)

$z(0) = 0$, $z(1) = 1$. For any $x_i = i/n$, $i = 1, \ldots, n - 1$, approximate the second order derivative using idea from Section 1.3.1:

Lagrange interpolation using x_{i-1}, x_i and x_{i+1} as interpolation points, we get

$$p(x) = z((i-1)/n)\frac{(x - i/n)(x - (i+1)/n)}{2/n^2} - z(i/n)\frac{(x - (i-1)/n)(x - (i+1)/n)}{1/n^2}$$

$$+z((i+1)/n)\frac{(x - i/n)(x - (i-1)/n)}{2/n^2}$$

So

$$p''(x_i) = n^2(z((i+1)/n) + z((i-1)/n) - 2z(i/n))$$

So the question reduces to a system of equations:

$$z(0) = 0$$

$$z(1) = 1$$

$$n^2(z((i+1)/n) + z((i-1)/n) - 2z(i/n)) = -z(i/n)$$

When $n = 3$, we get $z(1/3) = 81/208$, $z(2/3) = 153/208$. Error is about 0.0007.

5.2   Finite Element Method (Chap. 14)

Rewrite the differential equation into a **variational problem**, which is minimizing

$$\int_0^1 y'^2 - y^2 dx$$

Now pick values of $z(i/n)$, such that the linear spline g using x_i minimizes $\int_0^1 g'^2 - g^2 dx$. When $n = 3$, this gives us $z(1/3) = 3025/7791$, $z(2/3) = 5720/7791$. Error is about 0.0007.

6 Final review

(i) Interpolation

    (a) Lagrange and Hermite interpolation

    (b) Uniqueness

    (c) Error Estimate

(ii) Orthogonal Polynomials

    (a) p_d is a degree d orthogonal polynomial of weight w on $[a, b]$, iff for any polynomial q of degree no more than $d - 1$, $\int_a^b w p_d q \, dx = 0$.

    (b) Gram-Schmidt process

    (c) Example: Legendre and Chebyshev polynomials.

(iii) Numerical integration

    (a) Quadrature rule

        i. Error estimate

        ii. Newton-Cotes quadrature. (Special cases: trapezium rule, Simpson's rule)

            A. Improved error estimate for n even. (Special case: error bound for Simpson's rule)

        iii. Gauss quadrature (Special case: Gauss-Legendre quadrature)

            A. Error bound

            B. Positive weights

            C. Convergence

    (b) Composite methods

(iv) Numerical solution for IVP

    (a) Methods

        i. Linear Multistep Methods (Special case: AB, AM)

        ii. Runge-Kutta methods (Special case: RK2, RK4, Heun)

    (b) Analysis

        i. One step: power series expansion w.r.t. h

        ii. Global error

            A. Zero stability and Dahlquist theorem

            B. One step method: $z(t + h) = z(t) + h\Phi(z(t))$, needs Φ Lipschitz.

        iii. Domain of absolute stability

Example 6.1. *Consider the function $f(x) = \frac{1}{\pi}\cos(\pi x)$. Find the Hermite interpolation polynomial p with interpolation points at $\pm 1/2, \pm 3/2$. Find the error bound at $x = 0$ using the error bound of Hermite interpolation.*

Answer:

$$p = -\frac{(x + \frac{3}{2})(x - \frac{3}{2})^2(x^2 - 1/4)^2}{36} + \frac{(x + \frac{1}{2})(x - \frac{1}{2})^2(x^2 - \frac{9}{4})^2}{4}$$

$$-\frac{(x - \frac{1}{2})(x + \frac{1}{2})^2(x^2 - \frac{9}{4})^2}{4} + \frac{(x + \frac{3}{2})(x - \frac{3}{2})^2(x^2 - 1/4)^2}{36}$$

The error bound at 0 is

$$\frac{\max \|f^{(8)}\| \times \frac{1}{2}^4 \times \frac{3}{2}^4}{8!} = \frac{9\pi^7}{1146880} \approx 0.0237$$

The actual error is $0.00972 < 0.0237$.

Remark 6.2. *When one add two more interpolation points at $\pm(m - \frac{1}{2})$, the theoretical error bound is multiplied by*

$$\frac{\pi^4(m - \frac{1}{2})^4}{(4m - 3)(4m - 2)(4m - 1)(4m)} \approx \frac{\pi^4}{4^4} < 1$$

So if we do interpolation at $\pm 1/2, \ldots, \pm(m - 1/2)$, as $m \to \infty$ the value of the interpolation polynomial at $x = 0$ does converge to $\frac{1}{\pi}$, though very slowly.

Example 6.3. *Consider IVP $y' = f(y)$, $y(0) = y_0$, where $|f|, |f^{(k)}|$ are all bounded by 1. Consider the linear multistep method*

$$z(t+2h) = z(t) + \frac{hf(z(t))}{3} + \frac{4hf(z(t+h))}{3} + \frac{hf(z(t+2h))}{3}$$

Here z is an approximation of y. Find C such that for sufficiently small h, if $z(t) = y(t)$, $z(t+h) = y(t+h)$, then

$$|z(t+2h) - y(t+2h)| < Ch^5$$

Answer:

- Firstly, by error bound of Simpson's rule,

$$y(t+2h) = y(t) + \int_t^{t+2h} f(y(s))ds$$

$$= y(t) + \frac{hf(y(t))}{3} + \frac{4hf(y(t+h))}{3} + \frac{hf(y(t+2h))}{3} + E$$

Here

$$|E| \leq \frac{\max|(f \circ y)^{(4)}|(2h)^5}{2880}$$

$$\leq \frac{4h^5}{15}$$

This is because

$$|(f \circ y)^{(4)}| = |(ff'^4 + 11f^2f'^2f'' + 4f^3f''^2 + 7f^3f'f'' + f^4f''') \circ y| \leq 24$$

- Now, by assumption,

$$|y(t+2h) - z(t+2h)| \leq \frac{h|y(t+2h) - z(t+2h)|}{3} + |E|$$

So, for example, if $h < 1$, we have

$$|y(t+2h) - z(t+2h)| \leq \frac{2}{5}h^5$$

Example 6.4. *Suppose we use quadrature rule to estimate $\int_0^1 f dx$, $x_0 = -1$, $x_1 = 0$. Can you find a x_2 such that the quadrature rule gives accurate answer for any polynomial of degree 3?*

Answer: For any x_2, the quadrature rule with 3 quadrature points I_3 will give accurate answer to any polynomial of degree at most 2. Let $g_3 = x(x+1)(x-x_3)$ (you can also choose any other polynomial of degree 3, e.g. x^3), then any polynomial of degree 3 can be written as $cg_3 + h$, where c is a constant and h is a polynomial of degree at most 2. Hence

$$0 = I_3(f) - \int_0^1 f dx$$

$$= cI_3(g_3) + I_3(h) - c\int_0^1 g_3 dx - \int_0^1 h dx$$

$$= c(I_3(g_3) - \int_0^1 g_3 dx)$$

So we only need to have $I_3(g_3) = \int_0^1 g_3 dx$. Because g_3 is zero at all three quadrature points, this is equivalent to $\int_0^1 g_3 dx = 0$, by calculation $x_3 = 7/10$.

Example 6.5. *Suppose f and all its partial derivatives are bounded by* 1. *Consider IVP:* $y' = f(t, y)$, $y(0) = y_0$, *and the implicit one-step method*

$$z(0) = y(0), z((n+1)h) = z(nh) + hf((n+1)h, z((n+1)h))$$

- *Find the order of accuracy.*

- *Show that this method is A-stable, i.e. the domain of absolute stability covers the whole left half plane.*

Answer: By Taylor series:

$$y(nh + h) = y(nh) + hy'(nh) + h^2 \frac{y''(nh)}{2} + \dots$$

$$= y(nh) + hf(nh, y(nh)) + \frac{h^2}{2}(f_t(nh, y(nh)) + f(nh, y(nh))f_y(nh, y(nh)) + \dots$$

Now assume $z(nh) = y(nh)$, and do power series expansion for $z(nh+h)$: suppose $z(nh+h) = y(nh) + ah + bh^2 + O(h^3)$, then the formula for this implicit method becomes

$$y(nh)+ah+bh^2+O(h^3) = y(nh)+h(f(nh, y(nh))+f_t(nh, y(nh))h+f_y(nh, y(nh))(ah+bh^2+O(h^3))+O(h^2))$$

Here the right hand side is due to the fact that f is differentiable, hence if $t - nh = O(h)$, $y - y(nh) = O(h)$, then

$$f(t, y) = f(nh, y(nh)) + f_t(nh, y(nh))(t - nh) + f_y(nh, y(nh))(y - nh) + O(h^2)$$

Now substitute y with $(n + 1)h$ and y with $y(nh + h) = y(nh) + ah + bh^2 + O(h^3)4$.

Now compare coefficients on both sides, we get

$$a = f(nh, y(nh))$$

$$b = f_t(nh, y(nh)) + f(nh, y(nh))f_y(nh, y(nh))$$

So $y(nh + h) - z(nh + h) = O(h^2)$, the order of accuracy is 1.

To see domain of absolute stability, consider equation $y' = \lambda y$, $\lambda \in \mathbb{C}$ and $Re(\lambda) < 0$. The method becomes:

$$z(nh + h) = z(nh) + h\lambda z(nh + h)$$

Hence the solution converges to 0 iff $\left|\frac{1}{1-h\lambda}\right| < 1$, which is true for all $h > 0$ and all $Re(\lambda) < 0$.

Example 6.6. *Apply Euler's method to the IVP $y' = f(t)$, $y(0) = 0$. Bound the error at time $t = 1$ using the derivative of f and step size $1/n$.*

Answer:

- We can use the error bound for Euler's method discussed in the lecture, which is $Ch\frac{e^L-1}{L}$ where C is $\frac{\max(|f_t|)+\max(|f|)\max(|f_y|)}{2} = \frac{\max|f'|}{2}$, and L can be any positive number, which is because f is not dependent on y, hence is 0-Lip with respect to y. Now let $L \to 0$, we get the upper bound $\frac{\max|f'|h}{2} = \frac{\max|f'|}{2n}$.

- Alternatively, we can use Euler's method to get the estimate of y at time 1, which is

$$z(1) = \sum_{i=0}^{n-1} \frac{1}{n} f(\frac{i}{n})$$

So

$$|y(1) - z(1)| = \sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} |f(s) - f(i/n)| ds$$

$$\leq \sum_{i=0}^{n-1} \int_{i/n}^{(i+1)/n} \max|f'||s - i/n| ds$$

$$= n\frac{\max|f'|}{2n^2} = \frac{\max|f'|}{2n}$$

The inequality is due to mean value theorem.

Example 6.7. *Consider solving IVP* $y' = f(y), y(0) = 0$. *Here* $f \in C^2$, $|f| \le 1$, $|f'| \le 1$, $|f''| \le 1$

- *Show that* $y(t + h) = y(t) + \int_t^{t+h} f(y(s))ds$.

- *Write down the Hermite interpolation of* $f(y(s))$ *using interpolation point* $x_0 = t$.

- *Integrate this Hermite interpolation, and make a numerical method based on this integration.*

- *Find the order of accuracy of this method. Bound its truncated error.*

Answer:

- This is fundamental theorem of calculus.

- $p(s) = f(y(t)) + f(y(t))f'(y(t))(s - t)$

- $\int_t^{t+h} = hf(y(t)) + \frac{h^2}{2}f(y(t))f'(y(t))$. So we can approximate the IVP using

$$z(t + h) = z(t) + hf(z(t)) + \frac{h^2}{2}f(z(t))f'(z(t))$$

- When $z(t) = y(t)$, the Taylor series with remainder formula tells us:

$$y(t+h) = y(t) + hf(y(t)) + \frac{h^2}{2}f(y(t))f'(y(t)) + \frac{h^3}{6}(f(y(s))f'^2(y(s)) + f(y(s))f''(y(s)))$$

Where $s \in [t, t + h]$.

$$z(t + h) = y(t) + hf(y(t)) + \frac{h^2}{2}f(y(t))f'(y(t))$$

So the truncated error is bounded by

$$|\frac{y(t + h) - z(t + h)}{h}| \le \frac{\max|f|\max|f'|^2 + \max|f|\max|f''|}{6}h^2 = O(h^2)$$

So order of accuracy is 2.

7   HW Solutions

7.1   HW 4

1. Let $f(x) = x^3$, p be the Lagrange interpolation polynomial of f using interpolation points $x = 0$, $x = 1$. On the interval $[0, 1]$, find the point c that maximizes the interpolation error $|f(c) - p(c)|$, and find another point $s \in [0, 1]$ such that

$$f(c) - p(c) = f''(s)c(c - 1)/2$$

Answer:

$$p(x) = 0 \cdot \frac{x - 1}{0 - 1} + 1 \cdot \frac{x - 0}{1 - 0} = x$$

$$|f - p| = |x^3 - x|$$

So this is maximalized at point $c = \frac{\sqrt{3}}{3}$.

$$f(c) - p(c) = c^3 - c = 3sc(c - 1)$$

So

$$s = \frac{c + 1}{3} = \frac{\sqrt{3} + 3}{9}$$

2.   Let $f(x) = e^x$, p be the Lagrange interpolation polynomial of f on interval $[0, 2]$ using interpolation points $x_0 = 0$, $x_1 = 1$, $x_2 = 2$, find an upper bound for the L^∞ norm of $f(x) - p(x)$ on $[0, 2]$, using the error bound of Lagrange polynomial we covered in the lecture (Theorem 6.2 in textbook, Theorem 1.5 in lecture notes).

Answer:

The error bound of Lagrange polynomial is

$$|f(x) - p(x)| = \frac{|f'''(c)||x(x - 1)(x - 2)|}{3!}$$

129

When $c \in [0, 2]$, $|f'''(c)| \leq e^2$.

When $x \in [0, 2]$, $|x(x-1)(x-2)| \leq \frac{2\sqrt{3}}{9}$.

Hence an upper bound for this error is $\frac{e^2\sqrt{3}}{27}$.

It's ok if you get a slightly larger error bound, for example $4e^2/3$.

3. Suppose f is continuous and with continuous derivatives of order up to and including 5 on $[a, b]$, and there are three distinct points x_0, x_1, x_2 in $[a, b]$. Let $y_i = f(x_i)$, $i = 0, 1, 2$; $z_j = f'(x_j)$, $j = 0, 2$.

(i) Find a polynomial p of degree at most 4, such that $p(x_i) = y_i$, $i = 0, 1, 2$; $p'(x_j) = z_j$, $j = 0, 2$.

(ii) Use an argument similar to the error estimate of Hermite interpolation polynomial to show that for any $x \in [a, b]$, there is some number $s \in [a, b]$ such that

$$f(x) - p(x) = f^{(5)}(s)(x - x_0)^2(x - x_1)(x - x_2)^2/5!$$

Answer:

(i) • Approach I: We can find five polynomials p_0, p_1, p_2, q_0, q_2, such that

$$p_0(x_0) = p_1(x_1) = p_2(x_2) = q_0'(x_0) = q_2'(x_2) = 1$$

$$p_i(x_j) = 0 \text{ when } i \neq j$$

$$p_i'(x_j) = 0 \text{ when } j = 0, 2$$

$$q_0'(x_2) = q_2'(x_0) = 0$$

$$q_i(x_j) = 0$$

Then the answer can be written as

$$p = \sum_i y_i p_i + z_0 q_0 + z_2 q_2$$

130

To get p_0, from $p_0(x_1) = p_0(x_2) = p_0'(x_2) = 0$ we get $p_0 = (x - x_1)(x - x_2)^2(Ax + B)$, now use the remaining two conditions, $p_0(x_0) = 1$, $p_0'(x_0) = 0$, to solve for A and B, we get

$$p_0 = \frac{(x - x_1)(x - x_2)^2}{(x_0 - x_1)(x_0 - x_2)^2}(1 - (x - x_0)(\frac{1}{x_0 - x_1} + \frac{2}{x_0 - x_2}))$$

Similarly,

$$p_1 = \frac{(x - x_0)^2(x - x_2)^2}{(x_1 - x_0)^2(x_1 - x_2)^2}$$

$$p_2 = \frac{(x - x_0)^2(x - x_1)}{(x_2 - x_0)^2(x_2 - x_1)}(1 - (x - x_2)(\frac{2}{x_2 - x_0} + \frac{1}{x_2 - x_1}))$$

$$q_0 = \frac{(x - x_0)(x - x_1)(x - x_2)^2}{(x_0 - x_1)(x_0 - x_2)^2}$$

$$q_2 = \frac{(x - x_0)^2(x - x_1)(x - x_2)}{(x_2 - x_0)^2(x_2 - x_1)}$$

So

$$p = y_0 \frac{(x - x_1)(x - x_2)^2}{(x_0 - x_1)(x_0 - x_2)^2}(1 - (x - x_0)(\frac{1}{x_0 - x_1} + \frac{2}{x_0 - x_2}))$$

$$+ y_1 \frac{(x - x_0)^2(x - x_2)^2}{(x_1 - x_0)^2(x_1 - x_2)^2}$$

$$+ y_2 \frac{(x - x_0)^2(x - x_1)}{(x_2 - x_0)^2(x_2 - x_1)}(1 - (x - x_2)(\frac{2}{x_2 - x_0} + \frac{1}{x_2 - x_1}))$$

$$+ z_0 \frac{(x - x_0)(x - x_1)(x - x_2)^2}{(x_0 - x_1)(x_0 - x_2)^2}$$

$$+ z_2 \frac{(x - x_0)^2(x - x_1)(x - x_2)}{(x_2 - x_0)^2(x_2 - x_1)}$$

- We can also use Hermite interpolation polynomial: suppose $p'(x_1) = a$, use all information for $p(x_i)$, $i = 0, 1, 2$ and $p'(x_i)$, $i = 0, 1, 2$, we can write down the Hermite interpolation polynomial which is a polynomial of degree at most 5. The coefficient for x^5 is

$$-\frac{y_0}{(x_0 - x_1)^2(x_0 - x_2)^2}\left(\frac{2}{x_0 - x_1} + \frac{2}{x_0 - x_2}\right)$$

$$-\frac{y_1}{(x_1 - x_0)^2(x_1 - x_2)^2}\left(\frac{2}{x_1 - x_0} + \frac{2}{x_1 - x_2}\right)$$

$$-\frac{y_2}{(x_2 - x_1)^2(x_2 - x_0)^2}\left(\frac{2}{x_2 - x_1} + \frac{2}{x_2 - x_0}\right)$$

$$+\frac{z_0}{(x_0 - x_1)^2(x_0 - x_2)^2} + \frac{a}{(x_1 - x_0)^2(x_1 - x_2)^2}$$

$$+\frac{z_2}{(x_2 - x_0)^2(x_2 - x_1)^2}$$

Since we want p to be of degree no more than 4, we must set this coefficient to be 0. Hence

$$a = \frac{y_0(x_1 - x_2)^2}{(x_0 - x_2)^2}\left(\frac{2}{x_0 - x_1} + \frac{2}{x_0 - x_2}\right)$$

$$+y_1\left(\frac{2}{x_1 - x_0} + \frac{2}{x_1 - x_2}\right)$$

$$+\frac{y_2(x_1 - x_0)^2}{(x_2 - x_0)^2}\left(\frac{2}{x_2 - x_1} + \frac{2}{x_2 - x_0}\right)$$

$$-\frac{z_0(x_1 - x_2)^2}{(x_0 - x_2)^2} - \frac{z_2(x_1 - x_0)^2}{(x_2 - x_0)^2}$$

Now put this in the formula for Hermite interpolation polynomials, you'll get the exact same answer as above.

(ii) If $x = x_i$ it's trivially true. Now suppose $x \neq x_i$ for any i, consider

$$G(t) = f(t) - p(t) - \frac{(f(x) - p(x))(t - x_0)^2(t - x_1)(t - x_2)^2}{(x - x_0)^2(x - x_1)(x - x_2)^2}$$

$G(x) = G(x_i) = G'(x_0) = G'(x_2) = 0$, so G' is zero at at least 5 points, $G^{(5)}$ is zero at at least one point. Let that point be s, then $G^{(5)}(s) = 0$ implies the equation we need to prove.

4. Let $q_j = (1-x^2)^j$, $\varphi_j = q_j^{(j)}$, show that φ_j are orthogonal to each other in $L^2([-1,1])$. In other words, if $j \neq j'$, $\int_{-1}^1 \varphi_j \varphi_{j'} dx = 0$.

Answer:

Firstly we show that if $i < j$, then $q_j^{(i)}$ has a factor $(1 - x^2)^{j-i}$. Do induction on i. It is trivially true for $i = 0$. Now, suppose $q_j^{(i)} = (1 - x^2)^{j-i} h(x)$ where h is a polynomial, then, by product rule,

$$q_j^{(i+1)} = ((1-x^2)^{j-i}h(x))' = -2(j-i)x(1-x^2)^{j-i-1}h(x) + (1-x^2)^{j-i}h'(x)$$

$$= (1-x^2)^{j-i-1}(-2(j-i)xh(x) + (1-x^2)h'(x))$$

Hence by induction this statement is proved.

Now, because φ_i are all non-zero, they all have non-zero L^2 norms on $[-1, 1]$. We only need to show that when $i \neq j$, $\int_{-1}^1 \varphi_i \varphi_j dx = 0$. Without loss of generality assume $i < j$, then by integration by parts and the conclusion in the previous step,

$$\int_{-1}^1 \varphi_i \varphi_j dx = \int_{-1}^1 q_i^{(i)} q_j^{(j)} dx$$

$$= -\int_{-1}^1 q_i^{(i+1)} q_j^{(j-1)} dx$$

$$= \int_{-1}^{1} q_i^{(i+2)} q_j^{(j-2)} dx$$

$$= \cdots = (-1)^j \int_{-1}^{1} q_i^{(i+j)} q_j dx$$

However the degree of q_i is $2i < i + j$, hence $q_i^{(i+j)} = 0$, which implies that the integration is zero.

5. Find three distinct points x_0, x_1 and x_2 in $(-1, 1)$, such that for any polynomial function f of degree 3, the best approximation of f under L^2 norm on $[-1, 1]$ of degree at most 2 coincides with the Lagrange interpolation polynomial of f using interpolation points x_0, x_1 and x_2.

Answer:

Suppose f is the degree 3 Legendre polynomial $f_3 = x^3 - \frac{3}{5}x$, then, because it is orthogonal to the degree 0, 1, and 2 Legendre polynomials under $L^2([-1, 1])$, and these three Legendre polynomials form an orthogonal basis of the space V_2 of polynomials of degree no more than 2, the best approximation formula in inner product space implies that the best approximation of f on V_2 under $L^2([-1, 1])$ norm must be 0. By assumption, the Legendre interpolation of f_3 at x_0, x_1 and x_2 must also be zero, so these three points can only be the three roots of $x^3 - \frac{3}{5}x$, which are $0, \pm\sqrt{\frac{3}{5}}$.

Now suppose $f = \sum_{i=0}^{3} a_i x^i$ is any degree 3 polynomial. Then, because $f - a_3 f_3$ is of degree at most 2 and is identical to f at $0, \pm\sqrt{\frac{3}{5}}$, the Lagrange interpolation of f at x_i is $f - a_3 f_3$. On the other hand, let e_0, e_1, e_2 be any orthogonal basis of V_2, then the best approximation of f on V_2 under $L^2([-1, 1])$ norm is $\sum_i (f, e_i) e_i$. However, because f_3 is orthogonal to V_2, $(f, e_i) = (f - a_3 f_3, e_i)$, so the best approximation of f is the same as the best approximation of $f - a_3 f_3$, which must be $f - a_3 f_3$ itself as $f - a_3 f_3 \in V_2$. This proves that x_i being $0, \pm\sqrt{\frac{3}{5}}$ satisfies the requirement in the problem.

6. Let f be a continuous function on $[0, 1]$, p_n be the polynomial of best approximation of degree no more than n under the L^2 norm. Then, after studying Theorem 9.5 in the textbook, which proved that $f - p_n$ is zero at at least $n + 1$ distinct points in $(0, 1)$, find a function f such that $f - p_2$ is zero at 4 distinct points in $(0, 1)$.

Answer:

Let V_2 be the space of polynomials of degree no more than 2. If we pick f to be anything orthogonal to V_2 under the $L^2([0, 1])$ norm, then the best approximation of f on V_2 must be zero, so we just need to pick such a f with 4 or more zeros. So, for example, we can pick the degree 4 orthogonal polynomials with weight 1 on $[0, 1]$, which is $70x^4 - 140x^3 + 90x^2 - 20x + 1$.

7.2 HW 5

1. (Problem 7.6 in textbook) Consider the trapezium and Simpson's rule applied to $\int_0^1 (x^5 - Cx^4)dx$.

- Write down the error for trapezium and Simpson's rule, as functions of C.

- Find C that makes the error under trapezium rule is 0.

- Find the range of C where the trapezium rule is more accurate than Simpson's rule.

Answer:

- The true answer is $1/6 - C/5$. The result of the trapezium rule is $\frac{1-c}{2}$, and the result of Simpson's rule is $\frac{2}{3}(1/32 - C/16) + \frac{1}{6}(1 - c) = 3/16 - 5c/24$. So the error under trapezium rule is $|1/3 - 3C/10|$, the error under Simpson's rule is $|1/48 - C/120|$. It's ok if you do not write the absolute value.

- We need to have $|1/3 - 3C/10| = 0$, hence $C = 10/9$.

- This is the range of C such that $|1/3 - 3C/10| < |1/48 - C/120|$, hence $C \in (15/14, 85/74)$.

2. (Problem 7.11 in textbook) Suppose $f \in C^4([-1, 1])$. Let p be the Hermite interpolation polynomial of f at $x_0 = -1$, $x_1 = 1$.

- Calculate $\int_{-1}^{1} p\,dx$ and write it as a linear combination of $f(\pm 1)$, $f'(\pm 1)$.

- Prove that

$$|\int_{-1}^{1} f\,dx - \int_{-1}^{1} p\,dx| \leq \frac{2}{45} \max_{c \in [-1,1]} |f^{(4)}(c)|$$

Answer:

-

$$p(x) = f'(-1)\frac{(x+1)(x-1)^2}{4} + f'(1)\frac{(x-1)(x+1)^2}{4}$$

$$+f(-1)\frac{(x-1)^2(1+(x+1))}{4} + f(1)\frac{(x+1)^2(1-(x-1))}{4}$$

So

$$\int_{-1}^{1} p\,dx = \frac{f'(-1)}{3} - \frac{f'(1)}{3} + f(-1) + f(1)$$

- The error bound for Hermite interpolation tells us

$$|f(x) - p(x)| \leq \frac{\max |f^{(4)}|(x+1)^2(x-1)^2}{4!}$$

So

$$|\int_{-1}^{1} f\,dx - \int_{-1}^{1} p\,dx| \leq \int_{-1}^{1} |f - p|\,dx$$

$$\leq \max |f^{(4)}| \cdot \int_{-1}^{1} \frac{(x+1)^2(x-1)^2}{4!}\,dx$$

$$= \frac{2}{45} \max_{c \in [-1,1]} |f^{(4)}(c)|$$

3. (Problem 10.3 in textbook) Show that if $f \in C^2([0,1])$, then there is some point $c \in (0,1)$ such that

$$\int_0^1 xf\,dx = \frac{1}{2}f(2/3) + \frac{1}{72}f''(c)$$

136

Hint: use Gauss quadrature with weight x.

Answer: Because for any constant function C, $\int_0^1 xC(x-2/3)dx = 0$, $x-2/3$ is the degree-1 orthogonal polynomial on $[0,1]$ with weight x. Hence the weight x Gauss quadrature for $\int_0^1 xf dx$ should be the $I_0(f) = w_0 f(x_0)$, where x_0 is the root of $x - 2/3$ which is $2/3$, and

$$w_0 = \int_0^1 w(x) \frac{\prod_{i \neq 0}(x - x_i)}{\prod_{i \neq 0}(x_0 - x_i)} dx = \int_0^1 w dx = \frac{1}{2}$$

Now the error formula for Gauss quadrature tells us

$$\int_0^1 xf dx - I_0(f) = f''(c) \int_0^1 \frac{x(x - 2/3)^2}{2!} dx = \frac{f''(c)}{72}$$

4.

- Suppose f is continuous on $[0,1]$. Let I_n be the estimate of $\int_0^1 f dx$ using composite trapezium rule with n subintervals. Show that

$$\lim_{n \to \infty} | \int_0^1 f dx - I_n| = 0$$

  Hint: There are many possible approaches. You can use the fact that any continuous function on a closed interval is uniformly continuous (for any $\epsilon > 0$, there is some δ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$), or use the Weierstrass approximation theorem.

- (Optional) Find a continuous function f, such that there is $C > 0$ such that

$$| \int_0^1 f dx - I_n| \geq \frac{C}{n}$$

  Hint: if f has bounded second derivative then the error decays like $O(1/n^2)$, so you need to find some f that doesn't have second order derivative or has unbounded second order derivative.

137

Answer:

- – Approach I: f is continuous on $[0,1]$ hence uniformly continuous. For any $\epsilon > 0$, find δ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \epsilon$. Now let N be some integer larger than $1/\epsilon$. For any $n > N$, consider the composite trapezium rule I_n, then f sends each of the n subintervals to an interval of length no more than ϵ, hence the error of the trapezium rule on this subinterval is no more than ϵ/n, and the error of the composite trapezium rule is no more than ϵ, hence $\lim_{n \to \infty} I_n(f) = \int_0^1 f dx$ by definition of limit.

  – Approach II: For any $\epsilon > 0$, find polynomial p such that $|f - p|_\infty < \epsilon/3$, then $|\int_0^1 f dx - \int_0^1 p dx| < \epsilon/3$, and for any n, $|I_n(f) - I_n(p)| < \epsilon/3$. Now let N be large enough such that $\frac{\max |p''|}{12N^2} < \epsilon/3$, then for any $n > N$, $|I_n(p) - \int_0^1 p dx| < \epsilon/3$, hence

  $$|I_n(f) - \int_0^1 f dx| \leq |\int_0^1 f dx - \int_0^1 p dx| + |I_n(p) - \int_0^1 p dx| + |I_n(f) - I_n(p)| < \epsilon$$

- Let $f(x) = \sum_{i=1}^\infty \frac{1}{i(i+1)} \cos(2\pi i! x)$. Then $\int_0^1 f dx = 0$, and by trigonometry, $I_n(f) = \frac{1}{n}$.

7.3 HW 6

1. (Problem 12.7 in textbook) Consider solving $y' = f(t,y), y(0) = y_0$ using the trapezium method

$$z((n+1)h) = z(nh) + \frac{h}{2}(f((n+1)h, z((n+1)h)) + f(nh, z(nh)))$$

Suppose further that $|y'''|$ is uniformly bounded by M.

(i) Prove that

$$|\frac{y((n+1)h) - y(nh)}{h} - \frac{1}{2}(f((n+1)h, y((n+1)h)) + f(nh, y(nh)))| \leq \frac{Mh^2}{12}$$

Hint: You can prove it by applying integration by parts to $\int_{nh}^{(n+1)h}(x - hn)(x - nh - h)y'''(x)dx$.

(ii) Let $e_n = z(nh) - y(nh)$, and assume that f is L-Lipschitz with respect to the second parmeter, then

$$|e_{n+1}| \leq |e_n| + \frac{1}{2}hL(|e_{n+1}| + |e_n|) + \frac{h^3 M}{12}$$

Answer:

(i) • Approach I: Left hand side is the error for applying trapezium rule to calculate $\int_{nh}^{(n+1)h} f(s, y(s))ds$, and right hand side is the error bound we learned in class.

• Approach II:

$$\frac{h^3 M}{6} \geq |\int_{nh}^{(n+1)h} (x - hn)(x - nh - h)y'''(x)dx|$$

$$= |(x - hn)(x - nh - h)y''|_{nh}^{(n+1)h} - \int_{nh}^{(n+1)h} (2x - 2nh - h)y''(x)dx|$$

$$= |\int_{nh}^{(n+1)h} (2x - 2nh - h)y''(x)dx|$$

$$= |(2x - 2nh - h)y'|_{nh}^{(n+1)h} - \int_{nh}^{(n+1)h} 2y'dx|$$

$$= |h(f((n+1)h, y((n+1)h)) + f(nh, y(nh))) - 2(y((n+1)h) - y(nh))|$$

Divide h on both sides we get the required inequality.

(ii) From the inequality proved above, we have

$$y((n+1)h) = y(nh) + \frac{h}{2}(f((n+1)h, y((n+1)h)) + f(nh, y(nh))) + E$$

Where $|E| \leq \frac{Mh^3}{12}$. Hence,

$$|e_{n+1}| = |y((n+1)h) - z((n+1)h)| \leq |y(nh) - z(nh)| + \frac{hL}{2}(|y((n+1)h) - z((n+1)h)| + |y(nh) - z(nh)|) + |E|$$

$$\leq |e_n| + \frac{hL}{2}(|e_{n+1}| + |e_n|) + \frac{h^3 M}{12}$$

2. (Problem 12.12 in textbook) Consider solving the initial value problem $y' = f(t, y), y(0) = y_0$ via linear multistep method:

$$z((n+3)h) + bz((n+1)h) + az(nh) = hf((n+2)h, z((n+2)h))$$

(i) Find a, b such that the method is consistent.

(ii) Show that for such a, b, the method is not zero stable.

Answer:

(i) $1 + b + a = 0$, $3 + b = 1$, so $b = -2$, $a = 1$.

(ii) The first characteristic polynomial is now $z^3 - 2z + 1$, which has a root $\frac{-1-\sqrt{5}}{2}$ hence is not zero-stable.

7.4   Honors Assignment 2

1. (Exercise 7.13) Show that the composite Trapezium rule always give accurate answer to $\int_0^{2\pi} \sin(x)dx$.

Answer: The composite trapezium rule with n subintervals is

$$I_n = \frac{1}{n} \sum_{k=1}^{n-1} \sin(\frac{2\pi k}{n}) = \frac{1}{n} \sum_{k=1}^{n-1} (\sin(\frac{2\pi k}{n}) + \sin(\frac{2\pi(n-k)}{n}))/2$$

$$= 0 = \int_0^{2\pi} \sin(x)dx$$

2. (Exercise 10.7) Let $[a, b] = [-1, 1]$, let p_{n-1} be the degree $n-1$ orthogonal polynomial of weight $1 - x^2$, and let I_n be the quadrature rule where the quadrature points are roots of $(x^2 - 1)p_{n-1}(x)$.

- Show that if q is a polynomial of degree no more than $2n - 1$, then $\int_{-1}^{1} qdx = I_n(q)$.

- Show that all quadrature weights are positive.

- Suppose f is smooth, find a constant C such that

$$|\int_{-1}^{1} f dx - I_n(f)| \le C \max_{x \in [-1,1]} |f^{(2n)}(x)|$$

Answer:

- I_n has $n+1$ quadrature points hence gives accurate answer to any polynomial of degree up to n. If q is of degree no more than $2n-1$, by long division of polynomials we have $q = (x^2-1)p_{n-1}q_1 + r$, where r is of degree at most n, and q_1 is a polynomial of degree no more than $n-2$. Hence $I_n(q) = I_n(r) = \int_{-1}^{1} r dx = \int_{-1}^{1} q dx$.

- The j-th quadrature weight is

$$w_j = \int_{-1}^{1} \frac{\prod_{i \ne j}(x-x_i)}{\prod_{i \ne j}(x_j-x_i)} dx$$

If $j=0$ or $j=n$, the function being integrated is non negative, hence $w_j > 0$. Now suppose $0 < j < n$, then by calculation,

$$w_j = \int_{-1}^{1} \frac{x^2-1}{x_j^2-1} \cdot \frac{\prod_{i \ne j, 1 < i < n}(x-x_i)^2}{\prod_{i \ne j, 1 < i < n}(x_j-x_i)^2} dx$$

Which are all positive.

- Let p be the polynomial of degree no more than $2n-1$ such that $p(x_i) = f(x_i)$, and for all $1 < i < n$, $p'(x_i) = f'(x_i)$. Then by a similar argument to the error bound of Hermite interpolation polynomials we have

$$|f(x) - p(x)| \le \max |f^{(2n)}| \frac{(1-x^2)\prod_{1<i<n}(x-x_i)^2}{(2n)!}$$

So

$$C = \int_{-1}^{1} \frac{(1-x^2)\prod_{1<i<n}(x-x_i)^2}{(2n)!} dx$$

141

3. Consider the initial value problem $y' = \sin(y)$, $y(0) = 1$.

- Write down the formula for two step Adams-Bashforth.

- Show that the two step Adams-Bashforth has order of accuracy 2 for this problem.

- Suppose we use starting points $z(0) = 1$, $z(h) = 1 + h\sin(1)$ to carry out Adams-Bashforth till time $t = nh = 1$. Find number C such that

$$|z(1) - y(1)| \leq Ch^2$$

Answer:

- The quadrature weights for $\int_{t+h}^{t+2h}$, using $x_0 = t$, $x_1 = t + h$, are

$$w_0 = \int_{t+h}^{t+2h} \frac{(s - t - h)}{-h} ds = -h/2$$

$$w_1 = \int_{t+h}^{t+2h} \frac{(s - t)}{h} ds = 3h/2$$

So the 2nd order Adams-Bashforth is

$$z(t + 2h) - z(t + h) = h(\frac{3}{2}f(t + h, z(t + h)) - \frac{1}{2}f(t, z(t)))$$

- This can be done by doing power series expansion on both sides, or via the error formula for quadrature rules.

- Suppose $z_k(nh)$ satisfies

$$z_k(nh) = \begin{cases} y(nh) & n \leq k \\ z_k((n-1)h) + \frac{3h}{2}\sin(z_k((n-1)h)) - \frac{h}{2}\sin(z_k((n-2)h)) & n > k, n \geq 2 \\ 1 + h & n = 1, k = 0 \end{cases}$$

Then by analyzing the truncated error for Euler's and Adams-Bashforth methods, we get

$$|z_k((k+1)h) - z_{k+1}((k+1)h)| = |z_k((k+1)h) - y((k+1)h)| \leq \begin{cases} \frac{h^2}{2} & k = 0 \\ \frac{5h^3}{6} & k > 0 \end{cases}$$

You may be able to find better bounds.

Now we prove by induction on m that $|z_k((k+1+m)h) - z_{k+1}((k+1+m)h)| \leq (1+2h)^m |z_k((k+1)h) - z_{k+1}((k+1)h)|$: when $m = 0$ or $m = 1$ one can verify it directly. If $m > 1$, the left hand side is bounded by

$$|z_k((k+m)h) - z_{k+1}((k+m)h)| + \frac{3h}{2}|z_k((k+m)h) - z_{k+1}((k+m)h)| + \frac{h}{2}|z_k((k+m-1)h) - z_{k+1}((k+m-1)h)|$$

$$\leq ((1+2h)^{m-1} + \frac{3h}{2}(1+2h)^{m-1} + \frac{h}{2}(1+2h)^{m-2})|z_k((k+1)h) - z_{k+1}((k+1)h)|$$

$$\leq (1+2h)^m |z_k((k+1)h) - z_{k+1}((k+1)h)|$$

So

$$|z(nh) - y(nh)| \leq \sum_k |z_k(nh) - z_{k+1}(nh)|$$

$$\leq \sum_{k=0}^{n-1} (1+2h)^{n-k-1} h^2$$

$$\leq e^2 \frac{h^2}{2} + \frac{e^2 - 1}{2h} \frac{5h^3}{6}$$

$$= \frac{11e^2 - 5}{12} h^2$$

143

7.5　Final Review Questions

7.5.1　Basic Problems

1. Write down the Hermite interpolation polynomial $p(x)$ of $f(x) = \sin(x)$ at $x_0 = 0$, $x_1 = \pi$, and find an upper bound of $|f(x) - p(x)|$ using the error bound of Hermite interpolation.

　　Answer: The Hermite interpolation polynomial is

$$p(x) = \frac{1}{\pi^2}\left(x(x - \pi)^2 + x^2(\pi - x)\right)$$

And

$$|f(x) - p(x)| = \frac{|-\sin(s)|x^2(x - \pi)^2}{4!} \le \frac{x^2(x - \pi)^2}{24}$$

2. Find two points x_0 and x_1, such that for any polynomial f of degree no more than 3,

$$\int_0^\pi \sin(x)f(x)dx = f(x_0) + f(x_1)$$

And find constant c such that if $g \in C^6([0, \pi])$,

$$|\int_0^\pi \sin(x)g(x)dx - (g(x_0) + g(x_1))| \le C \max |g^{(6)}|$$

　　Answer: These two points are the Gauss quadrature points on interval $[0, \pi]$ with weight function $\sin(x)$, hence must be the root of the degree-2 orthogonal polynomial on $[0, \pi]$ with weight $\sin(x)$. Suppose this polynomial is $p_2 = x^2 + ax + b$, then

$$0 = \int_0^\pi \sin(x)p_2(x)dx = \pi^2 - 4 + a\pi + 2b$$

$$0 = \int_0^\pi \sin(x)xp_2(x)dx = \pi^3 - 6\pi + a(\pi^2 - 4) + b\pi$$

So $a = -\pi$, $b = 2$, $x_0 = \frac{\pi - \sqrt{\pi^2 - 8}}{2}$, $x_1 = \frac{\pi + \sqrt{\pi^2 - 8}}{2}$. And by Theorem 10.1 from the textbook or 3.19(iv) in the Lecture notes,

$$C = \frac{\int_0^\pi \sin(x)(x^2 - \pi x + 2)^2 dx}{6!} = \frac{10 - \pi^2}{180}$$

144

3. Estimate the solution of $y' = \sin(y)$, $y(0) = 1$ at time 0.1 using Euler's method, improved Euler's method, and rk4, using time step $h = 0.1$.

Answer:

- Euler's method gets $z(0.1) = 1 + 0.1 \times \sin(1) = 1 + \sin(1)/10 \approx 1.0841471$.

- Improved Euler's method gets $z(0.1) = 1 + \frac{1}{20}(\sin(1) + \sin(1 + \sin(1)/10)) \approx 1.0862688$

- Runge-Kutta 4-th order gets $k_1 = \sin(1)$, $k_2 = \sin(1 + k_1/20)$, $k_3 = \sin(1 + k_2/20)$, $k_4 = \sin(1 + k_3/10)$

$$z(0.1) = 1 + \frac{\sin(1)}{60} + \frac{\sin(1 + \sin(1)/20)}{30}$$

$$+ \frac{\sin(1 + \sin(1 + \sin(1)/20)/20)}{30}$$

$$+ \frac{\sin(1 + \sin(1 + \sin(1 + \sin(1)/20)/20)/10)}{60} \approx 1.0863557$$

The accurate answer is 1.0863558.

4. Consider explicit 2-step method for $y' = f(t, y)$:

$$z((n+2)h) = az((n+1)h) + bz(nh) + chf((n+1)h, z((n+1)h)) + dhf(nh, z(nh))$$

Where h is step size and $z(t)$ is the estimate for $y(t)$. Find all real numbers a, b, c, d such that the method is zero stable and has order of accuracy at least 2.

Answer: The first characteristic polynomial is

$$\rho(z) = z^2 - az - b$$

To make it consistent, $\rho(1) = 0$, $c + d = 2 - a$, so $1 - a - b = 0$, $b = 1 - a$, and the roots of ρ must be 1 and $a - 1$, so $0 \le a < 2$ and $b = 1 - a$.

Now let's calculate the order of accuracy. Firstly, let $t = nh$, suppose y is the solution of the IVP, then

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

145

$$y''(t) = \partial_t f(t, y(t)) + \partial_y f(t, y(t))y'(t)$$

Now let's do power series expansion, with respect to h, for

$$y(t + 2h) - ay(t + h) - (1 - a)y(t) - chf(t + h, y(t + h)) - dhf(t, y(t))$$

And after cancelling some terms, we get

$$2y''(t)h^2 - ay''(t)h^2/2 - c\partial_t f(t, y(t))h^2 - c\partial_y f(t, y(t))y'(t)h^2 + O(h^3)$$

So $c = 2 - a/2$, $d = -a/2$. Note that when $a = 1$ this is 2-step Adams-Bashforth.

7.5.2   More advanced problems

Problems like the ones below will account for no more than 10% of the final exam, so don't worry about them unless you have a lot of time during final review.

5. Suppose f is smooth and periodic with period 1, $|f^{(4)}| \leq 1$. Let I_n be the result of composite trapezium rule for $\int_0^1 f dx$ using n subintervals. Find a number C, such that

$$|\int_0^1 f dx - I_n(f)| \leq \frac{C}{n^4}$$

Answer: Consider the function $f_n(x) = \sum_{i=0}^{n-1}(x + i/n)$. Then f_n is periodic with period $1/n$, and it is easy to see that the composite trapezium rule for $\int_0^1 f dx$ using n subintervals is the same as the trapezium rule for $\int_0^{1/n} f_n dx$.

Now let p_n be the Hermite interpolation of f_n at 0 and $1/n$. Then because $f_n(0) = f_n(1/n)$, $f_n'(0) = f_n'(1/n)$, we have

$$p_n(x) = f_n(0) + 2f_n'(0)n^2 x(x - \frac{1}{2n})(x - \frac{1}{n})$$

$$\int_0^{1/n} p_n(x)dx = f_n(0)/n = I_n(f)$$

146

So

$$|\int_0^1 f dx - I_n(f)| \leq \int_0^{1/n} |f_n(x) - p_n(x)| dx \leq \int_0^{1/n} \frac{\max |f_n^{(4)}| x^2 (x - 1/n)^2}{24} dx \leq \frac{1}{720 n^4}$$

6. Consider the 3-step Adams-Bashforth method for $y' = \cos(y)$:

$$z(t + 3h) = z(t + 2h) + \frac{23h}{12} f(z(t + 2h)) - \frac{4h}{3} f(z(t + h)) + \frac{5h}{12} f(z(t))$$

Suppose $z(t) = y(t)$, $z(t + h) = y(t + h)$, $z(t + 2h) = y(t + 2h)$, find C such that

$$|z(t + 3h) - y(t + 3h)| \leq C h^4$$

Answer: Let $g(t) = y'(t) = \cos(y(t))$, p_3 be the Lagrange interpolation of g at t, $t + h$, $t + 2h$, then the 3-step Adams-Bashforth can be written as

$$z(t + 3h) = y(t + 2h) + \int_{t+2h}^{t+3h} p_3(s) ds$$

So

$$|z(t + 3h) - y(t + 3h)| \leq \int_{t+2h}^{t+3h} |g(s) - p_3(s)| ds$$

Now by error estimate of Lagrange interpolation,

$$|g(s) - p_3(s)| \leq \frac{\max |g^{(3)}|(s - t)(s - t - h)(s - t - 2h)}{6}$$

So after integration we get

$$|z(t + 3h) - y(t + 3h)| \leq \max |g^{(3)}| \cdot \frac{3h^4}{8}$$

$$g' = -y' \sin(y) = -\cos(y)\sin(y) = -\frac{\sin(2y)}{2}$$

$$g'' = -\cos(y)\cos(2y) = -\frac{\cos(3y) + \cos(y)}{2}$$

$$g''' = \frac{3\sin(3y)\cos(y) + \sin(y)\cos(y)}{2}$$

So $|g'''| \leq 7/4$, or you can use a better bound, and $C = 21/32$.

147

8 Notes on Prior Subjects

8.1 Notes on linear algebra

Recall that a (real) **Vector space** V is a set with an element 0, a "scalar multiplication" map $\mathbb{R} \times V \to V$ and an "addition" map $V \times V \to V$, such that, for any $x, y, z \in V$, any $a, b \in \mathbb{R}$, the followings are true:

(i) $x + y = y + x$

(ii) $(x + y) + z = x + (y + z)$

(iii) $0 + x = x$

(iv) $1x = x$

(v) $0x = 0$

(vi) $(a + b)x = ax + bx$

(vii) $a(x + y) = ax + ay$

(viii) $(ab)x = a(bx)$

If V is a vector space, any non empty subset $V' \subset V$ which is closed under addition and scalar multiplication is called a **subspace**.

The **span** of a set $S \subset V$ is the subset of V consisting of finite linear combinations of elements of S. We call $S \subset V$ a **linearly independent set** if for any finite collection of vectors $s_1, \ldots s_n \in S$, $\sum_i a_i s_i = 0 \implies a_i = 0 \forall i$. We call $S \subset V$ a **basis** of V if S is linearly independent and $V = span(S)$.

Any two basis of the same vector space have the same cardinality (number of elements). This cardinality is called the **dimension** of V.

Example 8.1. • *The set of n dimensional column vectors $\mathbb{R}^n$, under the usual addition and scalar multiplication, is a vector space, and it has dimension n.*

- *The set of polynomials of degree no more than n, under the usual addition and scalar multiplication, is also a vector space. A basis is $\{1, x, \ldots, x^n\}$ hence its dimension is $n + 1$.*

If $B = \{b_1, \ldots, b_n\}$ is a basis of a vector space V of dimension n, $v \in V$, the **coordinate** of v under B is a vector $x \in \mathbb{R}^n$ such that $v = \sum_i x_i b_i$ where x_i is the i-th entry of x.

A map between two vector spaces $T : V \to W$ is called a **linear transformation**, if

- $T(x + y) = T(x) + T(y)$

- $T(cx) = cT(x)$

If $T : V \to W$ is a linear transformation between two linear spaces, x is the coordinate of v in basis B, y is the coordinate of $T(v)$ under basis C, then $y = Ax$ where $A = [a_{ij}]$, and $T(b_j) = \sum_i a_{ij} c_i$.

The **inner product** on $\mathbb{R}^n$ is defined as $(x, y) = x^T y = \sum_i x_i y_i$. It is easy to check that this inner product satisfies the following properties:

(i) Symmetry: $(x, y) = (y, x)$

(ii) Bilinearity: $(ax + a'x', y) = a(x, y) + a'(x', y)$, $(x, by + b'y') = b(x, y) + b'(x, y')$.

(iii) Positive definiteness: $(x, x) \geq 0$ and $(x, x) = 0$ iff $x = 0$.

Two vectors are **orthogonal** to each other iff their inner product is 0. A set of vectors is orthogonal to another set if every vector in the first set is orthogonal to every vector in the second.

Let V be a subspace of $\mathbb{R}^n$. We call a basis of V **orthogonal** if the inner product of distinct basis vectors are all 0, **orthonomal** if in addition, the inner product of any basis vector with itself is 1.

Given any basis $\{x_1, \ldots, x_d\}$ of a subspace $V \subset \mathbb{R}^n$, we can make it into an orthogonal or orthonormal basis via the **Gram-Schmidt process**:

$$y_1 = x_1$$

$$y_i = x_i - \sum_{j<i}((y_j, x_i)/(y_j, y_j))y_j$$

Then $\{y_i\}$ is an orthogonal basis, and $\{(y_i, y_i)^{-1/2}y_i\}$ is an orthonormal basis.

If V is a subspace of $\mathbb{R}^n$, $x \in \mathbb{R}^n$, we call the **orthogonal projection** of x on V, denoted as $P_V(x)$, the unique vector that satisfies $P_V(x) \in V$ and $(x - P_V(x), y) = 0$ for all $y \in V$.

For any $x' \in V$, $(x - P_V(x), x - P_V(x)) \le (x - x', x - x')$ and equality happens iff $x' = P_V(x)$.

To calculate $P_V(x)$, we can use either of these formulas:

(i) If $\{x_i\}$ is an orthonormal basis of V, then $P_V(x) = \sum_i(x, x_i)x_i$.

(ii) If $\{x_i\}$ is an orthogonal basis of V, then $P_V(x) = \sum_i((x, x_i)/(x_i, x_i))x_i$.

(iii) If $\{x_i\}$ is just a basis of V, let $X = [x_1, \dots x_d]$ be a $n \times d$ matrix, then

$$P_V(x) = X(X^TX)^{-1}X^Tx = \sum_i(\sum_j a_{ij}(x_j, x))x_i$$

Where $A = [a_{ij}] = [(x_i, x_j)]^{-1}$ is a $d \times d$ matrix.

If one replace (x, y) with $(x, y)_A$ defined as x^TAy, where A is a symmetric matrix with all eigenvalues positive, then $(\cdot, \cdot)_A$ still satisfies symmetry, bilinearity and positive definiteness, and all the conclusions about $(\cdot, \cdot)$ above are still valid.

Furthermore, if V is any vector space and $(\cdot, \cdot)$ is a $\mathbb{R}$-valued function on $V \times V$ which is symmetric, bilinear and positive definite, all the conclusions above are valid as well.

8.2   Notes on ODE

This is a review on some basics of the theory of ordinary differential equations.

An **ordinary differential equation** is an equation relating a function on $\mathbb{R}$ and its derivatives. For example, the followings are ordinary differential equations:

$$y' = t\sin(y)$$

$$y' = sin(t)$$

$$y' = y\cos(t) + e^t$$

We can also have systems of equations like the following

$$y_1' = y_2, y_2' = -y_1$$

The **initial value problem** of an ordinary differential equation means finding a solution after specifying the value of the solution at some time t_0, which, for convenience, we can choose to be 0. For example

$$y' = y\cos(t) + e^t, y(0) = 0$$

In general the solution of an ODE can not be written down explicitly, however, in some situations we can get explicit solutions. For example, if the equation is of the form $y' = f(t)g(y)$, then the general solution is of the form

$$\int_0^y \frac{ds}{g(s)} = \int_0^t f(s)ds + C$$

This is called **separation of variables**.

The most important result in the theory of ODE is Picard's theorem:

Theorem 8.2. *If $f(t,y)$ is continuous and Lipschitz in the second parameter with Lipschitz constant L, then $y' = f(t,y)$, $y(0) = a$ always has a unique solution.*

Proof. Firstly let's show uniqueness: if y_1 and y_2 are two solutions, then $|y_1(t) - y_2(t)|e^{-L|t|}$ is non increasing when $t > 0$ and non decreasing when $t < 0$, hence must always be 0.

Now let's show existence: consider a sequence of functions defined as below:

$$y_0(t) = a$$

$$y_i(t) = a + \int_0^t f(s, y_{i-1}(s))ds$$

Then f being Lipschitz implies that

$$|y_i(t) - y_{i-1}(t)| \leq \int_0^t L|y_{i-1}(s) - y_{i-2}(s)|ds$$

$$\leq \int_0^t L^2(t-s)|y_{i-2}(s) - y_{i-3}(s)|ds$$

$$\leq \int_0^t L^3 \frac{(t-s)^2}{2}|y_{i-3}(s) - y_{i-4}(s)|ds$$

$$\leq \cdots \leq \int_0^t L^{i-1}\frac{(t-s)^{i-2}}{(i-2)!}|y_1(s) - y_0(s)|ds$$

$$\leq \frac{\max(|y_1 - y_0|)L^{i-1}t^{i-1}}{(i-1)!}$$

Hence the sequence converges uniformly on any finite interval, and fundamental theorem of calculus implies that the limiting function y is the solution. $\square$

The argument above can be used to show that if f is real analytic (i.e. has Taylor series convergent to itself), so is y.

8.3 Notes on Linear difference and differential equations and Taylor's theorem

8.3.1 Linear Difference Equations

A homogeneous linear difference equation is an iterative relationship:

$$z_{n+k} + a_{k-1}z_{n+k-1} + \cdots + a_0 z_n = 0$$

The general solution of a homogeneous linear difference equation can be obtained as follows:

- Firstly, define the characteristic polynomial $\chi(z) = z^k + a_{k-1}z^{k-1} + \cdots + a_0$.
- Let $\lambda_1, \ldots, \lambda_l$ be its distinct roots, $m_1, \ldots, m_l$ their multiplicities (hence $\sum_i m_i = k$).
- Then, the general solution can be written as

$$z_n = \sum_i p_i(n)\lambda_i^n$$

Where $p_i(n)$ is any polynomial of degree no more than $m_i - 1$.

8.3.2   Linear Differential Equations

Similarly, a homogeneous linear differential equation is

$$y^{(k)} + a_{k-1}y^{(k-1)} + \cdots + a_1 y' + a_0 y = 0$$

The general solution of a homogeneous linear differential equation is as follows:

- Firstly, define the characteristic polynomial $\chi(z) = z^k + a_{k-1}z^{k-1} + \cdots + a_0$.
- Let $\lambda_1, \ldots, \lambda_l$ be its distinct roots, $m_1, \ldots, m_l$ their multiplicities (hence $\sum_i m_i = k$).
- Then, the general solution can be written as

$$y(t) = \sum_i p_i(t)e^{\lambda_i t}$$

Where $p_i(n)$ is any polynomial of degree no more than $m_i - 1$.

8.3.3   Taylor Series

If $f \in C^{k+1}$, then the Taylor series of f at a, with Lagrange remainder, is

$$f(x) = f(a) + \sum_{j=1}^{k} \frac{f^{(j)}(a)(x-a)^j}{j!} + \frac{f^{(j+1)}(c)(x-a)^{j+1}}{(j+1)!}$$

Where c is in the closed interval between x and a.