AM213B Numerical Methods for the Solution of Differential Equations

Lecture 01

List of topics in this lecture

- Classification of differential equations
- Exact solutions of ODEs
- Existence and uniqueness of solution, Lipschitz continuity
- Numerical differentiation, numerical integration, discretization error, order of a numerical method
- Numerical error estimation

Reviews and Preliminaries

Classification of differential equations according to various aspects

1. Number of variables

   Single variable: $u(t)$

   ----> ODE (Ordinary Differential Equation)

   Example:

   $$u' = \sin(u^2 + t)$$

   Multiple variables: $u(x, t)$, $u(x, y)$, ...

   ----> PDE (Partial Differential Equation)

   Example:

   $$u_t = u_{xx} \,,$$
   $$u_{xx} + u_{yy} = 0$$

2. Number of unknown functions

   Scalar unknown $u(t)$

   ----> a scalar differential equation

Vector unknown $\vec{u}(t)=\begin{pmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_N(t) \end{pmatrix}$

----> a system of differential equations

Example:

$$\begin{cases} u_1{}' = u_2 \\ u_2{}' = -u_1 \end{cases}$$

3. Order of derivatives

Highest order involved

----> order of the differential equation

Example:

First order ODE: $u' = \sin(u^2+t)$

Second order ODE: $u'' = \sin(t)u' + u^2 + t$

Remark: A higher order differential equation can be converted into a first order system of differential equations.

Example: $u'' = \sin(t)u' + u^2 + t$ is a second order ODE.

Let $\vec{w}(t)=\begin{pmatrix} u(t) \\ u'(t) \end{pmatrix} \equiv \begin{pmatrix} w_1(t) \\ w_2(t) \end{pmatrix}$

The vector function $w(t)$ satisfies a first order ODE system

$$\begin{cases} w_1{}' = w_2 \\ w_2{}' = \sin(t)w_2 + w_1{}^2 + t \end{cases}$$

It has the form of

$$\vec{w}' = f(\vec{w},t)$$

General form of n-th order ODE

$$F\left(t, u(t), u'(t), \ldots, u^{(n)}(t)\right) = 0$$

Example:

$$\left(u'(t)\right)^2 + \sin(t)u'(t) + 3u(t) + \cos(t) = 0$$

A slightly less general form

$$u^{(n)}(t) = f\left(t, u(t), u'(t), \ldots, u^{(n-1)}(t)\right)$$

Example:

$$u'(t) = \sin\left(u(t)\right) + \cos(t)$$

4. <u>Linearity:</u>

ODE $F\left(t, u(t), u'(t), \ldots, u^{(n)}(t)\right) = 0$ is called linear if and only if

$F\left(t, w_0, w_1, \ldots, w_n\right)$ is a linear function of (w_0, w_1, ..., w_n).

<u>Caution:</u>

Definitions of <u>linear function</u> and <u>linear mapping</u> are slightly different.

<u>A linear function</u> is a polynomial of degree 1.

$$f(x_1, x_2, \ldots, x_n) = a_1 x_1 + a_2 x_2 + \cdots + a_n x_n + b$$

<u>A linear mapping</u> (linear operator, linear transformation) L satisfies

$$L(c_1 u + c_2 v) = c_1 L(u) + c_2 L(v)$$

Example:

$f(x) = 3x + 2$ is a linear function but is NOT a linear mapping.

5. <u>Homogeneity</u>

A differential equation is homogeneous if and only if

$u(t) \equiv 0$ is a solution.

Example:

$u'(t) = \sin(t)u(t) + t^2$ is a <u>first order</u>, <u>linear</u>, <u>nonhomogeneous</u> ODE,

6. <u>Dependence of differential equation on t</u>

$u' = f(u)$

Evolution equation is invariant with respect to t

----> autonomous

$u' = f(u, t)$

Evolution equation varies with t

----> non-autonomous

Remark: A non-autonomous system can be converted to a larger autonomous system.

Example:

$$\begin{cases} w_1' = w_2 \\ w_2' = \sin(t) w_2 + w_1^2 + \cos(t) \end{cases}$$

Let $w_3(t) = t$. We write the non-autonomous system as

$$\begin{cases} w_1' = w_2 \\ w_2' = \sin(w_3) w_2 + w_1^2 + \cos(w_3) \\ w_3' = 1 \end{cases}$$

which is an autonomous system.

Exact solutions of ODEs

Method of integrating factor

- Exact solution of first order, linear scalar ODE with constant coefficient

$$u' = au + g(t)$$

==> $u' - au = g(t)$

==> $e^{-at}u' - e^{-at}au = e^{-at}g(t)$

The key is $-e^{-at}a = \left(e^{-at}\right)'$

==> $e^{-at}u' + \left(e^{-at}\right)' u = e^{-at}g(t)$

==> $\left(e^{-at}u\right)' = e^{-at}g(t)$

Integrating from 0 to t, we get

$$e^{-at}u(t) - u(0) = \int_0^t e^{-as} g(s)\, ds$$

==> $$\boxed{u(t)=e^{at}u(0)+e^{at}\int_0^t e^{-as}g(s)ds}$$

- Exact solution of first order, linear ODE system with constant coefficients

$$\vec{u}' = A\vec{u}+\vec{g}(t)$$

The key is $-e^{-At}A=\frac{d}{dt}(e^{-At})$

...

==> $$\boxed{\vec{u}(t)=e^{At}\vec{u}(0)+e^{At}\int_0^t e^{-As}\vec{g}(s)ds}$$

where the exponential of a matrix is defined using the Taylor expansion

$$e^{At}=\sum_{n=0}^{\infty}\frac{1}{n!}(At)^n$$

In particular, we have

$$\frac{d}{dt}e^{At}=e^{At}A=Ae^{At}$$

- Exact solution of first order, linear, scalar ODE with variable coefficient

$$u'=a(t)u+g(t)$$

Let $p(t)=\int a(t)dt$

==> $u'-p'(t)u=g(t)$

==> $e^{-p(t)}u'-p'(t)e^{-p(t)}u=e^{-p(t)}g(t)$

The key is $-p'(t)e^{-p(t)}=\frac{d}{dt}\left(e^{-p(t)}\right)$

==> $\left(e^{-p(t)}u\right)'=e^{-p(t)}g(t)$

...

==> $$\boxed{u(t)=e^{p(t)}u(0)+e^{p(t)}\int_0^t e^{-p(s)}g(s)ds}$$

- Exact solution of first order, linear, ODE system with variable coefficients

  $$\vec{u}' = A(t)\vec{u} + \vec{g}(t)$$

  Unfortunately, we don't have an analytical expression for the exact solution.

  Caution: In general, matrix multiplication is not commutative.

  Let $P(t) = \int A(t)dt$

  $$\frac{d}{dt}\exp\big(P(t)\big) \neq -\exp\big(P(t)\big)A(t)$$

Existence and uniqueness of solution

We consider the initial value problem (IVP)

$$\begin{cases} u' = f(u,t) \\ u(t_0) = u_0 \end{cases} \qquad \text{(E01)}$$

We can set $t_0 = 0$.

The existence and uniqueness is governed by the Lipschitz continuity.

Definition (Lipschitz continuity)

$f(u,t)$: $R^{n+1} \to R^n$ is Lipschitz continuous (LC) in variable u over domain $\Omega \times [t_0, t_2]$ if there exists a constant C_L such that

$$\left\| f(u,t) - f(v,t) \right\| \le C_L \left\| u - v \right\|$$

for all $t \in [t_0, t_2]$, $u \in \Omega$, $v \in \Omega$

Theorem:

If $f(u, t)$ is Lipschitz continuous (LC), then IVP (E01) has a unique solution.

Note: when region Ω is bounded, it is easy to check Lipschitz continuity:

If region Ω is bounded and function $f(u, t)$ is continuously differentiable in $\Omega \times [t_0, t_2]$, then $f(u, t)$ is Lipschitz continuous (LC) over $\Omega \times [t_0, t_2]$.

Caution: This is true only over a bounded region Ω.

Example:

$$\begin{cases} u' = u^2 \\ u(0) = 1 \end{cases}$$

$f(u) = u^2$ is not Lipschitz continuous over R.

Let us look at the consequence of violating LC.

We use the method of separation of variables to solve it exactly.

$$\frac{du}{dt} = u^2$$

==> $$\frac{1}{u^2} du = dt$$

Integrating both sides, we have

$$\frac{-1}{u} = t + c$$

==> $$u = \frac{-1}{t+c}$$

Enforcing the initial condition $u(0) = 1$, we obtain

$$u(t) = \frac{1}{1-t}$$

The solution blows up at $t = 1$.

==> Solution does not exist beyond $t = 1$.

Let us see another example where the Lipschitz continuity (LC) is violated.

Example:

$$\begin{cases} u' = \sqrt{u} \\ u(0) = 0 \end{cases}$$

$f(u) = \sqrt{u}$ is not Lipschitz continuous in a neighborhood of $u = 0$:

$$\left| f(u) - f(0) \right| = \sqrt{u} = \frac{1}{\sqrt{u}} \cdot u$$

which is not bounded by a constant multiple of u near $u = 0$.

Let us look at the consequence of violating LC.

We use the method of separation of variables to solve it exactly.

$$\frac{du}{dt} = \sqrt{u}$$

==> $$\frac{1}{\sqrt{u}} du = dt$$

Caution:

when we divide both sides by $\sqrt{u}$, we restrict ourselves to non-zero solutions.

Integrating both sides, we get

$$2\sqrt{u} = t + c$$

==> $$u = \frac{(t+c)^2}{4}$$

Enforcing the initial condition $u(0) = 0$, we obtain

$$u(t) = \frac{t^2}{4}$$

It is straightforward to verify that $u(t) \equiv 0$ is also a solution.

==> Solution is not unique.

Numerical differentiation

Goal: to approximate $f'(x)$, $f''(x)$

A first order method for $f'(x)$

Recall the definition $f'(x) = \lim\limits_{h \to 0} \dfrac{f(x+h) - f(x)}{h}$

For small h, we have $f'(x) \approx \dfrac{f(x+h) - f(x)}{h}$

We adopt the general formulation below for all numerical approximations

$$\underbrace{\frac{f(x+h)-f(x)}{h}}_{\text{Numerical approximation}} = \underbrace{f'(x)}_{\text{Exact}} + \underbrace{e(h)}_{\text{Discretization error}}$$

Question: How to find the order of $e(h)$?

Answer: Taylor expansion

$$f(x+h) = f(x) + f'(x)h + f''(x)\frac{h^2}{2} + \cdots$$

==> $$e(h)=\frac{f(x+h)-f(x)}{h}-f'(x)=\frac{f'(x)h+f''(x)\frac{h^2}{2}+\cdots}{h}-f'(x)$$

$$=f''(x)\frac{h}{2}+\cdots=C\cdot h+\cdots=O(h)$$

Another first order method for $f'(x)$

$$\underbrace{\frac{f(x)-f(x-h)}{h}}_{\text{Numerical approximation}}=\underbrace{f'(x)}_{\text{Exact}}+\underbrace{e(h)}_{\text{Discretization error}}$$

$$e(h)=-f''(x)\frac{h}{2}+\cdots=C\cdot h+\cdots=O(h)$$

A second order method for $f'(x)$

$$\underbrace{\frac{f(x+h)-f(x-h)}{2h}}_{\text{Numerical approximation}}=\underbrace{f'(x)}_{\text{Exact}}+\underbrace{e(h)}_{\text{Discretization error}}$$

$$e(h)=f'''(x)\frac{h^2}{6}+\cdots=C\cdot h^2+\cdots=O(h^2)$$

A second order method for $f''(x)$

$$\underbrace{\frac{f(x+h)-2f(x)+f(x-h)}{h^2}}_{\text{Numerical approximation}}=\underbrace{f''(x)}_{\text{Exact}}+\underbrace{e(h)}_{\text{Discretization error}}$$

$$e(h)=f^{(4)}(x)\frac{h^2}{12}+\cdots=C\cdot h^2+\cdots=O(h^2)$$

Numerical integration

Goal: to approximate $\int_a^b f(x)dx$

Divide $[a, b]$ into N sub-intervals

$$h=\frac{b-a}{N}$$

$x_j = a + j\,h$

$x_0 = a, \qquad x_1 = a + h, \ \ldots, \qquad x_N = b$

$$\int_a^b f(x)dx = \sum_{j=1}^{N} \int_{x_{j-1}}^{x_j} f(x)dx$$

We only need to work out each piece (integral over a sub-interval).

<u>Trapezoidal rule:</u>

We use a linear approximation for $f(x)$.

$$f(x) \approx a\,x + b \quad \text{for } x \in [x_{j-1}, x_j]$$

…

$$\Longrightarrow \qquad \int_{x_{j-1}}^{x_j} f(x)dx \approx \left(f(x_{j-1}) + f(x_j)\right)\frac{h}{2}$$

<u>Notation:</u> $f_j \equiv f(x_j)$

We write the numerical approximation into <u>the general formulation</u>.

$$\underbrace{\left(f_{j-1} + f_j\right)\frac{h}{2}}_{\text{Numerical approximation}} = \underbrace{\int_{x_{j-1}}^{x_j} f(x)dx}_{\text{Exact}} + \underbrace{e_j(h)}_{\text{Discretization error}}$$

We can show that

$$e_j(h) = O(h^3)$$

To approximate $\int_a^b f(x)dx$, we sum over all sub-intervals.

<u>Composite trapezoidal rule</u>

$$\left(f_{j-1} + f_j\right)\frac{h}{2} = \int_{x_{j-1}}^{x_j} f(x)dx + e_j(h) \qquad \text{(Trapezoidal rule for sub-interval } j\text{)}$$

$$\Longrightarrow \qquad \sum_{j=1}^{N}\left(f_{j-1} + f_j\right)\frac{h}{2} = \sum_{j=1}^{N}\int_{x_{j-1}}^{x_j} f(x)dx + \sum_{j=1}^{N} e_j(h)$$

$$\Longrightarrow \qquad \underbrace{\left(f_0 + f_N + 2\sum_{j=1}^{N-1} f_j\right)\frac{h}{2}}_{\text{Numerical approximation}} = \underbrace{\int_a^b f(x)dx}_{\text{Exact}} + \underbrace{E(h)}_{\text{Discretization error}}$$

where the error $E(h)$ is

$$E(h)=\sum_{j=1}^{N} e_j(h)=N\cdot O(h^3)=O(h^2)$$

Composite trapezoidal rule is a <u>second order</u> method for approximating $\int_a^b f(x)dx$.

<u>Simpson's rule:</u>

We use a quadratic approximation for $f(x)$.

$$f(x)\approx ax^2+bx+c \quad \text{for } x\in[x_{j-1},x_j]$$

...

$$\Longrightarrow \quad \underbrace{\left(f_{j-1}+f_j+4f_{j-1/2}\right)\frac{h}{6}}_{\substack{\text{Numerical}\\ \text{approximation}}} = \underbrace{\int_{x_{j-1}}^{x_j} f(x)dx}_{\text{Exact}} + \underbrace{e_j(h)}_{\substack{\text{Discretization}\\ \text{error}}}$$

<u>Notation:</u> $f_{j-1/2}=f(x_{j-1/2})$, $x_{j-1/2}=a+\left(j-\frac{1}{2}\right)h$

We can show that

$$e_j(h)=O(h^5)$$

We sum over all sub-intervals and obtain the composite Simpson's rule.

<u>Composite Simpson's rule:</u>

$$\underbrace{\left(f_0+f_N+2\sum_{j=1}^{N-1}f_j+4\sum_{j=1}^{N}f_{j-1/2}\right)\frac{h}{6}}_{\substack{\text{Numerical}\\ \text{approximation}}} = \underbrace{\int_a^b f(x)dx}_{\text{Exact}} + \underbrace{E(h)}_{\substack{\text{Discretization}\\ \text{error}}}$$

where the error $E(h)$ is

$$E(h)=\sum_{j=1}^{N} e_j(h)=N\cdot O(h^5)=O(h^4)$$

Composite Simpson's rule is a <u>fourth order</u> method for approximating $\int_a^b f(x)dx$.

<u>Caution:</u>

Be careful with indices when implementing the composite Simpson's rule.

There are (N-1) <u>internal</u> grid-points in summation $\sum_{j=1}^{N-1} f_j$

There are N middle points of sub-intervals in summation $\sum_{j=1}^{N} f_{j-1/2}$

<u>Advantage of higher order methods</u>:

In numerical integration, when we refine the numerical resolution,

- Grid size: h ----> $h/2$
- Number of points: N ----> $2N$ (approximately)
  The amount of computation is doubled
- Error: $E(h)$ ----> $E(h/2)$

For a second order method,

$$E\left(\frac{h}{2}\right) \approx C \cdot \left(\frac{h}{2}\right)^2 = \frac{1}{2^2} C \cdot h^2 \approx \frac{1}{4} E(h)$$

The error is reduced by a factor of 4 (approximately).

For a fourth order method,

$$E\left(\frac{h}{2}\right) \approx C \cdot \left(\frac{h}{2}\right)^4 = \frac{1}{2^4} C \cdot h^4 \approx \frac{1}{16} E(h)$$

The error is reduced by a factor of 16 (approximately).

Numerical error estimation

We discuss it in the general formulation for all numerical approximations.

Let $T(h)$ be a numerical approximation to quantity I, obtained with step size h using a p-th order numerical method. Exact value of quantity I is unknown.

$$\underbrace{T(h)}_{\substack{\text{Numerical}\\ \text{approximation}}} = \underbrace{I}_{\text{Exact}} + \underbrace{E(h)}_{\text{Error}}$$

$$E(h) = C \cdot h^p + \cdots \approx C \cdot h^p$$

<u>Goal:</u> to estimate the error $E(h)$ when quantity I is unknown.

<u>Example:</u> We write the composite trapezoidal rule as

$$\underbrace{T(h)}_{\substack{\text{Numerical}\\\text{approximation}}} = \underbrace{I}_{\text{Exact}} + \underbrace{E(h)}_{\text{Error}}$$

where

$$\underbrace{I}_{\text{Exact}} = \int_a^b f(x)dx\,, \qquad \underbrace{T(h)}_{\substack{\text{Numerical}\\\text{approximation}}} = \left(f_0 + f_N + 2\sum_{j=1}^{N-1} f_j\right)\frac{h}{2}$$

We like to estimate the error in $T(h)$ without knowing the value of quantity I.

Strategy:

We carry out computation with two numerical resolutions.

Specifically, we calculate $T(h)$ and $T(h/2)$.

$$T(h) = I + C \cdot h^p + \cdots$$

$$T\left(\frac{h}{2}\right) = I + C \cdot \left(\frac{h}{2}\right)^p + \cdots$$

==> $$T(h) - T\left(\frac{h}{2}\right) = C \cdot h^p \left(1 - \frac{1}{2^p}\right) + \cdots$$

==> $$C \cdot h^p \approx \frac{T(h) - T\left(\frac{h}{2}\right)}{1 - \frac{1}{2^p}}$$

$$\boxed{E(h) \approx \frac{T(h) - T\left(\frac{h}{2}\right)}{1 - \frac{1}{2^p}}}$$

This is the method for estimating the error numerically when the exact quantity I is unknown. If the order of the method p is also unknown, we simply use

$$E(h) \approx T(h) - T\left(\frac{h}{2}\right)$$

This simple formula may under-estimate the error by a factor of up to 2.

$$\frac{1}{1 - \frac{1}{2^p}} = 2 \qquad \text{when } p = 1$$

$$\frac{1}{1-\frac{1}{2^p}}=\frac{4}{3} \qquad \text{when } p = 2$$

$$\frac{1}{1-\frac{1}{2^p}}=\frac{16}{15} \qquad \text{when } p = 4$$