Math 514

December 3, 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 $x_0, \dots x_n$, 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 $0 \leq k \leq m-1$, $f^{(k)} = 0$ 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. □

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,

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

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

□

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 $x_0, \dots x_n$. 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

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

□

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. $f(x) = \cos(x)$, $x_i = 5i/n$, $i = 0, 1, 2, \ldots n$

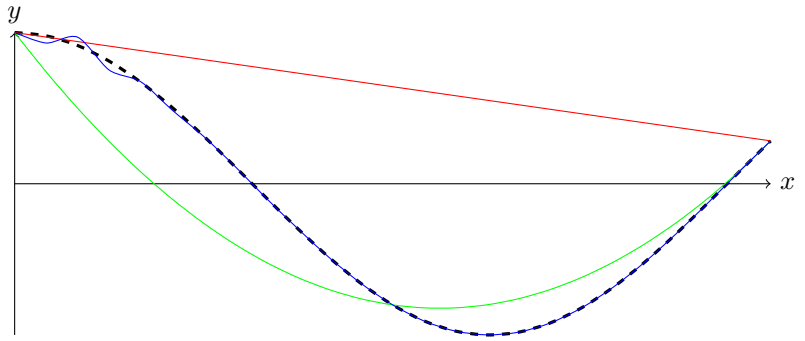


Figure 1: Black dashed line: $y = \cos(x)$. Red line: Lagrange interpolation with 2 points. Green line: Lagrange interpolation with 3 points. Blue line: Lagrange interpolation with 11 points.

Example 1.7. $f(x) = 1/(1+2(x-2)^2)$, $x_i = 5i/n$, $i = 0, 1, 2, \ldots n$.

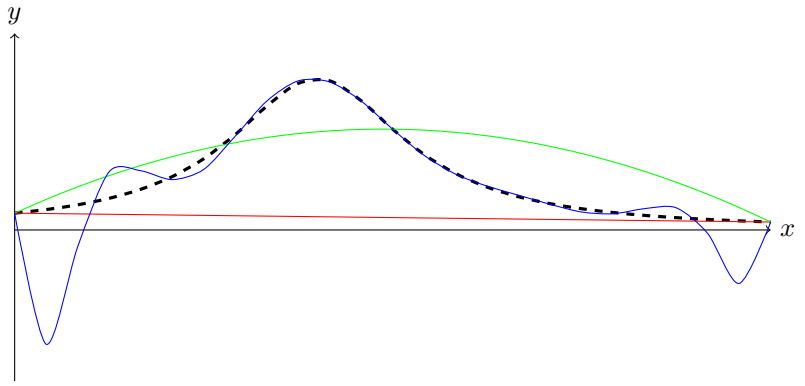


Figure 2: Black dashed line: $y = 1/(1+2(x-2)^2)$. Red line: Lagrange interpolation with 2 points. Green line: Lagrange interpolation with 3 points. Blue line: Lagrange interpolation with 11 points.

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 $p(x_i) = y_i$, $p'(x_i) = z_i$, $i = 0, \dots n$, 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 $p'(x_k) = z_k$. □

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 $p(x_i) = y_i$, $p'(x_i) = z_i$, $i = 0, \dots n$, 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 $r = 0$. □

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 $m = 2n+2$, $k = 2n+1$. □

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 $d_1, \dots d_n$, 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:*

$$p(x+t) = y_{-1}t(t-h)/(2h^2) - y_0(t+h)(t-h)/h^2 + y_1 t(t+h)/(2h^2)$$

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 $f'(x)$.

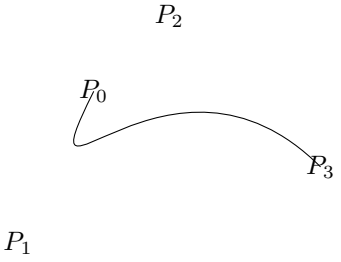
However, this approach is generally unstable. If f is complex analytic one can use complex analysis to do it which is stable, which we will not cover in this class.

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):



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*

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

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* $x \in [a,b]$, $|f(x) - s_L(x)| \leq \frac{1}{8}h^2M$.

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

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

□

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 $i = 1, \ldots, n-1$,

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

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

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

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. □

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 $F_x(y) = \|y - x\| \geq \|x\| + 1 > F_x(0) \geq F_x(x_{R_0}^*)$ $\square$

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 $(f,g) = \int_a^b fgdx$. 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 $(f,g)_w = \int_a^b wfgdx$.*

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

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

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

When the inner product is the L^2_w 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, $j = 0, 1, 2, \dots$ 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^2_w 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 $\{1, x, x^2, x^3, \dots\}$.

$$\phi_0 = 1$$

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

□

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 $L^2_w([a,b])$ can become arbitrarily accurate. In other words, the orthogonal polynomials form an orthonormal basis of $L^2_w([a,b])$. (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

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

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

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.*

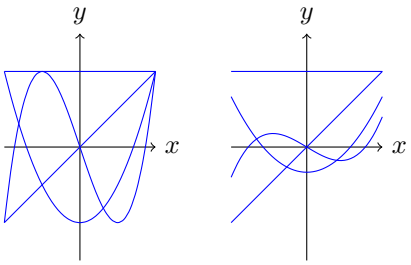


Figure 3: Chebyshev polynomials and Lagrange polynomials

- 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. □

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^2_w 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 $0, \pm 1$.*
- *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 $x^* = \sum_i ((x, x_i)/(x_i, x_i))x_i$:

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

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

$$= \frac{15e - 105e^{-1}}{4} x^2 + 3e^{-1}x + \frac{-3e + 33e^{-1}}{4}$$

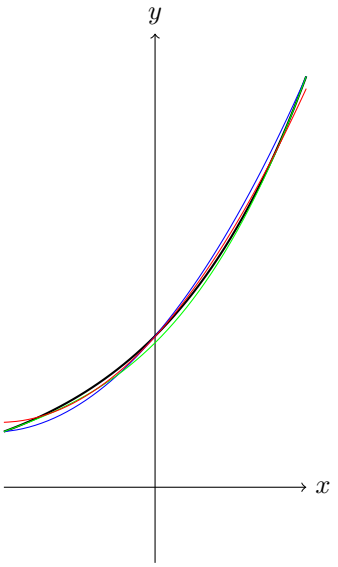


Figure 4: Black line: $y = e^x$. Blue line: Lagrange interpolation. Green line: Hermite interpolation. Red line: L^2 best approximation

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. □

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

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

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:

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

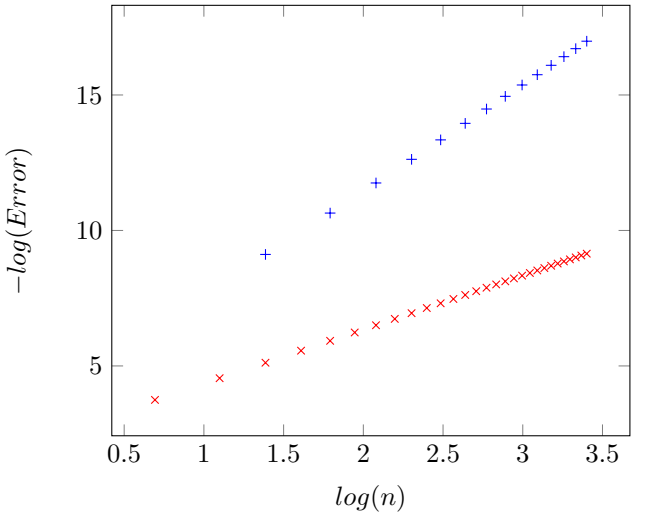
Algorithm 1: Composite Trapezium rule

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

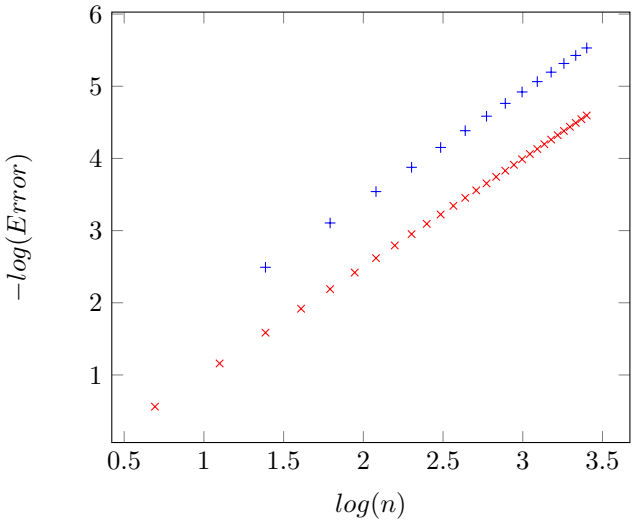
Algorithm 2: Composite Simpson's rule

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

```
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
```


15
10
5
0.5
1
1.5
2
2.5
3
3.5
$-log(Error)$
$log(n)$

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 $I(g) = \int_0^\pi g dx$.

  (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 $\int_0^\pi p dx = I(p) = I(f)$.

(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:

- $p = c^3(1-c) \cdot \frac{x(x-1))}{c(c-1)} = -c^2x(x-1)$.
- $I = c^2/6$.
- $\int_0^1 f(x)dx = 1/20$, 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

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

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 $L^2([-1,1])$, $\int_{-1}^{1} g dx = 0$

□

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. □

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

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

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

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

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

□

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 $n > \deg(p)$,

$$\begin{aligned} |\int_a^b f dx - I_n(f)| &\leq |\int_a^b f dx - \int_a^b p dx| + |\int_a^b p dx - I_n(p)| + |I_n(p) - I_n(f)| \\ &\leq \epsilon(b-a) + 0 + \epsilon(b-a) \end{aligned}$$

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. □

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 $I = \int f w dx$.
- 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 $w = e^{-x}$, $L^2_w([0,\infty))$ be the space of functions f such that $\int_0^\infty wf^2 dx$ exists. Show that $1, x, x^2 \in L^2_w([0,\infty))$.

(b) Use the inner product $(f,g) = \int_0^\infty wfgdx$, 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 $\int_0^\infty wgdx = w_0 g(x_0) + w_1 g(x_1)$ for any polynomial g of degree 1. This is called the Gauss-Laguerre quadrature.

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)

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

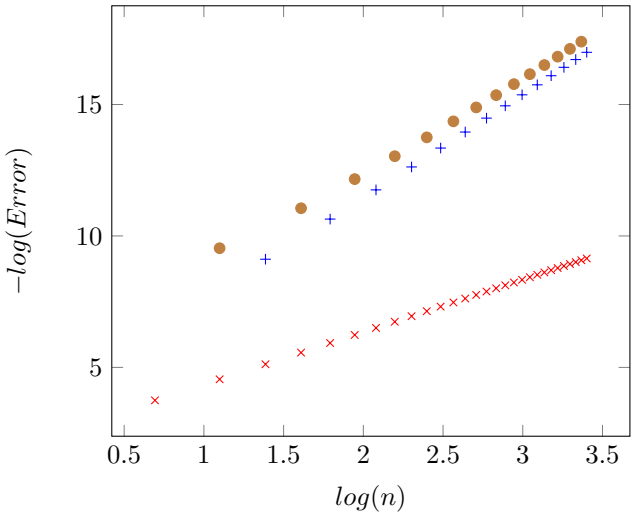
3.7 Composite Gauss quadrature (Section 10.5)

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 predetermined 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

$$z_k(kh) = y(kh) = e^{kh}$$

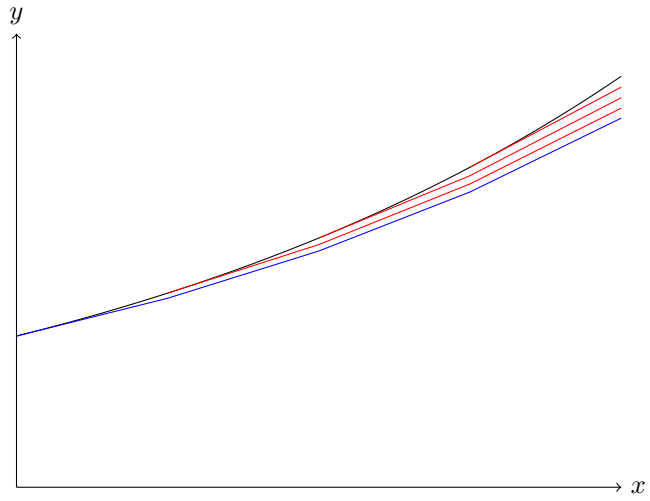
$$z_k((n+1)h) = z_k(nh) + hz_k(nh)$$

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

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

□

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)

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

(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 $\frac{Ch^2(e^{Lt}-1)}{e^h-1}$.

(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.

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

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)

$$z(t) = z(t-h) + \int_{t-h}^{t} p(s)ds$$
$$= z(t-h) + h(\frac{55}{24}f(t-h, z(t-h)) - \frac{59}{24}f(t-2h, z(t-2h))$$
$$+\frac{37}{24}f(t-3h, z(t-3h)) - \frac{3}{8}f(t-4h, z(t-4h)))$$

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

$$w_0 = 0$$

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

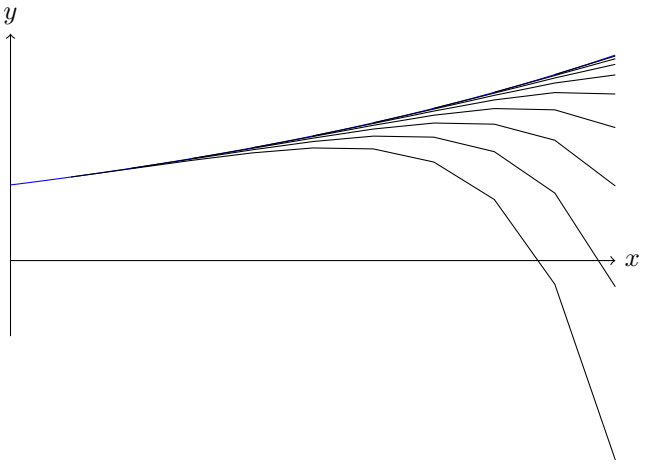
$$w_{m+2} - 3w_{m+1} + 2w_m = O(h)$$

(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.

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 $z(0.25)$:

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

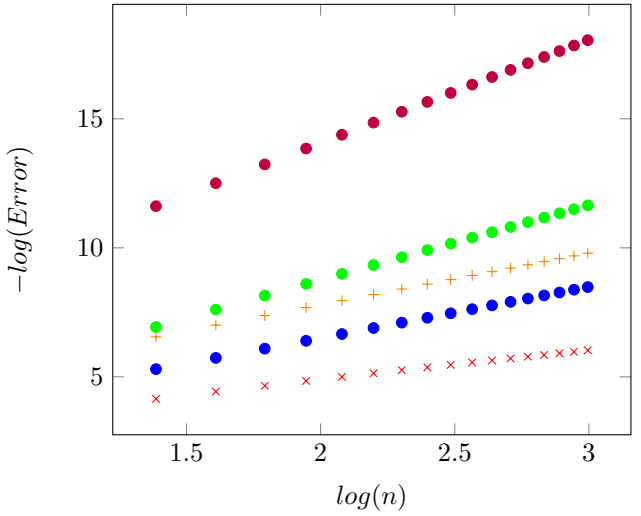
$$k_3 = \sin(y(0) + \frac{h}{2}k_2) = 0.8964504, k_4 = \sin(y(0) + hk_3) = 0.9405047$$

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

(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:

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

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

- 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

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* $z((n+1)h) = z(nh) + h\Phi(nh, z(nh))$, $z(0) = y(0)$, $|y((n+1)h) - h\Phi(y(nh))| \leq Th$, *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

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

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

□

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

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

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

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

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

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 $L + hL^2$-Lipschitz.

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

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

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.

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

(iii) Estimate $y(h)$ using Euler's method, then $y(2h)$ using this linear multistep method.

(iv) Find the region of absolute stability of this linear multistep method.

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.