MA 8020: Numerical Analysis II
Numerical Ordinary Differential Equations

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 320317, Taiwan
E-mail: syyang@math.ncu.edu.tw
http://www.math.ncu.edu.tw/∼syyang/

First version: April 06, 2019 Last updated: May 8, 2024

Initial-value problem (IVP)

- **Initial-value problem:** find $x(t)$ such that

$$\begin{cases} x'(t) = f(t,x), \\ x(t_0) = x_0, \end{cases}$$

  where $f(t,x), t_0, x_0 \in \mathbb{R}^1$ are given.

- **Example 1:**

$$\begin{cases} x'(t) = x \tan(t+3), \\ x(-3) = 1. \end{cases}$$

  One can verify that the analytic solution of this IVP is $x(t) = \sec(t+3)$. Since $\sec t$ becomes ∞ at $t = \pm\frac{\pi}{2}$, the solution is valid only for $-\frac{\pi}{2} < t+3 < \frac{\pi}{2}$.

- **Example 2:**

$$\begin{cases} x'(t) = x, \\ x(0) = 1. \end{cases}$$

  Try $x(t) = ce^{rt} \Rightarrow cre^{rt} = ce^{rt} \Rightarrow r = 1,\ x = ce^{t}$ (general solution)

  Use $x(0) = 1 \Rightarrow x = e^{t}$ (a particular solution)

Existence of solution

- **Existence:** do all IVPs have a solution? **Answer:** No!

  Some assumptions must be made about f, and even then we can expect the solution to exist only in a neighborhood of $t = t_0$.

- **Example:**

$$\begin{cases} x'(t) &= 1 + x^2, \\ x(0) &= 0. \end{cases}$$

  Try $x(t) = \tan t$, then $x(0) = 0$.

  LHS: $(\tan t)' = \dfrac{\cos^2 t + \sin^2 t}{\cos^2 t}$; RHS: $1 + \tan^2 t = 1 + \dfrac{\sin^2 t}{\cos^2 t}$.

  Hence $x(t) = \tan t$ is a solution of the IVP.

- If $t \to (\pi/2)^-$ then $x(t) \to \infty$. For the solution starting at $t = 0$, it has to "stop the clock" before $t = \pi/2$. Here we can only say that there exists a solution for a limited time.

Existence theorem

Consider the IVP:

$$\begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0, \end{cases}$$

If f is continuous in a rectangle R centered at (t_0, x_0), say

$$R = \{(t,x) : \ |t - t_0| \le \alpha, \ |x - x_0| \le \beta\},$$

then the IVP has a solution $x(t)$ for

$$|t - t_0| \le \min\{\alpha, \beta/M\},$$

where M is maximum of $|f(t,x)|$ in the rectangular R.

Example

Prove that

$$\begin{cases} x'(t) &= (t+\sin x)^2, \\ x(0) &= 3 \end{cases}$$

has a solution in the interval $-1 \le t \le 1$.

Solution:

(1) Consider $f(t,x) = (t+\sin x)^2$, where $(t_0, x_0) = (0,3)$.

(2) Let $R = \{(t,x) : |t| \le \alpha, |x-3| \le \beta\}$. Then $|f(t,x)| \le (\alpha+1)^2 := M$.

(3) We want $|t-0| \le 1 \le \min\{\alpha, \beta/M\}$.

(4) Let $\alpha = 1$ then $M = (1+1)^2 = 4$ and force $\beta \ge 4$. By the existence theorem, the IVP has a solution in the interval $|t-t_0| \le \min\{\alpha, \beta/M\} = 1$, that is, $-1 \le t \le 1$. $\square$

Uniqueness

- If f is continuous, we may still have more than one solution, e.g.,

$$\begin{cases} x'(t) = x^{2/3}, \\ x(0) = 0. \end{cases}$$

  Note that $x(t) = 0$ is a solution for all t. Another solution is $x(t) = t^3/27$.

- To have a unique solution, we need to assume somewhat more about f.

Uniqueness theorem

Consider the IVP:

$$\begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0. \end{cases}$$

If f *and* $\frac{\partial f}{\partial x}$ *are continuous in the rectangle R centered at* (t_0, x_0),

$$R = \{(t,x) : |t - t_0| \leq \alpha, |x - x_0| \leq \beta\},$$

then the IVP has a unique solution $x(t)$ *for*

$$|t - t_0| \leq \min\{\alpha, \beta/M\},$$

where M is maximum of $|f(t,x)|$ *in the rectangular R.*

Another uniqueness theorem

Consider the IVP:

$$\begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0, \end{cases}$$

If f is continuous in $a \le t \le b$, $-\infty < x < \infty$ and satisfies

$$|f(t,x_1) - f(t,x_2)| \le L|x_1 - x_2|, \qquad (\star)$$

then the IVP has a unique solution $x(t)$ in the interval $[a,b]$.

Note: $(\star)$ is called the Lipschitz condition of $f(t,x)$ in the variable x.

Example

Prove that

$$\begin{cases} x'(t) &= 1 + t\sin(tx), \\ x(0) &= 0 \end{cases}$$

has a solution on the interval $0 \le t \le 2$.

Solution:

(1) Since $f(t,x) = 1 + t\sin(tx)$, we have $|\frac{\partial f}{\partial x}(t,x)| = |t^2\cos(tx)| \le 4$ for $0 \le t \le 2$ and $-\infty < x < \infty$.

(2) By the mean value theorem, $\exists\ \xi$ between x_1 and x_2 such that

$$f(t,x_2) - f(t,x_1) = \frac{\partial f(t,\xi)}{\partial x}(x_2 - x_1).$$

$$\Longrightarrow |f(t,x_2) - f(t,x_1)| \le 4|x_2 - x_1|.$$

$\Longrightarrow f$ satisfies $(\star)$ with $L = 4$ and f is continuous in $0 \le t \le 2$, $-\infty < x < \infty$.

$\Longrightarrow$ the IVP has a unique solution $x(t)$ for $a \le t \le b$. $\square$

Numerical methods

- Consider the IVP:

$$\begin{cases} x'(t) & = & f(t,x), \\ x(t_0) & = & x_0. \end{cases}$$

- **Strategy:** instead of finding $x(t)$ for all t in some interval containing t_0, we approximate $x(t)$ at some discrete points.

(insert a graph here!)

Taylor-series method

- For the Taylor-series method, it is necessary to assume that various partial derivatives of f exist.
- We use a concrete example to illustrate the method. Consider an IVP as

$$\begin{cases} x'(t) &= \cos t - \sin x + t^2, \\ x(-1) &= 3. \end{cases}$$

- Assume that we know $x(t)$ and we wish to compute $x(t+h)$. By the Taylor expansion of x, we have

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x''(t) + \frac{h^3}{3!}x'''(t) + \frac{h^4}{4!}x^{(4)}(t) + O(h^5).$$

Taylor-series method (cont'd)

- How to compute $x'(t)$, $x''(t)$, $x'''(t)$ and $x^{(4)}(t)$?

$$\begin{cases} x'(t) &= \cos t - \sin x + t^2, \\ x''(t) &= -\sin t - (\cos x)x' + 2t, \\ x'''(t) &= -\cos t + \sin x(x')^2 - (\cos x)x'' + 2, \\ x^{(4)}(t) &= \sin t + (\cos x)(x')^3 + 3(\sin x)x'x'' - (\cos x)x'''. \end{cases}$$

- If we truncate at h^4 then the local truncation error for obtaining $x(t+h)$ is $O(h^5)$. We say the method is of order 4.
- **Definition:** *The order of the Taylor-series method is n if terms up to and include $h^n x^{(n)}(t)/n!$ are used.*
- Let $t_k := t_0 + kh$ and $x_k \approx x(t_k)$. Then the Taylor-series method for this example is defined as

$$x_{k+1} = x_k + h\widetilde{x}'(t_k) + \frac{h^2}{2!}\widetilde{x}''(t_k) + \frac{h^3}{3!}\widetilde{x}'''(t_k) + \frac{h^4}{4!}\widetilde{x}^{(4)}(t_k),\ k \geq 0,$$

$$\widetilde{x}'(t_k) := f(t_k, x_k),\ \widetilde{x}''(t_k) := f_t(t_k, x_k) + f_x(t_k, x_k)f(t_k, x_k), \cdots.$$

Algorithm

Starting $t = -1$ with $h = 0.01$, we can compute the solution in $[-1, 1]$ with 200 steps:

input $M \leftarrow 200, h \leftarrow 0.01, t \leftarrow -1, x \leftarrow 3$
output $0, t, x$
for $k = 1$ **to** M **do**

$$
\begin{aligned}
x' &\leftarrow \cos t - \sin x + t^2 \\
x'' &\leftarrow -\sin t - (\cos x)x' + 2t \\
x''' &\leftarrow -\cos t + \sin x(x')^2 - (\cos x)x'' + 2 \\
x^{(4)} &\leftarrow \sin t + (\cos x)(x')^3 + 3(\sin x)x'x'' - (\cos x)x''' \\
x &\leftarrow x + h(x' + \tfrac{h}{2}(x'' + \tfrac{h}{3!}(x''' + \tfrac{h}{4!}x^{(4)}))) \\
t &\leftarrow t + h
\end{aligned}
$$

output k, t, x
end do

Error estimate

- The estimate of the local truncation error is given by

$$E_n := \frac{1}{(n+1)!} h^{n+1} x^{(n+1)}(t+\theta h) \quad \text{for some } \theta \in (0,1).$$

Hence

$$E_4 = \frac{1}{5!} h^5 x^{(5)}(t+\theta h) \quad \text{for some } \theta \in (0,1).$$

- We can replace $x^{(5)}(t+\theta h)$ by a simple finite difference,

$$E_4 \approx \frac{1}{5!} h^5 \left(\frac{x^{(4)}(t+h) - x^{(4)}(t)}{h} \right) = \frac{h^4}{120} \left(x^{(4)}(t+h) - x^{(4)}(t) \right).$$

- Suppose that the local truncation error (LTE) is $O(h^{n+1})$.
  An error of this sort is present in each step of the numerical solution. The accumulation of all LTEs gives the global truncation error (GTE). Roughly speaking, we have

$$GTE \approx \frac{T - t_0}{h} O(h^{n+1}) = O(h^n),$$

and we say the numerical method is of $O(h^n)$.

Advantages and disadvantages of the Taylor-series method

- **Disadvantages:**
  (1) The method depends on repeated differentiation of the differential equation, unless we intend to use only the method of order 1.
  $\Longrightarrow f(t,x)$ must have partial derivatives of sufficient high order in the region where are solving the problem. Such an assumption is not necessary for the existence of a solution.
  (2) The various derivatives formula need to be programmed.

- **Advantages:**
  (1) Conceptual simplicity.
  (2) Potential for high precision: If we get, e.g. 20 derivatives of $x(t)$, then the method is order 20 (i.e., terms up to and including the one involving h^{20}).

Euler's method (Taylor-series method of order 1)

- If $n = 1$, the Taylor series method reduces to Euler's method.

$$x_{k+1} = x_k + hf(t_k, x_k), \quad k \geq 0.$$

Disadvantage of the method is that the necessity of taking small value for h to gain acceptable precision.

Advantage is not to require any differentiation of f.

- **In-class exercise:** Consider the following IVP:

$$\begin{cases} x'(t) & = & \cos t - \sin x + t^2, \\ x(0) & = & 3. \end{cases}$$

Derive Euler's method based on the Taylor series and compute $x(0.1)$ when $h = 0.1$.

Basic concepts of Runge-Kutta methods

We wish to approximate the following IVP:

$$\begin{cases} x'(t) = f(t,x), \\ x(t_0) = x_0. \end{cases}$$

- Suppose that f is sufficiently smooth. From the Taylor theorem, we have

$$x(t+h) = x(t) + hx'(t) + \frac{h^2}{2!}x''(t) + O(h^3).$$

- By the chain rule, we obtain

$$x''(t) = f_t(t,x) + f_x(t,x)x'(t) = f_t(t,x) + f_x(t,x)f(t,x).$$

Basic concepts of Runge-Kutta methods (cont'd)

- In the Taylor expansion, we have

$$\begin{aligned} x(t+h) &= x(t) + hf(t,x) + \frac{h^2}{2}(f_t(t,x) + f_x(t,x)f(t,x)) + O(h^3) \\ &= x(t) + \frac{h}{2}f(t,x) + \frac{h}{2}\Big[f(t,x) + hf_t(t,x) + hf_x(t,x)f(t,x)\Big] \\ &\quad +O(h^3) \\ &= x(t) + \frac{h}{2}f(t,x) + \frac{h}{2}f(t+h, x+hf(t,x)) + O(h^3). \end{aligned}$$

- Note that the last equality above is valid by the Taylor expansion in two variables,

$$f(t+h, x+hf(t,x)) = f(t,x) + hf_t(t,x) + hf(t,x)f_x(t,x) + O(h^2).$$

A second-order Runge-Kutta method

- Then a 2nd-order Runge-Kutta (RK) method is given by

$$x(t+h) \approx x(t) + \frac{h}{2} f(t,x) + \frac{h}{2} f(t+h, x+hf(t,x)),$$

or alternating

$$\begin{aligned} x(t+h) &\approx x(t) + \frac{1}{2}(F_1 + F_2), \\ F_1 &= hf(t,x), \\ F_2 &= hf(t+h, x+F_1). \end{aligned}$$

It is also known as Heun's method.

- In practice, let $x_n \approx x(t_n)$, then we define Heun's method as

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{2}(F_1 + F_2), \quad n \geq 0, \\ F_1 &:= hf(t_n, x_n), \\ F_2 &:= hf(t_{n+1}, x_n + F_1). \end{aligned}$$

The general second-order Runge-Kutta method

- In general, the 2nd order RK method needs

$$\begin{aligned} x(t+h) &= x(t) + \omega_1 hf + \omega_2 hf(t+\alpha h, x+\beta hf) + O(h^3), \\ &= x(t) + \omega_1 hf + \omega_2 h\left[f + \alpha h f_t + \beta h f f_x\right] + O(h^3). \end{aligned}$$

- Comparing with

$$x(t+h) = x(t) + hf + \frac{h^2}{2}(f_t + f_x f) + O(h^3),$$

we have

$$\begin{aligned} \omega_1 + \omega_2 &= 1, \\ \omega_2 \alpha &= 1/2, \\ \omega_2 \beta &= 1/2. \end{aligned}$$

Modified Euler method

- The previous method (Heun's method) is obtained by setting

$$\begin{cases} \omega_1 = \omega_2 = 1/2, \\ \alpha = \beta = 1. \end{cases}$$

- Setting

$$\begin{cases} \omega_1 = 0, \\ \omega_2 = 1, \\ \alpha = \beta = 1/2, \end{cases}$$

  we obtain the following modified Euler method:

$$\begin{aligned} x_{n+1} &= x_n + F_2, \quad n \geq 0, \\ F_1 &:= hf(t_n, x_n), \\ F_2 &:= hf(t_n + \frac{1}{2}h, x_n + \frac{1}{2}F_1). \end{aligned}$$

Fourth-order RK methods

- The derivations of higher order RK methods are tedious. However, the formulas are rather elegant and easily programmed once they have been derived.
- The most popular 4th order RK is:

$$\begin{aligned} x(t+h) &\approx x(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4), \\ F_1 &= hf(t,x), \qquad F_2 = hf(t+\frac{h}{2}, x+\frac{1}{2}F_1), \\ F_3 &= hf(t+\frac{h}{2}, x+\frac{1}{2}F_2), \qquad F_4 = hf(t+h, x+F_3). \end{aligned}$$

That is, the 4th order RK is defined as

$$\begin{aligned} x_{n+1} &= x_n + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4), \quad n \geq 0, \\ F_1 &:= hf(t_n, x_n), \qquad F_2 := hf(t_n+\frac{h}{2}, x_n+\frac{1}{2}F_1), \\ F_3 &:= hf(t_n+\frac{h}{2}, x_n+\frac{1}{2}F_2), \qquad F_4 := hf(t_{n+1}, x_n+F_3). \end{aligned}$$

Homework

- Use the most popular 4th order RK with $h = 1/128$ to solve the following IVP for $t \in [1, 3]$ and then plot the piecewise linear approximate solution:

$$\begin{cases} x'(t) &= t^{-2}(tx - x^2), \\ x(1) &= 2. \end{cases}$$

- Also plot the exact solution:

$$x(t) = (1/2 + \ln t)^{-1} t.$$

Algorithm

input $M \leftarrow 256, t \leftarrow 1.0, h \leftarrow 0.0078125, x \leftarrow 2.0$
define $f(t,x) = (tx - x^2)/t^2$
define $u(t) = t/(1/2 + \ln t)$
$e \leftarrow |u(t) - x|$
output $0, t, x, e$
for $k = 1$ **to** M **do**

$$\begin{aligned}
F_1 &\leftarrow hf(t,x) \\
F_2 &\leftarrow hf(t + \tfrac{h}{2}, x + \tfrac{1}{2}F_1) \\
F_3 &\leftarrow hf(t + \tfrac{h}{2}, x + \tfrac{1}{2}F_2) \\
F_4 &\leftarrow hf(t + h, x + F_3) \\
x &\leftarrow x + \tfrac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4) \\
t &\leftarrow t + h \\
e &\leftarrow |u(t) - x|
\end{aligned}$$

output k, t, x, e
end do

How to estimate the local truncation error of RK4?

- For RK4, the local truncation error is of $O(h^5)$. The local truncation error at the first step is

$$x^*(t_0 + h) - x_1 = O(h^5),$$

where $x^*(t_0 + h)$ is exact value and x_1 is computed value. That is, the truncation error behaves like Ch^5 for small h. Here C is a number independent of h but dependent on t_0 and x^*.

- Let v be the value of the approximate solution at $t_0 + h$ obtained by taking one step of length h from t_0. Let u be the approximate solution at $t_0 + h$, obtained by taking two steps of size $h/2$ from t_0. Then we have

$$x^*(t_0 + h) \approx v + Ch^5 \quad \text{and} \quad x^*(t_0 + h) \approx u + 2C(h/2)^5.$$

By substraction, we obtain

$$\text{local truncation error} = Ch^5 \approx \frac{u - v}{1 - 2^{-4}} \approx u - v.$$

Basic concepts of multistep methods

- Taylor-series and RK methods are examples of single-step methods, i.e. use information only at t to get $t+h$.
- Consider the IVP: $x'(t) = f(t,x)$ and $x(t_0) = x_0$. Assume that we want to approximate $x(t)$ at $t_0, t_1, \cdots, t_i, \cdots$. Let x_i be the approximate solution of $x(t_i)$. Then by the *Fundamental Theorem of Calculus*, we have

$$\int_{t_n}^{t_{n+1}} x'(t)dt = x(t_{n+1}) - x(t_n)$$

  and then

$$x(t_{n+1}) - x(t_n) = \int_{t_n}^{t_{n+1}} f(t, x(t))dt.$$

- One of the basic idea of the multistep method is to interpolate the integrand $f(t, x(t))$ by using $t_n, t_{n-1}, \cdots$. Then we have

$$x_{n+1} = x_n + af_n + bf_{n-1} + cf_{n-2} + \cdots, \quad \text{where } f_i := f(t_i, x_i).$$

  An equation of this type is called an Adams-Bashforth formula.

Adams-Bashforth formula of order 5

- To derive the A.-B. formula of order 5, we consider (on equally spaced points: $t_i = t_0 + ih$)

$$\int_{t_n}^{t_{n+1}} f(t, x(t))dt \approx h\Big(Af_n + Bf_{n-1} + Cf_{n-2} + Df_{n-3} + Ef_{n-4}\Big).$$

- We wish the numerical integration is exact for polynomials of degree ≤ 4.

  Without loss of generality, we may consider $t_n = 0$ and $h = 1$ ($\Rightarrow t_{n+1} = 1$).

  Then apply the method of undetermined coefficients.

Adams-Bashforth formula of order 5 (cont'd)

- As a basis for Π_4, we consider

$$\begin{aligned} p_0(t) &= 1, \\ p_1(t) &= t, \\ p_2(t) &= t(t+1), \\ p_3(t) &= t(t+1)(t+2), \\ p_4(t) &= t(t+1)(t+2)(t+3). \end{aligned}$$

- For each of these polynomials the following formula should be exact

$$\begin{aligned} \int_0^1 p_n(t)dt &= Ap_n(0) + Bp_n(-1) + Cp_n(-2) \\ &\quad + Dp_n(-3) + Ep_n(-4). \end{aligned}$$

Adams-Bashforth formula of order 5 (cont'd)

- By direct computations, we have

$$\begin{aligned} p_0(t) &= 1 \Longrightarrow \quad A+B+C+D+E=1, \\ p_1(t) &= t \Longrightarrow \quad -B-2C-3D-4E=1/2, \\ p_2(t) &= t(t+1) \Longrightarrow \quad C+6D+12E=5/6, \\ p_3(t) &= t(t+1)(t+2) \Longrightarrow \quad -6D-24E=9/4, \\ p_4(t) &= t(t+1)(t+2)(t+3) \Longrightarrow \quad 24E=251/30. \end{aligned}$$

  By backward substitution, we obtain

$$E=\frac{251}{720}, \quad D=-\frac{1274}{720}, \quad C=\frac{2616}{720}, \quad B=-\frac{2774}{720}, \quad A=\frac{1901}{720}.$$

- Therefore, we have for $n \geq 4$

$$x_{n+1}=x_n+\frac{1}{720}\left(1901f_n-2774f_{n-1}+2616f_{n-2}-1274f_{n-3}+251f_{n-4}\right),$$

$x_n \approx x(t_n)=x(0)$, $x_{n+1}\approx x(t_{n+1})=x(1)$, and $f_i := f(t_i, x_i)$.

Adams-Bashforth formula of order 5 (cont'd)

- We need to change the interval from $[0,1]$ to $[t_n, t_{n+1}]$ with

$$\lambda(s) = \frac{t_{n+1} - t_n}{1-0} s + \frac{t_n}{1-0} = hs + t_n.$$

Then $\lambda'(s) = h$. Hence,

$$\int_{t_n}^{t_{n+1}} f(t,x)dt = \int_0^1 f(\lambda(s), x(\lambda(s)))\lambda'(s)ds.$$

- Finally, we have the Adams-Bashforth formula of order 5: for $n \geq 4$

$$x_{n+1} = x_n + \frac{h}{720}\left(1901f_n - 2774f_{n-1} + 2616f_{n-2} - 1274f_{n-3} + 251f_{n-4}\right),$$

where $x_n \approx x(t_n)$, $x_{n+1} \approx x(t_{n+1})$, and $f_i := f(t_i, x_i)$.

Adams-Moulton formula of order 5

- A formula of the type

$$x_{n+1} = x_n + af_{n+1} + bf_n + cf_{n-1} + \cdots$$

  is called an Adams-Moulton formula, where $f_i := f(t_i, x_i)$.

- To derive the A.-M. formula of order 5, we consider (on equally spaced points: $t_i = t_0 + ih$)

$$\int_{t_n}^{t_{n+1}} f(t, x(t))dt \approx h\Big(Af_{n+1} + Bf_n + Cf_{n-1} + Df_{n-2} + Ef_{n-3}\Big).$$

- We wish the numerical integration is exact for polynomials of degree ≤ 4.

  Without loss of generality, we may consider $t_n = 0$ and $h = 1$.

  Then apply the method of undetermined coefficients.

Adams-Moulton formula of order 5 (cont'd)

- As a basis for Π_4, we consider

$$\begin{aligned} p_0(t) &= 1, \\ p_1(t) &= t-1, \\ p_2(t) &= (t-1)t, \\ p_3(t) &= (t-1)t(t+1), \\ p_4(t) &= (t-1)t(t+1)(t+2). \end{aligned}$$

- For each of these polynomial the following formula should be exact

$$\begin{aligned} \int_0^1 p_n(t)dt &= Ap_n(1) + Bp_n(0) + Cp_n(-1) \\ &\quad + Dp_n(-2) + Ep_n(-3). \end{aligned}$$

Adams-Moulton formula of order 5 (cont'd)

- Thus we have

$$\begin{aligned} p_0(t) &= 1 \Longrightarrow \quad A+B+C+D+E=1, \\ p_1(t) &= t-1 \Longrightarrow \quad -B-2C-3D-4E=-1/2, \\ p_2(t) &= t^2-t \Longrightarrow \quad 2C+6D+12E=-1/6, \\ p_3(t) &= t^3-t \Longrightarrow \quad -2D-24E=-1/4, \\ p_4(t) &= t^4+2t^3-t^2-2t \Longrightarrow \quad 2E=-19/30. \end{aligned}$$

  By backward substitution, we obtain

$$E=-\frac{19}{720},\quad D=\frac{106}{720},\quad C=-\frac{264}{720},\quad B=\frac{646}{720},\quad A=\frac{251}{720}.$$

- By changing of variable, we finally have

$$x_{n+1}=x_n+\frac{h}{720}\Big(251f_{n+1}+646f_n-264f_{n-1}+106f_{n-2}-19f_{n-3}\Big).$$

A predictor-corrector method

- In multistep methods, we need a start-up method to get started. Here, for example, we can use RK method of order 4 to get x_1, x_2, x_3, x_4.
- Note that in the A.-M. method, x_{n+1} occurs on both sides of the equation! $\because f_{n+1} = f(t_{n+1}, x_{n+1})$.
- **First strategy:**

  use the A.-B. formula of order 5 as a predictor to compute x^*_{n+1} and then use the A.-M formula of order 5 as corrector with $f_{n+1} = f(t_{n+1}, x^*_{n+1})$.

  This method is known as a predictor-corrector method.

Second strategy: a fixed-point method

- Define the mapping

$$\varphi(z) := \frac{251}{720} hf(t_{n+1}, z) + T,$$

  where T is composed of all the other terms in the A.-M. formula.

- Then this reduces to a fixed-point problem:

$$z_{k+1} = \varphi(z_k) = \frac{251}{720} hf(t_{n+1}, z_k) + T \quad (k \geq 0).$$

  It will converge to a fixed point of φ under appropriate hypotheses.

- Thus, if ξ is the fixed point, z_0 should be in the interval centered at ξ such that $|\phi'(z)| < 1$, where

$$\phi'(z) = \frac{251}{720} h \frac{\partial f(t_{n+1}, z)}{\partial z}.$$

  This can be made less than 1 by setting h is small enough.

Linear multistep methods

- Linear multistep methods (LMMs) are methods of the form

  $$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h\{b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}\}, \ (\star)$$

  where $a_k \neq 0, f_i := f(t_i, x_i)$ and $t_i = t_0 + ih$. This a k-step method if $a_0 \neq 0$ or $b_0 \neq 0$.

- $(\star)$ is used to compute x_n assuming that $x_{n-k}, \cdots, x_{n-1}$ are already known. If $b_k = 0$, the method is said to be explicit. Otherwise, the method is said to be implicit.

- To define the order of a linear multistep method, let us consider the linear functional L over differentiable functions $x(t)$,

  $$Lx = \sum_{i=0}^{k} \Big(a_i x(ih) - h b_i x'(ih) \Big). \quad \leftarrow \ \textit{local truncation error}$$

  Here we take $k = n$ for simplicity and assume the first value begins at $t = t_0 = 0$ rather than at $t = t_{n-k}$.

Analysis of linear multistep methods

- By using the Taylor series for x, one can express L as

$$Lx = d_0 x(0) + d_1 h x'(0) + d_2 h^2 x''(0) + \cdots$$

- To compute the coefficients, d_i, we write the Taylor series for x and x':

$$x(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j)}(0) \quad \text{and} \quad x'(ih) = \sum_{j=0}^{\infty} \frac{(ih)^j}{j!} x^{(j+1)}(0).$$

- By the comparison of coefficients, we obtain

$$\begin{aligned} d_0 &= \sum_{i=0}^{k} a_i, \qquad d_1 = \sum_{i=0}^{k} (i a_i - b_i), \qquad d_2 = \sum_{i=0}^{k} (\frac{1}{2} i^2 a_i - i b_i), \\ &\vdots \\ d_j &= \sum_{i=0}^{k} \left\{ \frac{i^j}{j!} a_i - \frac{i^{j-1}}{(j-1)!} b_i \right\} \quad (j \geq 1). \end{aligned}$$

Theorem on linear multistep method

The following three properties of the linear multistep method are equivalent:

1. $d_0 = d_1 = \cdots = d_m = 0$.
2. $Lp = 0$ for $p \in \Pi_m$.
3. Lx is $O(h^{m+1})$ for all $x \in C^{m+1}$.

Proof:

- $(1) \Rightarrow (2)$: Since $d_0 = d_1 = \cdots = d_m = 0$, we have $Lx = d_{m+1}h^{m+1}x^{(m+1)}(0) + \cdots$
  If $x \in \Pi_m$ then $x^{(m+1)} = x^{(m+2)} = \cdots = 0$, which implies $Lx = 0$.
- $(2) \Rightarrow (3)$: If $x \in C^{m+1}$, then Taylor theorem implies $x = p + r$, where $p \in \Pi_m$ and r is a function with $r^{(k)}(0) = 0$ for $0 \leq k \leq m$. Hence $Lx = Lr = d_{m+1}h^{m+1}r^{m+1}(0) + \cdots = O(h^{m+1})$.
- $(3) \Rightarrow (1)$: $Lx = d_0x(0) + d_1hx'(0) + d_2h^2x''(0) + \cdots$ reduces $Lx = d_{m+1}h^{m+1}x^{(m+1)}(0) + \cdots$. Hence $d_0 = d_1 = \cdots = d_m = 0$.

□

Order of a linear multistep method

- Define the order of an LMM to be the number m such that

$$d_0 = d_1 = \cdots = d_m = 0 \neq d_{m+1}.$$

- **Example:** what is the order of the LMM:

$$x_n - x_{n-2} = \frac{1}{3}h(f_n + 4f_{n-1} + f_{n-2})?$$

Solution:

$(a_0, a_1, a_2) = (-1, 0, 1)$ and $(b_0, b_1, b_2) = (1/3, 4/3, 1/3)$.

$d_0 = d_1 = d_2 = d_3 = d_4 = 0.$

$d_5 = (1/120 a_1 - 1/24 b_1) + (4/15 a_2 - 2/3 b_2) = -1/90.$

The order of the method is 4.

Vector space of infinite sequences

- A complex sequence is a complex-valued function $x : \mathbb{N} \to \mathbb{C}$. We write $x = [x_1, x_2, \cdots, x_n, \cdots]$.
- Let V be the set of all infinite sequences of complex numbers. Then there is a 0 element in V, namely, $0 = [0, 0, 0, \cdots]$. We define two operations $+ : V \times V \to V$ and $\cdot : \mathbb{C} \times V \to V$., for $x = [x_1, x_2, \cdots, x_n, \cdots], y = [y_1, y_2, \cdots, y_n, \cdots] \in V$ and $\alpha \in \mathbb{C}$,

$$\begin{aligned} x + y &:= [x_1 + y_1, x_2 + y_2, \cdots, x_n + y_n, \cdots], \\ \alpha x &:= [\alpha x_1, \alpha x_2, \cdots, \alpha x_n, \cdots]. \end{aligned}$$

  or more compactly $(x + y)_n := x_n + y_n$ and $(\alpha x)_n := \alpha x_n$.
- *V is a vector space and its dimension is infinite.*

  The set of vectors is linearly independent: $\{v^{(1)} = [1, 0, 0, 0, \cdots],$ $v^{(2)} = [0, 1, 0, 0, \cdots], v^{(3)} = [0, 0, 1, 0, \cdots], \cdots\}$

Linear difference operator

- Consider the following linear operator $E : V \to V$ defined by

$$Ex = [x_2, x_3, x_4, \cdots], \quad \text{where } x = [x_1, x_2, x_3, x_4 \cdots].$$

  We call E the shift operator or displacement operator. Thus, $(Ex)_n = x_{n+1}$ and $(EEx)_n = x_{n+2}$. In general, $(E^k x)_n = x_{n+k}$.

- We define a linear difference operator as a linear combination of powers of E,

$$L = \sum_{i=0}^{m} c_i E^i,$$

  where E^0 is the identity operator, i.e., $(E^0 x)_n = (Ix)_n = x_n$.
  L is a polynomial in E, i.e., $L = p(E)$, where p is called the characteristic polynomial of L and defined by $p(\lambda) = \sum_{i=0}^{m} c_i \lambda^i$.

- The set $\{x \in V : Lx = 0\}$ is a linear subspace of V and it is called the null space (kernel) of L. So we need to find a basis that spans the null space in order to solve $Lx = 0$.

Example: $Lx = 0$

- Let

$$L = \sum_{i=0}^{m} c_i E^i, \quad \text{with } c_0 = 2, c_1 = -3, c_2 = 1, c_i = 0 \text{ for } i \geq 3.$$

  We have the linear difference equation, which can be written in three forms:

$$\begin{aligned} (E^2 - 3E^1 + 2E^0)x &= 0, \\ x_{n+2} - 3x_{n+1} + 2x_n &= 0 \quad (n \geq 1), \\ p(E)x &= 0 \quad p(\lambda) = \lambda^2 - 3\lambda + 2. \end{aligned}$$

- How to solve it? Putting $x_n = \lambda^n$, we get

$$\begin{aligned} \lambda^{n+2} - 3\lambda^{n+1} + 2\lambda^n &= 0 \\ \lambda^n p(\lambda) &= 0 \\ \lambda^n(\lambda - 1)(\lambda - 2) &= 0 \end{aligned}$$

Example: $Lx = 0$ (cont'd)

- $\lambda = 0$: trivial solution;
  $\lambda = 1$: $u_n := 1^n = 1$;
  $\lambda = 2$: $v_n := 2^n$.
  We can show that u_n and v_n form a basis for the solution space of $Lx = 0$, i.e., any solution is a linear combination of them

  $$x_n = \alpha \cdot 1 + \beta 2^n.$$

  (By induction, see page 30 for the details)
  Once we specify the starting values x_1 and x_2, then x_n is determined uniquely. In general, we have following theorem:
- **Theorem:** *If p is a polynomial and λ is a zero of p then one solution of the difference equation $p(E)x = 0$ is $[\lambda, \lambda^2, \lambda^3, \cdots]$. If all the zeros of p are simple and nonzero, then each solution of difference equation is a linear combination of such special solutions.*

  (see page 31 for the proof)

Multiple zeros

- Let $x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$. If p is any polynomial then

$$p(E)x(\lambda) = p(\lambda)x(\lambda).$$

  Differentiating with respect to λ, we get

$$p(E)x'(\lambda) = p'(\lambda)x(\lambda) + p(\lambda)x'(\lambda).$$

- If λ is a multiple zero of p, then $p(\lambda) = p'(\lambda) = 0$. Hence, $x(\lambda)$ and $x'(\lambda)$ are solutions of the difference equation $p(E)x = 0$. That is,

$$x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots] \quad \text{and} \quad x'(\lambda) = [1, 2\lambda, 3\lambda^2, \cdots]$$

  are solutions of $p(E)x = 0$.

- If $\lambda \neq 0$, then $x(\lambda)$ and $x'(\lambda)$ are linearly independent.

Multiple zeros (cont'd)

- Similarly, if λ is a zero of p having multiplicity k, then the following are solutions of the difference equation $p(E)x = 0$.

$$\begin{aligned} x(\lambda) &= [\lambda, \lambda^2, \lambda^3, \cdots], \\ x'(\lambda) &= [1, 2\lambda, 3\lambda^2, \cdots], \\ x''(\lambda) &= [0, 2, 6\lambda, \cdots], \\ &\vdots \\ x^{(k-1))}(\lambda) &= \frac{d^{(k-1)}}{d\lambda^{k-1}}[\lambda, \lambda^2, \lambda^3, \cdots]. \end{aligned}$$

- **Theorem:** *Let p be a polynomial satisfying $p(0) \neq 0$. Thus a basis for null space of $p(E)$ is obtained as follows: with each zero λ of p having multiplicity k, associate the k solutions, $x(\lambda), x'(\lambda), \cdots, x^{(k-1)}(\lambda)$, where $x(\lambda) = [\lambda, \lambda^2, \lambda^3, \cdots]$.*

An example

Find general solution of $4x_n + 7x_{n-1} + 2x_{n-2} - x_{n-3} = 0$.

Solution:

The characteristic polynomial is $p(\lambda) = 4\lambda^3 + 7\lambda^2 + 2\lambda - 1 = 0$.

Roots are $\lambda_1 = \lambda_2 = -1$ and $\lambda_3 = 1/4$.

The basic solutions are

$$\begin{aligned} x(-1) &= [-1, 1, -1, 1, \cdots], \\ x'(-1) &= [1, -2, 3, -4, \cdots], \\ x(1/4) &= [1/4, 1/16, 1/64, \cdots]. \end{aligned}$$

The general solution is

$$x_n = \alpha(-1)^n + \beta n(-1)^{n-1} + \gamma(1/4)^n.$$

Stable difference equations

- **Definition:** *An element* $x = [x_1, x_2, x_3, \cdots] \in V$ *is bounded if* $\exists\, c > 0$ *such that* $|x_n| \le c,\ \forall\, n \ge 1$, *i.e.,* $\sup_{n\ge 1} |x_n| < \infty$.
- **Definition:** *A difference equation of the form* $p(E)x = 0$ *is said to be stable if all of its solution is bounded.*

  Example: $x_{n+2} - 3x_{n+1} + 2x_n = 0,\ n \ge 1.$

  The general solution is $x_n = \alpha \cdot 1 + \beta 2^n$. Since 2^n is not bounded, so the difference equation is unstable.
- **Theorem on stable difference equations:** *For any polynomial* p *satisfying* $p(0) \ne 0$, *the following are equivalent:*
  1. *The difference equation* $p(E)x = 0$ *is stable.*
  2. *All zeros of* p *satisfy* $|z| \le 1$ *and all multiple zeros satisfy* $|z| < 1$.

Linear multistep methods

- Recall the IVP:

$$\begin{cases} x'(t) & = & f(t, x(t)), \\ x(t_0) & = & x_0. \end{cases}$$

The LMM can be written as

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h\{b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}\},$$

where $a_k \neq 0$, $f_i = f(t_i, x_i)$, and $t_i = t_0 + ih$.

- We assume $x_0, x_1, \cdots, x_{k-1}$ have been obtained by some other method (e.g., RK4).
  - (1) if $b_k \neq 0$ then the method is implicit. e.g., A-M formula of order 5 (4-step method):
    $x_n - x_{n-1} =$
    $h\{\frac{251}{720} f_n + \frac{646}{720} f_{n-1} - \frac{264}{720} f_{n-2} + \frac{106}{720} f_{n-3} - \frac{19}{720} f_{n-4}\}.$
  - (2) if $b_k = 0$ then the method is explicit. e.g., A-B formula of order 5 (5-step method):
    $x_n - x_{n-1} =$
    $h\{\frac{1901}{720} f_{n-1} - \frac{2774}{720} f_{n-2} + \frac{2616}{720} f_{n-3} - \frac{1274}{720} f_{n-4} + \frac{251}{720} f_{n-5}\}.$

Convergence

- **Definition:** *The LMM is said to be convergent if*

$$\lim_{h\to 0} x(h,t) = x(t), \quad (t \text{ fixed}) \quad (\star)$$

where $x(h,t)$ is the approximate solution using the step size h and $x(t)$ is exact solution, $\forall t \in [t_0, t_m]$, provided that starting values obey the same equation, that is,

$$\lim_{h\to 0} x(h, t_0 + nh) = x_0 \quad (0 \le n < k) \quad (\star\star)$$

and f satisfies the hypotheses of the existence-uniqueness theorem: f is continuous in the strip $t_0 \le t \le t_m$, $-\infty < x < \infty$ and satisfies a Lipschitz condition in the second variable.

Stability and consistency

- Consider the following polynomials associated with the LMM:

$$\begin{aligned} p(z) &= a_k z^k + a_{k-1} z^{k-1} + \cdots + a_0, \\ q(z) &= b_k z^k + b_{k-1} z^{k-1} + \cdots + b_0. \end{aligned}$$

  It can be shown that certain desirable properties of the LMM depend on the location of the roots of the polynomials p and q.

- **Definition:** *The LMM is stable if all the roots of p lie in the disk $|z| \le 1$ and if each root of modulus 1 is simple.*
- **Definition:** *The LMM is consistent if $p(1) = 0$ and $p'(1) = q(1)$.*

Main theorem of the LMM

For the LMM to be convergent, it is necessary and sufficient that it be stable and consistent.

Proof: (stability is necessary)

- Suppose that the method is not stable. Then either p has a root λ satisfying $|\lambda| > 1$ or p has a root λ satisfying $|\lambda| = 1$ and $p'(\lambda) = 0$.
- In either case we consider a simple IVP whose solution is $x(t) = 0$:

$$\begin{cases} x'(t) = 0, \\ x(0) = 0. \end{cases}$$

In this case, the LMM becomes

$$a_k x_n + a_{k-1} x_{n-1} + \cdots a_0 x_{n-k} = 0. \quad (\star\star\star)$$

This is a linear difference equation. One solution is $x_n = h\lambda^n$.

Proof: stability is necessary (cont'd)

- Assume that $|\lambda| > 1$ implies for $0 \le n < k$

$$|x(h, nh)| = h|\lambda^n| < h|\lambda|^k \to 0 \quad \text{as } h \to 0.$$

  Thus the condition $(\star\star)$ is verified.

- However, if $t = nh$ then $h = tn^{-1}$ and

$$|x(h, t) = |x(h, nh)| = tn^{-1}|\lambda|^n \to \infty \quad \text{as } h \to 0,$$

  since $n \to \infty$ as $h \to 0$ and $|\lambda| > 1$. Thus, $(\star)$ is violated.

- Now assume $|\lambda| = 1$ and $p'(\lambda) = 0$, i.e., λ is a multiple roots, then a solution of $(\star\star\star)$ is $x_n = hn\lambda^{n-1}$. Again $(\star\star)$ is satisfied, since for $0 \le n < k$ we have

$$|x(h, nh)| = hn|\lambda|^{n-1} = hn < hk \to 0 \quad \text{as } h \to 0.$$

- However, the condition $(\star)$ is violated because

$$|x(h, t)| = (tn^{-1})n|\lambda|^{n-1} = t \ne 0$$

  and does not go to zero as $h \to 0$.

Therefore, if the LMM is convergent then it is stable.

Proof: consistency is necessary

- Suppose that the method is convergent. Consider a simple IVP problem whose solution is $x(t) = 1$.

$$\begin{cases} x'(t) &= 0, \\ x(0) &= 1. \end{cases}$$

- For this IVP, the LMM becomes $(\star\star\star)$ again. One solution is obtained by setting $x_0 = x_1 = \cdots = x_{k-1} = 1$ and then use $(\star\star\star)$ to generate the remaining values, $x_k, x_{k+1}, \cdots$.
- Since the method is convergent, $\lim_{n\to\infty} x_n = 1$. Substituting this into $(\star\star\star)$ implies

$$a_k + a_{k-1} + \cdots + a_0 = 0 \quad \text{or} \quad p(1) = 0.$$

Proof: consistency is necessary (cont'd)

- Now consider a simple IVP problem whose solution is $x(t) = t$:

$$\begin{cases} x'(t) & = & 1, \\ x(0) & = & 0. \end{cases}$$

For this IVP, the LMM becomes

$$a_k x_n + a_{k-1} x_{k-1} + \cdots + a_0 x_{n-k} = h\{b_k + b_{k-1} + \cdots + b_0\}. \quad (\star\star\star)$$

- Since the method is convergent, it is stable by the preceding proof which implies $p(1) = 0$ and $p'(1) \neq 0$, i.e., no multiple roots of size 1.

Proof: consistency is necessary (cont'd)

- Let us verify that $x_n = (n+k)h\gamma$ with $\gamma := q(1)/p'(1)$ is a solution of $(\star\star\star)$:

$$
\begin{aligned}
& h\gamma\{a_k(n+k) + a_{k-1}(n+k-1) + \cdots + a_0 n\} \\
&= nh\gamma \underbrace{(a_k + a_{k-1} + \cdots + a_0)}_{p(1)=0} + h\gamma \underbrace{(ka_k + (k-1)a_{k-1} + \cdots + a_1)}_{p'(1)\neq 0} \\
&= h\gamma p'(1) = h\frac{q(1)}{p'(1)}p'(1) = h\{b_k + b_{k-1} + \cdots + b_0\}.
\end{aligned}
$$

- Notice that the starting values in this numerical solution are consistent with the initial value $x(0) = 0 = x_0$ because $\lim_{h\to 0}(n+k)h\gamma = 0 = x_0$ for $n = 0, 1, \cdots, k-1$. That is, $(\star\star)$ holds.
- The convergence condition demands that $\lim_{n\to\infty} x_n = t$ if $nh = t$. Hence we have $\lim_{n\to\infty}(n+k)h\gamma = t$. We can conclude $\gamma = 1$ or $p'(1) = q(1)$ because $\lim_{n\to\infty} kh = 0$.

Example

Consider the Milne method

$$x_n - x_{n-2} = h\left(\frac{1}{3}f_n + \frac{4}{3}f_{n-1} + \frac{1}{3}f_{n-2}\right).$$

- $p(z) = z^2 - 1 = 0 \Rightarrow z = \pm 1$: simple root. Hence, the method is stable.
- $p'(z) = 2z$ and $q(z) = \frac{1}{3}z^2 + \frac{4}{3}z + \frac{1}{3}$. Then $p'(1) = 2 = q(1)$ and $p(1) = 0$. Hence, the method is consistent.

Therefore we can conclude that the method is convergent.

Local truncation error

- Assume that all previous steps of the LMM are computed correctly, i.e., $x_i = x(t_i)$ for $n-k \le i \le n-1$. Here $x(t)$ denotes the exact solution of the IVP. We now want to to compute x_n.

  Definition: *The local truncation error is defined as $x(t_n) - x_n$. Note that the round-off error is not included.*

- **Theorem:** *If the LMM is of order m, and if $x \in C^{m+2}$ and $\frac{\partial f}{\partial x}$ is continuous, then under the assumption above we have*

$$x(t_n) - x_n = \left(\frac{d_{m+1}}{a_k}\right) h^{m+1} x^{(m+1)}(t_{n-k}) + O(h^{m+2}).$$

  The coefficient d_k are defined in Section 8.4, p. 553.

  Proof: see page 561.

  The theorem states that if the method has order m, then the local truncation error will be $O(h^{m+1})$.

Global truncation error

- The question is how do local truncation errors propagate during the solution process. Consider the IVP

$$\begin{cases} x'(t) &= f(t, x(t)), \\ x(0) &= s. \end{cases}$$

Assume that $f_x(t, x)$ is continuous and $f_x(t, x) \le \lambda$ in $[0, T] \times \mathbb{R}$.

- To see how the solution is affected by a change in the initial value s, first write the solution of the IVP as $x(t; s)$. Assume that $x(t; s)$ is smooth. Then define $u(t) := \frac{\partial x(t;s)}{\partial s}$.
- Differentiate the IVP with respect to s, we obtain the variational equation:

$$\begin{cases} u'(t) &= f_x(t, x)u, \\ u(0) &= 1. \end{cases}$$

Solving for u, we see how a change in s can affect the solution to the IVP.

Example

Find u for the following IVP:

$$\begin{cases} x'(t) &= x^2, \\ x(0) &= s. \end{cases}$$

Solution:

Here $f(t,x) = x^2 \Rightarrow f_x = 2x$. The variational equation is:

$$\begin{cases} u'(t) &= 2xu, \\ u(0) &= 1. \end{cases}$$

Since the solution to the first IVP is $x(t) = s(1-st)^{-1}$, we then have

$$u'(t) = 2s(1-st)^{-1}u(t) \Rightarrow u(t) = (1-st)^{-2}.$$

Theorem on variational equation

If $f_x \le \lambda$, the solution to the variational equation satisfies $|u(t)| \le e^{\lambda t}$ for $t \ge 0$.

Proof: Recall the variational equation

$$\begin{cases} u'(t) &= f_x(t,x)u, \\ u(0) &= 1. \end{cases}$$

From the variational equation,

$$u'/u = f_x = \lambda - \alpha(t),$$

where $\alpha(t) \ge 0$. Integrating

$$\ln(|u|) = \lambda t - \int_0^t \alpha(\tau) d\tau = \lambda t - A(t).$$

Since $t \ge 0 \Rightarrow A \ge 0 \Rightarrow \ln(|u|) \le \lambda t \Rightarrow |u| \le e^{\lambda t}$. $\square$

Theorem on solution curves

Assume that $f_x \le \lambda$. If the IVP

$$\begin{cases} x'(t) &= f(t,x), \\ x(0) &= s \end{cases}$$

is solved with initial values s and $s+\delta$, then the solution curves at t differ by at most $|\delta| e^{\lambda t}$.

Proof: By the MVT, the definition of u, and the above Theorem, we have

$$\begin{aligned} |x(t;s) - x(t;s+\delta)| &= \left| \frac{\partial}{\partial s} x(t; s+\theta\delta) \right| |\delta| \\ &= |u(t)||\delta| \le |\delta| e^{\lambda t}. \end{aligned}$$

□

Theorem on global truncation error bound

If the local truncation errors at $t_1, t_2, \cdots, t_n$ do not exceed δ in magnitude, then the global truncation error at t_n does not exceed

$$\frac{\delta(e^{n\lambda h} - 1)}{(e^{\lambda h} - 1)}.$$

Proof: Let truncation errors of $\delta_1, \delta_2, \cdots$ be associated with numerical solution at $t_1, t_2, \cdots$. In computing x_2 there was an error of δ_1 in the initial condition, by above Theorem, the effect at t_2 is at most $|\delta_1|e^{\lambda h}$. Thus, the global truncation error at t_2 is at most

$$|\delta_1|e^{\lambda h} + |\delta_2|.$$

The effect of this error at t_3 is no greater than

$$(|\delta_1|e^{\lambda h} + |\delta_2|)e^{\lambda h}.$$

The global truncation error at t_3 is at most

$$(|\delta_1|e^{\lambda h} + |\delta_2|)e^{\lambda h} + |\delta_3|.$$

Theorem on global truncation error bound (cont'd)

Continuing in this way, we find that the global truncation error at t_n is no greater than

$$\begin{aligned}
\sum_{k=1}^{n} |\delta_k| e^{(n-k)\lambda h} &\leq \delta \sum_{k=1}^{n} e^{(n-k)\lambda h} \\
&= \delta \sum_{k=0}^{n-1} e^{(n-k-1)\lambda h} \\
&= \delta e^{(n-1)\lambda h} \sum_{k=0}^{n-1} e^{-k\lambda h} \\
&= \delta e^{(n-1)\lambda h} \left(\frac{1 - e^{-n\lambda h}}{1 - e^{-\lambda h}} \right) \\
&= \delta \frac{e^{n\lambda h} - 1}{e^{\lambda h} - 1}.
\end{aligned}$$

□

Theorem on global truncation error approximation

If the local truncation errors in the numerical solution are $O(h^{m+1})$, then the global truncation error is $O(h^m)$.

Proof: By the above Theorem, set $\delta = O(h^{m+1})$. Then

$$\begin{aligned}
\text{GTE} &\leq O(h^{m+1})\left(\frac{e^{nz}-1}{e^z-1}\right) \qquad (z := \lambda h) \\
&\approx O(h^{m+1})\frac{nz}{z} \qquad (e^z = 1 + z + \frac{1}{2!}z^2 + \cdots) \\
&= O(h^{m+1})\frac{t}{h} \qquad (nh = t) \\
&= O(h^m)t.
\end{aligned}$$

Stiff equations: introduction

- Euler's method for the IVP

$$\begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0, \end{cases}$$

is given by

$$x_{n+1} = x_n + hf(t_n, x_n) \quad n \geq 0.$$

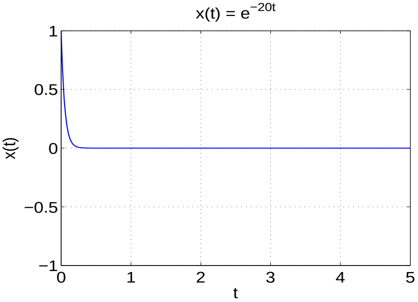
- Consider the results of Euler's method on the simple test problem: $x'(t) = \lambda x$ and $x(0) = 1$. The exact solution is $x(t) = e^{\lambda t}$.

Solution: Euler's method produces the numerical solution:

$$\begin{aligned} x_0 &= 1, \\ x_{n+1} &= x_n + h\lambda x_n \\ &= (1 + h\lambda)x_n \\ &= \cdots = (1 + h\lambda)^{n+1} x_0 \\ \Longrightarrow x_n &= (1 + h\lambda)^n. \end{aligned}$$

Stiff equations (cont'd)

- For $\lambda < 0$, the exact solution is exponentially decaying. The numerical solution will tend to 0 if and only if $|1 + h\lambda| < 1 \iff -1 < 1 + h\lambda < 1 \iff h < -2/\lambda$.
- For example, if $\lambda = -20$, we have to take $h < 0.1$. Thus, the numerical solution must proceed with small steps in a region where the nature of the exact solution indicates that large steps may be taken.



Implicit Euler's method

- Implicit Euler's method for the IVP

$$\begin{cases} x'(t) &= f(t,x), \\ x(t_0) &= x_0, \end{cases}$$

is given by

$$x_{n+1} = x_n + hf(t_{n+1}, x_{n+1}) \quad n \geq 0.$$

- Consider the results of implicit Euler's method on the problem: $x'(t) = \lambda x$ and $x(0) = 1$. The exact solution is $x(t) = e^{\lambda t}$.

  Solution: Implicit Euler's method produces

$$\begin{aligned} x_0 &= 1, \\ x_{n+1} &= x_n + h\lambda x_{n+1}. \\ x_{n+1} &= (1 - h\lambda)^{-1} x_n. \\ x_n &= (1 - h\lambda)^{-n}. \end{aligned}$$

  For $\lambda < 0$, we have $1 - h\lambda > 1$ and then $|1 - h\lambda|^{-1} < 1 \; \forall \, h > 0$.

- Explicit Euler's method is cheap but conditionally stable. Implicit Euler's method is expensive but unconditionally stable.

General linear multistep methods

- The LMM has the form:

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h\{b_k f_n + b_{k-1} f_{n-1} + \cdots + b_0 f_{n-k}\}.$$

- When this is applied to the test problem: $x'(t) = \lambda x$ and $x(0) = 1$, we obtain

$$a_k x_n + a_{k-1} x_{n-1} + \cdots + a_0 x_{n-k} = h\lambda\{b_k x_n + b_{k-1} x_{n-1} + \cdots + b_0 x_{n-k}\}.$$

- Thus, our numerical solution will solve the homogeneous linear difference equation:

$$(a_k - h\lambda b_k) x_n + (a_{k-1} - h\lambda b_{k-1}) x_{n-1} + \cdots + (a_0 - h\lambda b_0) x_{n-k} = 0.$$

General linear multistep methods (cont'd)

- The solutions of the homogeneous linear difference equation are determined by the roots of the characteristic polynomial:

$$\varphi(z) := (a_k - h\lambda b_k)z^k + (a_{k-1} - h\lambda b_{k-1})z^{k-1} + \cdots + (a_0 - h\lambda b_0).$$

e.g., If r is a zero of $\varphi(z)$, then $x_n = r^n$ is a solution of the linear difference equation.

- Note that

$$\varphi(z) = p(z) - h\lambda q(z),$$

where

$$\begin{aligned} p(z) &= a_k z^k + a_{k-1} z^{k-1} + \cdots + a_1 z + a_0, \\ q(z) &= b_k z^k + b_{k-1} z^{k-1} + \cdots + b_1 z + b_0. \end{aligned}$$

A-stability

- If $\lambda < 0$, then the solution $x(t) = e^{\lambda t}$ of the test problem is exponentially decaying. It is necessary that all roots of the polynomial $\varphi = p(z) - h\lambda q(z)$ lie in the disk $|z| < 1$. If $\lambda = \mu + i\nu$ is complex,

$$x(t) = e^{\lambda t} = e^{\mu t} e^{i\nu t} = e^{\mu t}(\cos \nu t + i \sin \nu t).$$

  In this case, exponential decay means $\mu < 0$.

- **Definition:** *We say the LMM is A-stable if the roots of ϕ to be interior to the unit disk whenever $h > 0$ and $Re(\lambda) < 0$.*

- **Definition:** *The region of absolute stability of the LMM is the set of complex numbers ω such that the roots of $p - \omega q$ lie in the interior of the unit disk.*

- An LMM is A-stable if and only if its region of absolute stability contains the left half-plane.

Examples

- By definition, the implicit Euler method is A-stable. Another example is the implicit trapezoid method defined by

$$x_n - x_{n-1} = \frac{1}{2}h\{f_n + f_{n-1}\},$$

then $\phi(z) = z - 1 - \lambda h\{\frac{1}{2}z + \frac{1}{2}\}$.

Root: $z(1 - \frac{\lambda h}{2}) = 1 + \frac{\lambda h}{2} \Rightarrow z = \frac{2 + \lambda h}{2 - \lambda h}$.

When $h > 0$ and $Re(\lambda) < 0$, we have $|z| < 1 \Rightarrow$ A-stable.

- What about the explicit Euler method? Here

$$x_n - x_{n-1} = hf_{n-1}.$$

$p(z) = z - 1$ and $q(z) = 1$.
$\phi(z) = z - 1 - \lambda h = 0 \Rightarrow z = 1 + \lambda h \Rightarrow |1 + \omega| < 1$, a disk of radius 1 centered at -1. It is not A-stable.

Remarks

- WARNING: If you are not using an A-stable method, you have to make sure that λh lies in the region of absolute stability for the method.
- An important theorem, due to Dahlquist [1963], states that an A-stable LMM must be an implicit method, and its order cannot exceed 2. This result places a severe restriction on A-stable methods.
- The implicit trapezoid rule is often used on stiff equations because it has the least truncation error among all A-stable linear multistep methods.

Homework

Consider the LMM

$$x_{n+1} = x_{n-1} + 2hf_n$$

to approximate the IVP: $x'(t) = f(t,x)$ and $x(t_0) = x_0$.

Is the method

- stable?
- consistent?
- convergent?
- A-stable?

A system of first-order differential equations

The standard form for a system of first-order ODEs is given by

$$\begin{cases} x_1'(t) &= f_1(t, x_1, x_2, \cdots, x_n), \\ x_2'(t) &= f_2(t, x_1, x_2, \cdots, x_n), \\ &\vdots \\ x_n'(t) &= f_n(t, x_1, x_2, \cdots, x_n). \end{cases} \qquad (\star)$$

There are n unknown functions, $x_1, x_2, \cdots, x_n$ to be determined. Here $x_i'(t) := \frac{dx_i(t)}{dt}$.

Example

Consider the system of first-order differential equations:

$$\begin{cases} x'(t) &= x + 4y - e^t, \\ y'(t) &= x + y + 2e^t. \end{cases}$$

The general solution:

$$\begin{cases} x(t) &= 2ae^{3t} - 2be^{-t} - 2e^t, \\ y(t) &= ae^{3t} + be^{-t} + 1/4e^t, \end{cases}$$

where $a, b \in \mathbb{R}$. If the system of differential equations with the initial conditions, e.g., $x(0) = 4$ and $y(0) = 5/4$, then the solution is unique, and

$$\begin{cases} x(t) &= 4e^{3t} + 2e^{-t} - 2e^t, \\ y(t) &= 2e^{3t} - e^{-t} + 1/4e^t. \end{cases}$$

Vector notation and higher-order ODEs

- **Vector notation:** let $X := [x_1, x_2, \cdots, x_n]^\top$ and $F := [f_1, f_2, \cdots, f_n]^\top$, where $X \in \mathbb{R}^n$ and $F : \mathbb{R}^{n+1} \to \mathbb{R}^n$.

  Then an IVP associated with the system of ODEs $(\star)$ is given by

$$\begin{cases} X'(t) &= F(t, X(t)), \\ X(t_0) &= X_0 \in \mathbb{R}^n. \end{cases}$$

- A higher-order ODE can be converted to a first-order system.

  Consider $y^{(n)}(t) = f(t, y, y', \cdots, y^{(n-1)})$ and introduce $x_1 = y, x_2 = y', \cdots, x_n = y^{(n-1)}$. Then we have

$$\begin{cases} x_1'(t) &= x_2, \\ x_2'(t) &= x_3, \\ &\vdots \\ x_{n-1}'(t) &= x_n, \\ x_n'(t) &= f(t, x_1, x_2, \cdots, x_n). \end{cases}$$

Example

Convert the higher-order IVP

$$(\sin t)y''' + \cos(ty) + \sin(y'' + t^2) + (y')^3 = \log t$$

with $y(2) = 7, y'(2) = 3, y''(2) = -4$ to a system of 1st-order equations with initial values.

Solution: Let $x_1(t) = y(t), x_2(t) = y'(t), x_3(t) = y''(t)$. Then,

$$\begin{cases} x_1'(t) &= x_2, \\ x_2'(t) &= x_3, \\ x_3'(t) &= \{\log t - x_2^3 - \sin(t^2 + x_3) - \cos(tx_1)\}/\sin t, \end{cases}$$

with $x_1(2) = 7, x_2(2) = 3, x_3(2) = -4$.

In-class exercise

Convert the system

$$\begin{cases} (x'')^2 + te^y + y' & = & x' - x, \\ y'y'' - \cos(xy) + \sin(tx'y) & = & x \end{cases}$$

to a system of 1st-order equations.

Taylor-series method for systems

For each variable, use the Taylor-series method

$$x_i(t+h) \approx x_i(t) + hx_i'(t) + \frac{h^2}{2!}x_i''(t) + \frac{h^3}{3!}x_i'''(t) + \cdots + \frac{h^n}{n!}x_i^{(n)}(t),$$

or in the vector form

$$X(t+h) \approx X(t) + hX'(t) + \frac{h^2}{2!}X''(t) + \frac{h^3}{3!}X'''(t) + \cdots + \frac{h^n}{n!}X^{(n)}(t).$$

Homework

Write the Taylor-series codes of order 3 for the following IVP using $h = -0.1$ and plot the solution $-2 \le t \le 1$:

$$\begin{cases} x'(t) & = & x + y^2 - t^3, \\ y'(t) & = & y + x^3 + \cos t \end{cases}$$

with $x(1) = 3$ and $y(1) = 1$.

Autonomous systems

- From the theoretical standpoint, there is no loss of generality in assuming that the equations in system $(\star)$ do not contain t explicitly. We can take $x_0(t) = t$, $x_0'(t) = 1$. Then $x_i' = f_i(x_0, x_1, \cdots, x_n)$, $i = 0, 1, \cdots, n$, or $X'(t) = F(X)$, where $X(t) = (x_0(t), x_1(t), \cdots, x_n(t))^\top$.
- **Example:** convert the following IVP to an autonomous system

$$(\sin t)y''' + \cos(ty) + \sin(y'' + t^2) + (y')^3 = \log t,$$

with $y(2) = 7, y'(2) = 3, y''(2) = -4$.

Solution: Let $x_0(t) = t$. Then $x_0'(t) = 1$. Let $x_1'(t) = x_2$ and $x_2'(t) = x_3$. Then we have

$$\begin{cases} x_0'(t) &= 1, \\ x_1'(t) &= x_2, \\ x_2'(t) &= x_3, \\ x_3'(t) &= \{\log x_0 - x_2^3 - \sin(x_0^2 + x_3) - \cos(x_0 x_1)\} / \sin x_0, \end{cases}$$

with the initial condition $X(2) = (2, 7, 3, -4)^\top$.

RK4 method for $X'(t) = F(X)$

- For an autonomous system of equations, $X'(t) = F(X)$, we have 4th-order Runge-Kutta method:

$$
\begin{aligned}
X(t+h) &\approx X(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4), \\
F_1 &= hF(X), \quad F_2 = hF(X + 1/2F_1), \\
F_3 &= hF(X + 1/2F_2), \quad F_4 = hF(X + F_3).
\end{aligned}
$$

In other words, the 4th order RK is defined as

$$
\begin{aligned}
X_{k+1} &= X_k + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4), \quad k \geq 0, \\
F_1 &:= hF(X_k), \quad F_2 := hF(X_k + 1/2F_1), \\
F_3 &:= hF(X_k + 1/2F_2), \quad F_4 := hF(X_k + F_3), \\
X_k &:= [x_{1k}, x_{2k}, \cdots, x_{nk}]^\top, x_{ik} \approx x_i(t_0 + kh) \text{ for } 1 \leq i \leq n.
\end{aligned}
$$

- Other methods, they are all similar to the single equation case.

Boundary-value problems

- For an IVP, the auxiliary conditions are prescribed at the same point, $t = a$, e.g.,

$$\begin{cases} x''(t) &= f(t, x, x'), \\ x(a) &= \alpha, \\ x'(a) &= \beta. \end{cases}$$

- For a boundary-value problem (BVP), the auxiliary conditions are prescribed at the different points, $t = a$ and $t = b$, e.g.,

$$\begin{cases} x''(t) &= f(t, x, x'), \\ x(a) &= \alpha, \\ x(b) &= \beta. \end{cases}$$

  This particular example is a so-called two-point BVP.

Existence of solutions

- Assume that f is nice function. It is not enough for existence of a solution. Consider the BVP:

$$\begin{cases} x''(t) &= -x, \\ x(0) &= 3, \\ x(\pi) &= 7. \end{cases}$$

- The general solution is (recall from ODE course)

$$x(t) = A \sin t + B \cos t.$$

- Using the boundary conditions, we have

$$\begin{aligned} x(0) = 3 &\quad\Rightarrow\quad B = 3, \\ x(\pi) = 7 &\quad\Rightarrow\quad B = -7. \end{aligned}$$

No solution!

Existence of solutions (cont'd)

- Note that we could also have infinite number of solutions. Consider the BVP:

$$\begin{cases} x''(t) &= -x, \\ x(0) &= 0, \\ x(\pi) &= 0. \end{cases}$$

- The general solution is given by

$$x(t) = A\sin t + B\cos t.$$

- Using the boundary conditions,

$$\begin{aligned} x(0) = 0 \quad &\Rightarrow \quad B = 0, \\ x(\pi) = 0 \quad &\Rightarrow \quad B = 0. \end{aligned}$$

  We have

$$x(t) = A\sin t, \quad \forall\, A \in \mathbb{R}.$$

Existence and uniqueness theorem (Keller, 1968)

The BVP

$$\begin{cases} x''(t) &= f(t,x), \\ x(0) &= 0, \\ x(1) &= 0 \end{cases}$$

has a unique solution if $\frac{\partial f}{\partial x}$ *is continuous, nonnegative, and bounded in the strip* $0 \le t \le 1$ *and* $-\infty < x < \infty$.

Note: Existence and uniqueness theorems for solutions of the two-point BVP are more complicated than the IVP.

Example

Use the previous theorem to show the following BVP has a unique solution

$$\begin{cases} x''(t) &= (5x + \sin 3x)e^t, \\ x(0) &= x(1) = 0. \end{cases}$$

Solution: We have

$$2 \leq \frac{\partial f}{\partial x} = (5 + 3\cos 3x)e^t \leq 8e$$

for $0 \leq t \leq 1$, $-\infty < x < \infty$, and it is a continuous function, nonnegative since $3\cos 3x \geq -3$.

$\Longrightarrow$ all assumptions of above theorem are satisfied.

$\Longrightarrow$ the BVP has a unique solution.

Theorem for more general BVPs

In order to use the above theorem for more general BVPs, we can use change of variable, e.g., if we have to solve

$$\begin{cases} x''(t) &= f(t,x), \\ x(a) &= \alpha, \\ x(b) &= \beta, \end{cases}$$

then consider $t := a + (b-a)s := a + \lambda s$, i.e., $s := \frac{t-a}{b-a}$. Define

$$\begin{aligned} y(s) &:= x(a+\lambda s), \\ y'(s) &= \lambda x'(a+\lambda s), \\ y''(s) &= \lambda^2 x''(a+\lambda s) = \lambda^2 f(a+\lambda s, y(s)). \end{aligned}$$

BCs: $y(0) = x(a) = \alpha$ and $y(1) = x(b) = \beta$.

First theorem on two-point BVPs

Consider these two-point BVPs:

$$\begin{cases} x''(t) &= f(t,x), \\ x(a) &= \alpha, \\ x(b) &= \beta; \end{cases} \qquad (\star)$$

$$\begin{cases} y''(s) &= \lambda^2 f(a+\lambda s, y(s)) := g(s, y(s)), \\ y(0) &= \alpha, \\ y(1) &= \beta. \end{cases} \qquad (\star\star)$$

- *If* $x(t)$ *is a solution of* $(\star)$ *then* $y(s) = x(a+(b-a)s)$ *is a solution of* $(\star\star)$.
- *If* $y(s)$ *is a solution of* $(\star\star)$ *then* $x(t) = y((t-a)/(b-a))$ *is a solution of* $(\star)$.

Second theorem on two-point BVPs

Consider these two-point BVPs:

$$\begin{cases} y''(t) &= g(t,y), \\ y(0) &= \alpha, \\ y(1) &= \beta; \end{cases} \qquad (\star\star)$$

$$\begin{cases} z''(t) &= h(t,z), \\ z(0) &= 0, \\ z(1) &= 0, \end{cases} \qquad (\star\star\star)$$

where $h(t,z) = g(t, z + \alpha + (\beta - \alpha)t)$.

- *If* z *solves* $(\star\star\star)$ *then* $y(t) = z(t) + \alpha + (\beta - \alpha)t$ *solves* $(\star\star)$.
- *If* y *solves* $(\star\star)$ *then* $z(t) = y(t) - \{\alpha + (\beta - \alpha)t\}$ *solves* $(\star\star\star)$.

Example

Convert the following two-point BVP to an equivalent one with 0 boundary values on $[0, 1]$:

$$\begin{cases} x''(t) = x^2 + 3 - t^2 - xt, \\ x(3) = 7, \quad x(5) = 9. \end{cases}$$

Solution: By the first theorem, we have

$$\begin{cases} y''(t) = g(t, y), \\ y(0) = 7, \quad y(1) = 9, \end{cases}$$

$g(t, y) = (5-3)^2 f(3+2t, y) = 4\{y^2 + 3 - (3+2t)^2 - y(3+2t)\}$. By the second theorem, we get

$$\begin{cases} z''(t) = h(t, z), \\ z(0) = 0, \quad z(1) = 0, \end{cases}$$

$$\begin{aligned} h(t, z) &= g(t, z + 7 + 2t) \\ &= 4\{(z+7+2t)^2 + 3 - (3+2t)^2 + (z+7+2t)(3+2t)\}. \end{aligned}$$

Finite-difference methods: linear case

- Consider the linear BVP

$$\begin{cases} x''(t) &= u(t) + v(t)x + w(t)x', \\ x(a) &= \alpha, \\ x(b) &= \beta. \end{cases}$$

- Recall that

$$\begin{aligned} x'(t) &= \frac{1}{2h}\Big(x(t+h) - x(t-h)\Big) - \frac{h^2}{6}x'''(\xi), \\ x''(t) &= \frac{1}{h^2}\Big(x(t+h) - 2x(t) + x(t-h)\Big) - \frac{h^2}{12}x^{(4)}(\xi). \end{aligned}$$

- Let $t_i = a + ih$, where $0 \le i \le n+1$, and $h = (b-a)/(n+1)$.
- Set $u_i = u(t_i)$, $v_i = v(t_i)$, $w_i = w(t_i)$ and use $y_i \approx x(t_i)$.

Finite-difference methods: linear case (cont'd)

- Then the differential equation is approximated by

$$\left(\frac{y_{i+1}-2y_i+y_{i-1}}{h^2}\right)=u_i+v_iy_i+w_i\left(\frac{y_{i+1}-y_{i-1}}{2h}\right).$$

- Multiply by $-h^2$ and rearrange to obtain

$$\begin{cases} (-1-\frac{1}{2}hw_i)y_{i-1}+(2+h^2v_i)y_i+(-1+\frac{1}{2}hw_i)y_{i+1} & = & -h^2u_i, \\ i=1,2,\cdots n, & & \\ y_0 & = & \alpha, \\ y_{n+1} & = & \beta. \end{cases}$$

- Let

$$\begin{aligned} a_i &= -1-\frac{1}{2}hw_{i+1}, \quad 0\le i\le n-1, \\ d_i &= 2+h^2v_i, \quad 1\le i\le n, \\ c_i &= -1+\frac{1}{2}hw_i, \quad 1\le i\le n, \\ b_i &= -h^2u_i, \quad 1\le i\le n. \end{aligned}$$

A system of linear equations

We obtain

$$\begin{bmatrix} d_1 & c_1 & & & & \\ a_1 & d_2 & c_2 & & & \\ & a_2 & d_3 & c_3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & a_{n-2} & d_{n-1} & c_{n-1} \\ & & & & a_{n-1} & d_n \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} b_1 - a_0\alpha \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n - c_n\beta \end{bmatrix}.$$

- This is a tridiagonal system, and can be solved by a special Gaussian algorithm. Also the matrix is strictly diagonally dominant if $v_i > 0$ and h is small enough so that $|\frac{1}{2}hw_i| < 1$, which implies that Gaussian elimination algorithm does not require pivoting.
- The following equality will be needed later:

$$|d_i| - |c_i| - |a_{i-1}| = 2 + h^2 v_i - (1 - \frac{1}{2}hw_i) - (1 + \frac{1}{2}hw_i) = h^2 v_i > 0.$$

Existence-uniqueness theorem (Keller, 1968)

The BVP

$$\begin{cases} x''(t) & = & f(t,x,x'), \\ c_{11}x(a) + c_{12}x'(a) & = & c_{13}, \\ c_{21}x(b) + c_{22}x'(b) & = & c_{23} \end{cases}$$

has a unique solution on the interval $[a,b]$ *provided that*

- f *and its first partial derivatives* f_t, f_x *and* $f_{x'}$ *are continuous on* $D = [a,b] \times \mathbb{R} \times \mathbb{R}$;
- $f_x > 0$, $|f_x| \leq M$ *and* $|f_{x'}| \leq M$ *on* D;
- $|c_{11}| + |c_{12}| > 0$, $|c_{21}| + |c_{22}| > 0$, $|c_{11}| + |c_{21}| > 0$ *and* $c_{11}c_{12} \leq 0 \leq c_{21}c_{22}$.

Convergence analysis

- Let us go back to the linear BVP:

$$\begin{cases} x''(t) &= u(t) + v(t)x + w(t)x', \\ x(a) &= \alpha, \\ x(b) &= \beta. \end{cases}$$

  Assume that $u, v, w \in C^1[a, b]$ and $v > 0$. Then the BVP has a unique solution.

- We wish to estimate $|x(t_i) - y_i|$ as $h \to 0$, where $x(t_i)$ is the exact solution at t_i and y_i is the corresponding discrete solution, which depends on h.

Convergence analysis (cont'd)

- The exact solution $x(t)$ satisfies the following system:

$$\left(\frac{x(t_{i-1}) - 2x(t_i) + x(t_{i+1})}{h^2}\right) - \frac{1}{12}h^2 x^{(4)}(\tau_i)$$
$$= u_i + v_i x(t_i) + w_i\left(\frac{x(t_{i+1}) - x(t_{i-1})}{2h}\right) - \frac{1}{6}h^2 x^{(3)}(\eta_i).$$

- The discrete solution y_i satisfies the following system:

$$\left(\frac{y_{i-1} - 2y_i + y_{i+1}}{h^2}\right) = u_i + v_i y_i + w_i\left(\frac{y_{i+1} - y_{i-1}}{2h}\right).$$

- Subtracting above system from the first and writing $e_i = x(t_i) - y_i$, we obtain

$$\left(\frac{e_{i-1} - 2e_i + e_{i+1}}{h^2}\right) = v_i e_i + w_i\left(\frac{e_{i+1} - e_{i-1}}{2h}\right) + h^2 g_i,$$

where $g_i := \frac{1}{12}x^{(4)}(\tau_i) - \frac{1}{6}x^{(3)}(\eta_i)$.

Convergence analysis (cont'd)

- After multiplying by $-h^2$ and collecting terms, we have

$$(-1 - \frac{1}{2}hw_i)e_{i-1} + (2 + h^2 v_i)e_i + (-1 + \frac{1}{2}hw_i)e_{i+1} = -h^4 g_i.$$

- This is identical to the matrix problem we have for the discrete problem. Using the coefficients introduced earlier, we write this in the form

$$a_{i-1}e_{i-1} + d_i e_i + c_i e_{i+1} = -h^4 g_i. \qquad (\star)$$

- Let $\lambda = \|e\|_\infty$ and take an index i such that $|e_i| = \|e\|_\infty = \lambda$, where $e = (e_1, e_2, \cdots, e_n)^\top$. From $(\star)$, we get

$$|d_i||e_i| \leq h^4|g_i| + |c_i||e_{i+1}| + |a_{i-1}||e_{i-1}|.$$

Note that $d_i = 2 + h^2 v_i > 0$.

Convergence analysis (cont'd)

- From the previous slide, we have

$$|d_i||e_i| \le h^4|g_i| + |c_i||e_{i+1}| + |a_{i-1}||e_{i-1}|.$$

Hence

$$\begin{aligned} |d_i|\lambda &\le h^4\|g\|_\infty + c_i\lambda + |a_{i-1}|\lambda, \\ \lambda(|d_i| - |c_i| - |a_{i-1}|) &\le h^4\|g|_\infty, \\ h^2 v_i \lambda &\le h^4\|g\|_\infty, \\ \|e\|_\infty &\le h^2(\|g\|_\infty / \inf v(t)). \end{aligned}$$

- Note that $\|g\|_\infty \le \|x^{(4)}\|_\infty/12 + \|x^{(3)}\|_\infty/6$. The expression $\|g\|_\infty / \inf v(t)$ is a bound independent of h. Thus, we see that $\|e\|_\infty$ is $O(h^2)$.

Collocation method

Suppose that we have a linear differential operator L and we wish to solve the equation:

$$Lu(t) = f(t), \quad a < t < b,$$

where f is given and u is sought.

- Let $\{v_1, v_2, \cdots, v_n\}$ be a set of functions that are linearly independent. Suppose that

$$u(t) \approx c_1 v_1(t) + c_2 v_2(t) + \cdots + c_n v_n(t), \quad c_i \in \mathbb{R}.$$

- Then solve $L(\sum_{j=1} c_j v_j(t)) = f(t)$. How to determine c_j?
- Let t_i, $i = 1, 2, \cdots, n$, be n prescribed points (collocation points) in the domain of u and f. Then we require the following equations to determine c_j, $j = 1, 2, \cdots, n$:

$$\sum_{j=1}^{n} c_j (Lv_j)(t_i) = f(t_i), \quad i = 1, 2, \cdots, n.$$

- This is a system of n linear equations in n unknowns c_j. The functions v_j and the points t_i should be chosen so that the matrix with entries $(Lv_j)(t_i)$ is nonsingular.

Collocation method for Sturm-Liouville BVPs

- Consider a Sturm-Liouville two-point BVP:

$$\begin{cases} u''(t) + p(t)u'(t) + q(t)u(t) &= f(t), \quad 0 < t < 1, \\ u(0) &= 0, \\ u(1) &= 0, \end{cases} \qquad (\star)$$

  where p, q, f are given continuous functions on $[0,1]$

- Let $Lu := u'' + pu' + qu$. Define the vector space

$$V = \{u \in C^2(0,1) \cap C[0,1] : u(0) = u(1) = 0\}.$$

  If u is an exact solution of $(\star)$, then $u \in V$.

- One set of functions is given by

$$v_{jk}(t) = t^j(1-t)^k \in C^2[0,1], \quad 1 \le j \le m, 1 \le k \le n.$$

Variational formulation of a 1-dim model problem

Consider the following two-point boundary value problem (BVP):

$$\begin{cases} -u''(x) = f(x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \qquad (D)$$

where f is a given function in $C[0,1]$.

Remark: *Problem (D) has a unique classical solution* $u \in C^2(0,1) \cap C[0,1]$.

Some notation and definitions

- Define $(v,w) := \int_0^1 v(x)w(x)dx$ for real-valued piecewise continuous and bounded functions v and w on $[0,1]$.
- Define $V := \{v |\ v \in C[0,1], v(0) = v(1) = 0, v'$ *is piecewise continuous and bounded on* $[0,1]\}$.
- $F : V \to \mathbb{R}$,
  $F(v) := \frac{1}{2}(v',v') - (f,v) = \frac{1}{2}\int_0^1 (v'(x))^2 dx - \int_0^1 f(x)v(x)dx.$

  (represents the total potential energy)
- Define the following minimization and variational problems:

$$\text{Find } u \in V \text{ such that } F(u) \le F(v), \quad \forall\, v \in V. \qquad (M)$$

$$\text{Find } u \in V \text{ such that } (u',v') = (f,v), \quad \forall\, v \in V. \qquad (V)$$

(D) $\Rightarrow$ (V)

The solution of problem (D) is also a solution of problem (V):

$$\because -u''(x) = f(x), \quad 0 < x < 1.$$

$$\therefore \int_0^1 -u''(x)v(x)dx = \int_0^1 f(x)v(x)dx, \quad \forall\, v \in V.$$

$$\therefore (-u'', v) = (f, v), \quad \forall\, v \in V.$$

$$\therefore (u', v') - u'(x)v(x)\Big|_0^1 = (f, v), \quad \forall\, v \in V. \qquad \text{(integration by parts)}$$

$$\therefore (u', v') = (f, v), \quad \forall\, v \in V.$$

(V) ⇔ (M)

Problems (V) and (M) have the same solutions:

- (V) $\Rightarrow$ (M): Let u be a solution of problem (V). Let $v \in V$ and $w = v - u \in V$. Then $v = u + w$ and

$$\begin{aligned} F(v) &= F(u+w) = \frac{1}{2}((u+w)', (u+w)') - (f, u+w) \\ &= \frac{1}{2}(u', u') + (u', w') + \frac{1}{2}(w', w') - (f, u) - (f, w) \\ &= \frac{1}{2}(u', u') + \frac{1}{2}(w', w') - (f, u) \\ &\geq \frac{1}{2}(u', u') - (f, u) = F(u). \end{aligned}$$

- (M) $\Rightarrow$ (V): Let u be a solution of problem (M). Then for any $v \in V$, $\varepsilon \in \mathbb{R}$, we have $F(u) \leq F(u + \varepsilon v)$, since $u + \varepsilon v \in V$. Define

$$\begin{aligned} g(\varepsilon) &:= F(u+\varepsilon v) = \frac{1}{2}((u+\varepsilon v)', (u+\varepsilon v)') - (f, u+\varepsilon v) \\ &= \frac{1}{2}(u', u') + \frac{1}{2}\varepsilon^2(v', v') + \varepsilon(u', v') - (f, u) - \varepsilon(f, v). \end{aligned}$$

$\because g'(\varepsilon) = (u', v') + \varepsilon(v', v') - (f, v)$ and $g'(0) = 0$.
$\therefore 0 = g'(0) = (u', v') - (f, v)$.

Both problems (V) & (M) have at most one solution

It suffices to prove that problem (V) has at most one solution.
Suppose that u_1 and u_2 are solutions of problem (V). Then

$$\begin{aligned} (u_1', v') &= (f, v) \quad \forall\, v \in V, \\ (u_2', v') &= (f, v) \quad \forall\, v \in V. \end{aligned}$$

$\therefore (u_1' - u_2', v') = 0 \quad \forall\, v \in V.$

Taking $v = u_1 - u_2$, we have $(u_1' - u_2', u_1' - u_2') = 0$.

$\therefore \int_0^1 (u_1'(x) - u_2'(x))^2 dx = 0.$

$\therefore u_1'(x) - u_2'(x) = 0, x \in [0,1]$ a.e.

$\therefore u_1 - u_2$ is a step function on $[0,1]$.

$\because u_1 - u_2$ is continuous on $[0,1]$.

$\therefore u_1 - u_2$ is a constant function on $[0,1]$.

$\because u_1(0) = u_1(1) = 0$ and $u_2(0) = u_2(1) = 0$.

$\therefore u_1 - u_2 \equiv 0$ on $[0,1]$.

That is, $u_1(x) = u_2(x), \forall\, x \in [0,1]$.

(V) + smoothness $\Rightarrow$ (D)

Let u be a solution of problem (V). Then $(u', v') = (f, v), \forall\, v \in V$.

$\therefore \int_0^1 u'(x)v'(x)dx - \int_0^1 f(x)v(x)dx = 0, \quad \forall\, v \in V.$

Suppose that u'' exists and continuous on $[0,1]$, i.e., $u \in C^2[0,1]$.

Then $-\int_0^1 u''(x)v(x)dx - \int_0^1 f(x)v(x)dx = 0, \quad \forall\, v \in V.$

$\therefore -\int_0^1 (u''(x) + f(x))v(x)dx = 0, \quad \forall\, v \in V.$

By the sign-preserving property for continuous functions, we can conclude that
$u''(x) + f(x) = 0, \forall\, x \in [0,1].$

$\therefore u$ is a solution of problem (D).

FEM for the model problem with piecewise linear functions

Construct a finite-dimensional space V_h (finite element space):

Let $0 = x_0 < x_2 < \cdots < x_M < x_{M+1} = 1$ be a partition of $[0,1]$.

[Insert partition figure here!]

Define

- $I_j := [x_{j-1}, x_j], \quad j = 1, 2, \cdots, M+1.$
- $h_j := x_j - x_{j-1}, \quad j = 1, 2, \cdots, M+1.$
- $h := \max_{j=1,2,\cdots,M+1} h_j.$ (a measure of how fine the partition is)

Define

$$V_h := \{v_h \in V \mid v_h \text{ is linear on each subinterval } I_j, v_h(0) = v_h(1) = 0\}.$$

Notice that $V_h \subseteq V$.

Construct a basis of V_h

Here is a typical $v_h \in V_h$:

[Insert v_h figure here!]

For $j = 1, 2, \cdots, M$, we define $\varphi_j \in V_h$ such that

$$\varphi_j(x_i) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

[Insert φ_j figure here!]

Then we have

- $\{\varphi_j\}_{j=1}^M$ is a basis of the finite-dimensional vector space V_h.
- For each $v_h \in V_h$, v_h can be written as a unique linear combination of φ_j's: $v_h(x) = \sum_{j=1}^M \eta_j \varphi_j(x)$, where $\eta_j = v_h(x_j)$.

Numerical methods for solution of problem (D)

We now define the following two numerical methods for approximating the solution of problem (D):

- Ritz method:

$$\textit{Find } u_h \in V_h \textit{ such that } F(u_h) \le F(v_h), \quad \forall\, v_h \in V_h. \tag{M_h}$$

- Galerkin method (finite element method):

$$\textit{Find } u_h \in V_h \textit{ such that } (u_h', v_h') = (f, v_h), \quad \forall\, v_h \in V_h. \tag{V_h}$$

One can claim that $(M_h) \Leftrightarrow (V_h)$**.**

- $(V_h) \Longleftrightarrow$ *Find* $u_h \in V_h$ *such that* $(u_h', \varphi_i') = (f, \varphi_i), 1 \le i \le M$.

  Proof.

  $(\Rightarrow)$: trivial!

  $(\Leftarrow)$: For any $v_h \in V_h$, we have $v_h = \sum_{i=1}^{M} \eta_i \varphi$, for some $\eta_i \in \mathbb{R}, 1 \le i \le M$.

  $$\therefore (u_h', v_h') = (u_h', \sum_{i=1}^{M} \eta_i \varphi_i') = \sum_{i=1}^{M} \eta_i (u_h', \varphi_i')$$

  $$= \sum_{i=1}^{M} \eta_i (f, \varphi_i) = (f, \sum_{i=1}^{M} \eta_i \varphi_i) = (f, v_h).$$

- *Find* $u_h \in V_h$ *such that* $(u_h', \varphi_i') = (f, \varphi_i),\ 1 \le i \le M \Longleftrightarrow A\xi = b$.

  Proof. Let $u_h(x) = \sum_{j=1}^{M} \xi_j \varphi_j(x)$, where $\xi_j = u_h(x_j), 1 \le j \le M$, are unknown. Then

  $$(u_h', \varphi_i') = (f, \varphi_i),\ 1 \le i \le M \Leftrightarrow (\sum_{j=1}^{M} \xi_j \varphi_j', \varphi_i') = (f, \varphi_i),\ 1 \le i \le M$$

  $$\Leftrightarrow \sum_{j=1}^{M} \xi_j (\varphi_j', \varphi_i') = (f, \varphi_i),\ 1 \le i \le M \Leftrightarrow A\xi = b.$$

$A\xi = b$

$A = (a_{ij})_{M\times M}$: stiffness matrix; $b = (b_i)_{M\times 1}$: load vector; $\xi = (\xi_i)_{M\times 1}$: unknown vector.

$$\begin{bmatrix} (\varphi_1', \varphi_1') & (\varphi_2', \varphi_1') & \cdots & (\varphi_M', \varphi_1') \\ (\varphi_1', \varphi_2') & (\varphi_2', \varphi_2') & \cdots & (\varphi_M', \varphi_2') \\ \vdots & \vdots & \vdots & \vdots \\ (\varphi_1', \varphi_M') & (\varphi_2', \varphi_M') & \cdots & (\varphi_M', \varphi_M') \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{bmatrix} = \begin{bmatrix} (f, \varphi_1) \\ (f, \varphi_2) \\ \vdots \\ (f, \varphi_M) \end{bmatrix}.$$

Some remarks

- $\because (\varphi_j', \varphi_i') = 0$ if $|i-j| > 1 \quad \therefore A$ is a tri-diagonal matrix.
- $\because a_{ij} = (\varphi_j', \varphi_i') = (\varphi_i', \varphi_j') = a_{ji} \quad \therefore A$ is symmetric!
- Claim: A is positive definite.
  For any given $\eta = (\eta_1, \eta_2, \cdots, \eta_M)^\top \in \mathbb{R}^M$, define
  $v_h(x) := \sum_{i=1}^{M} \eta_i \varphi_i(x)$. Then

$$0 \le (v_h', v_h') = (\sum_{i=1}^{M} \eta_i \varphi_i', \sum_{j=1}^{M} \eta_j \varphi_j') = \sum_{i,j=1}^{M} \eta_i (\varphi_i', \varphi_j') \eta_j = \eta \cdot A\eta.$$

  If $(v_h', v_h') = 0$, then $\int_0^1 (v_h'(x))^2 dx = 0. \Longrightarrow v_h'(x) = 0$ a.e.
  $\because v_h \in V_h$, v_h is continuous on $[0,1]$ and $v_h(0) = v_h(1) = 0$.
  $\therefore v_h \equiv 0$ on $[0,1]$, i.e., $\eta = \mathbf{0}$. $\therefore \eta \cdot A\eta > 0, \forall\, \eta \in \mathbb{R}^M, \eta \ne \mathbf{0}$.
- $\because A$ is SPD $\therefore A$ is nonsingular $\therefore A\xi = b$ has a unique solution!

Evaluate a_{jj} and $a_{j-1,j}$

[Insert a figure of φ_{j-1} and φ_j here!]

For $j = 1, 2, \cdots, M$, we have

$$\begin{aligned}
(\varphi_j', \varphi_j') &= \int_{x_{j-1}}^{x_j} (\varphi_j')^2 dx + \int_{x_j}^{x_{j+1}} (\varphi_j')^2 dx \\
&= \int_{x_{j-1}}^{x_j} \frac{1}{h_j^2} dx + \int_{x_j}^{x_{j+1}} \frac{1}{h_{j+1}^2} dx = \frac{1}{h_j} + \frac{1}{h_{j+1}}, \\
(\varphi_j', \varphi_{j-1}') &= (\varphi_{j-1}', \varphi_j') = -\int_{x_{j-1}}^{x_j} \frac{1}{h_j^2} dx = -\frac{1}{h_j}.
\end{aligned}$$

For uniform partition: $h_j = h = \frac{1-0}{M+1}$. Then $A\xi = b$ becomes

$$\frac{1}{h} \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix} \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_M \end{bmatrix} = \begin{bmatrix} (f, \varphi_1) \\ (f, \varphi_2) \\ \vdots \\ (f, \varphi_M) \end{bmatrix}.$$

Taylor's theorem with Lagrange remainder

If $f \in C^n[a, b]$ and $f^{(n+1)}$ exists on (a, b), then for any points c and x in $[a, b]$ we have

$$f(x) = P_n(x) + E_n(x),$$

where the n-th Taylor polynomial $P_n(x)$ is given by

$$P_n(x) = \sum_{k=0}^{n} \frac{1}{k!} f^{(k)}(c)(x - c)^k$$

and the remainder (error) term $E_n(x)$ is given by

$$E_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi)(x - c)^{n+1}$$

for some point ξ between c and x (means that either $c < \xi < x$ or $x < \xi < c$).

Numerical differentiation

Assume that $u \in C^4[0,1]$ and $0 = x_0 < x_2 < \cdots < x_M < x_{M+1} = 1$ is a uniform partition of $[0,1]$. Then $h_j = h = \frac{1-0}{M+1}$ for $j = 1, 2, \cdots, M+1$.
For $i = 1, 2, \cdots, M$, we have

$$u(x_i + h) = u(x_i) + u'(x_i)h + \frac{1}{2}u''(x_i)h^2 + \frac{1}{6}u^{(3)}(x_i)h^3 + \frac{1}{24}u^{(4)}(\xi_{i1})h^4,$$
$$u(x_i - h) = u(x_i) - u'(x_i)h + \frac{1}{2}u''(x_i)h^2 - \frac{1}{6}u^{(3)}(x_i)h^3 + \frac{1}{24}u^{(4)}(\xi_{i2})h^4,$$

for some $\xi_{i1} \in (x_i, x_i + h)$ and $\xi_{i2} \in (x_i - h, x_i)$.

$$\therefore u(x_i + h) + u(x_i - h) = 2u(x_i) + u''(x_i)h^2 + \frac{1}{24}\{u^{(4)}(\xi_{i1}) + u^{(4)}(\xi_{i2})\}h^4.$$

$$\therefore u''(x_i) = \frac{1}{h^2}\{u(x_i + h) - 2u(x_i) + u(x_i - h)\} - \frac{1}{24}h^2\{u^{(4)}(\xi_{i1}) + u^{(4)}(\xi_{i2})\}.$$

$\because u \in C^4[0,1]$ and $\frac{1}{2}\{u^{(4)}(\xi_{i1}) + u^{(4)}(\xi_{i2})\}$ between $u^{(4)}(\xi_{i1})$ and $u^{(4)}(\xi_{i2})$.

$\therefore$ By IVT, $\exists\ \xi_i$ between ξ_{i1} and ξ_{i2} ($\Rightarrow \xi_i \in (x_i - h, x_i + h)$) such that

$$u^{(4)}(\xi_i) = \frac{1}{2}\{u^{(4)}(\xi_{i1}) + u^{(4)}(\xi_{i2})\}.$$

$$\therefore u''(x_i) = \frac{1}{h^2}\{u(x_i + h) - 2u(x_i) + u(x_i - h)\} - \frac{1}{12}h^2 u^{(4)}(\xi_i),$$

for some $\xi_i \in (x_i - h, x_i + h)$.

Finite difference method for problem (D)

Consider the BVP:

$$\begin{cases} -u''(x) = f(x), & 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \qquad (D)$$

For $i = 1, 2, \cdots, M$, we have

$$-\frac{1}{h^2}\{u(x_i + h) - 2u(x_i) + u(x_i - h)\} + \frac{1}{12}h^2 u^{(4)}(\xi_i) = f(x_i).$$

$$\Rightarrow -\frac{1}{h^2}\{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})\} + \frac{1}{12}h^2 u^{(4)}(\xi_i) = f(x_i).$$

We wish to find $U_i \simeq u(x_i)$ for $i = 1, 2, \cdots, M$ and $U_0 = U_{M+1} := 0$ such that

$$\begin{aligned} -\frac{1}{h^2}\{U_0 - 2U_1 + U_2)\} &= f(x_1). && (i = 1) \\ -\frac{1}{h^2}\{U_1 - 2U_2 + U_3)\} &= f(x_2). && (i = 2) \\ &\vdots \\ -\frac{1}{h^2}\{U_{M-1} - 2U_M + U_{M+1})\} &= f(x_M). && (i = M) \end{aligned}$$

Finite difference method for problem (D) (cont'd)

Finally, we reach at the following linear system:

$$\frac{1}{h^2}\begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}\begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_M \end{bmatrix} = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_M) \end{bmatrix}.$$

A comparison: what is the difference between FEM with piecewise linear basis functions and FDM for problem (D)? **Answer:** They are essentially the same!

Consider the first component in the right hand side:

- Finite difference method: $h^2 f(x_1)$.
- Finite element method:

$$h(f,\varphi_1) = h\int_{x_0}^{x_2} f(x)\varphi_1(x)dx \simeq hf(x_1)\int_{x_0}^{x_2}\varphi_1(x)dx = h^2 f(x_1).$$

Homework

Consider the following 1-D reaction-convection-diffusion problem:

$$\begin{cases} -\varepsilon u''(x) + u'(x) + u(x) = 1 & \text{for } x \in (0,1), \\ u(0) = 0, \ u(1) = 0. \end{cases} \qquad (\star)$$

Write the computer codes for numerical solution of problem $(\star)$ by using the following methods on the uniform mesh of $[0,1]$ with mesh size h:

- **Finite difference methods:**
  - Replace $u''(x_i) \approx \frac{U_{i+1}-2U_i+U_{i-1}}{h^2}$ and $u'(x_i) \approx \frac{U_{i+1}-U_{i-1}}{2h}$ with $(\varepsilon, h) = (0.01, 0.1)$ and $(\varepsilon, h) = (0.01, 0.01)$. Plot u_h.
  - Replace $u''(x_i) \approx \frac{U_{i+1}-2U_i+U_{i-1}}{h^2}$ and $u'(x_i) \approx \frac{U_i-U_{i-1}}{h}$ (upwinding) with $(\varepsilon, h) = (0.01, 0.1)$ and $(\varepsilon, h) = (0.01, 0.01)$. Plot u_h.
- **Finite element method:** use piecewise linear finite elements with $(\varepsilon, h) = (0.01, 0.1)$ and $(\varepsilon, h) = (0.01, 0.01)$. Plot u_h.