# MA 3021: Numerical Analysis I
# Numerical Ordinary Differential Equations

Suh-Yuh Yang (楊肅煜)

Department of Mathematics, National Central University
Jhongli District, Taoyuan City 32001, Taiwan

syyang@math.ncu.edu.tw
http://www.math.ncu.edu.tw/∼syyang/

# Initial-value problems

1. Initial-value problem (IVP): 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.

2. Example 1:
$$\begin{cases} x'(t) &= x\tan(t+3), \\ x(-3) &= 1. \end{cases}$$
The analytic solution of this IVP is $x(t) = \sec(t+3)$. The solution is valid only for $-\frac{\pi}{2} < t+3 < \frac{\pi}{2}$.

3. 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 $x(0) = 1 \Rightarrow x = e^t$ particular solution

## Existence and uniqueness

1. Existence: do all IVPs has a solution? No! Some assumptions must be made about $f$, and even then we can expect the solution to exist only in a neighborhood of $t = t_0$.

2. Example:

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

Try $x(t) = \tan t \Rightarrow x(0) = 0$

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 $x \to \infty$. 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 on the interval $-1 \le t \le 1$.

- Consider $f(t,x) = (t + \ sinx)^2$, where $(t_0, x_0) = (0, 3)$.
- Let $R = \{(t,x) : |t| \le \alpha, |x-3| \le \beta\}$. Then $|f(t,x)| \le (\alpha+1)^2 := M$.
- We want $|t-0| \le 1 \le \min\{\alpha, \beta/M\}$.
- Let $\alpha = 1$ then $M = (1+1)^2 = 4$ and force $\beta \ge 4$. By the Existence Theorem, the IVP has a solution on $|t-t_0| \le \min\{\alpha, \beta/M\} = 1$.

# Uniqueness

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

2. 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 $(t_0, x_0)$,*

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

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

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

*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*

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

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

**Note:** Inequality (*) is called the Lipschitz condition in the 2nd variable.

## Example

Prove that

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

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

- 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 $-\infty < x < \infty$.
- By the MVT, $\exists\ \xi$ 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 (*) 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 $a \le t \le b$.

## Numerical methods

1. Consider the IVP:

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

2. Strategy: Instead of finding $x(t)$ for all $t$ in some interval containing $t_0$, we find $x(t)$ at some fixed points.

## Taylor-series method

1. For the Taylor-series method, it is necessary to assume that various partial derivatives of $f$ exist.
2. We use a concrete example to illustrate the method. Consider an IVP as
$$\begin{cases} x'(t) &= \cos t - \sin x + t^2, \\ x(-1) &= 3. \end{cases}$$
3. Assume that we know $x(t)$ and we wish to compute $x(t+h)$. By the Taylor series for $x$, 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)

1. How to compute $x'(t)$, $x''(t)$, $x'''(t)$ and $x^{(4)}(t)$ in the last equation?

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

2. 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.
3. **Definition:** The order of the Taylor-series method is $n$ if terms up to and include $h^n x^{(n)}(t)/n!$ are used.

# 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

1. Estimate of the local truncation error can be done by looking at

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

   Hence

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

2. Replace $x^{(5)}(t+\theta h)$ by a simple finite-difference approximation

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

3. Suppose that the local truncation error (LTE) is $O(h^{n+1})$. The accumulation of all many LTEs gives rise the global truncation error (GTE).

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

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

# Advantages and disadvantages of the Taylor-series method

1. **Disadvantages:**
   - 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.
   - The various derivatives formula need to be programmed.
2. **Advantages:**
   - Conceptual simplicity.
   - 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

1. If $n = 1$, the Taylor series method reduces to Euler's method.
2. Advantage of the method is not to require any differentiation of $f$.
3. Disadvantage of the method is that the necessity of taking small value for $h$ to gain acceptable precision.
4. 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:

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

1. From the Taylor theorem, we have

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

2. By the chain rule, we obtain

$$\begin{cases} x''(t) &= f_t + f_x x' = f_t + f_x f, \\ x'''(t) &= f_{tt} + f_{tx} f + (f_t + f_x f) f_x + f(f_{xt} + f_{xx} f). \end{cases}$$

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

- In the Taylor expansion, we have

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

- Note that the term in the square blankets above can be obtained 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

1. 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

$$x(t+h) \approx x(t) + \frac{1}{2}(F_1 + F_2),$$

where

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

2. It is also known as Heun's method.

## The general second-order Runge-Kutta method

1. In general, the 2nd order RK method needs

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

2. Compare with

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

we have

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

# The modified Euler method

1. The previous method (Heun's method) is obtained by setting
$$\begin{cases} \omega_1 = \omega_2 = 1/2, \\ \alpha = \beta = 1. \end{cases}$$

2. Setting
$$\begin{cases} \omega_1 = 0, \\ \omega_2 = 1, \\ \alpha = \beta = 1/2, \end{cases}$$
we obtain the following modified Euler method:
$$x(t+h) \approx x(t) + F_2,$$
where
$$F_1 = hf(t,x), \quad F_2 = hf(t + \frac{1}{2}h, x + \frac{1}{2}F_1).$$

# 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:

$$x(t+h) \approx x(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4),$$

where

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

# Computer project

1. 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:

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

2. 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**

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

**output** $k, t, x, e$
**end do**

## A system of first-order differential equations

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

$$\left\{\begin{array}{rcl} x_1'(t) & = & f_1(t, x_1, x_2, \cdots, x_n), \\ x_2'(t) & = & f_2(t, x_1, x_2, \cdots, x_n), \\ & \vdots & \\ x_n'(t) & = & f_n(t, x_1, x_2, \cdots, x_n). \end{array}\right. \qquad (*)$$

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

# 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

1. **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 $(*)$ is given by

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

2. 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 1

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

## Example 2

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

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

# Autonomous systems

1. From the theoretical standpoint, there is no loss of generality in assuming that the equations in system $(*)$ 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 $X(t) = (x_0(t), x_1(t), \cdots, x_n(t))^\top$.
2. **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 $X'(t) = F(X)$

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

$$X(t+h) \approx X(t) + \frac{1}{6}(F_1 + 2F_2 + 2F_3 + F_4),$$

where

$$\begin{aligned} F_1 &= hF(X), \\ F_2 &= hF(X + \frac{1}{2}F_1), \\ F_3 &= hF(X + \frac{1}{2}F_2), \\ F_4 &= hF(X + F_3). \end{aligned}$$

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

# 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.

1. Let $\{v_1, v_2, \cdots, v_n\}$ be a set of functions that are linearly independent. Suppose that $u(t) \approx c_1 v_1(t) + c_2 v_2(t) + \cdots + c_n v_n(t)$, where $c_i \in \mathbb{R}$.
2. Then solve $L(\sum_{j=1} c_j v_j(t)) = f(t)$. How to determine $c_j$, $j = 1, 2, \cdots, n$?
3. 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 $c_j$, $j = 1, 2, \cdots, n$:

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

# Collocation method for Sturm-Liouville BVPs

1. Consider a Sturm-Liouville two-point BVP:

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

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

2. 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 $(*)$, then $u \in V$.

3. One set of functions is given by

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

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

## Some notation and definitions

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

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

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

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

## (D) $\Rightarrow$ (V)

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

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

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

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

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

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

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

1. (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

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

2. (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

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

Define

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

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 $\varphi_j$ figure here!]

Then we have

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

1. Ritz method:

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

2. Galerkin method (finite element method):

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

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

1. $(V_h) \Longleftrightarrow$ *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 $\eta_i \in \mathbb{R}$, $1 \le i \le M$.

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

2. *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

1. $\because (\varphi_j', \varphi_i') = 0$ if $|i-j| > 1 \quad \therefore A$ is a tri-diagonal matrix.
2. $\because a_{ij} = (\varphi_j', \varphi_i') = (\varphi_i', \varphi_j') = a_{ji} \quad \therefore A$ is symmetric!
3. 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., $\eta = \mathbf{0}$.

   $\therefore \eta \cdot A\eta > 0, \forall\, \eta \in \mathbb{R}^M, \eta \neq \mathbf{0}$.
4. $\because A$ 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 $\varphi_{j-1}$ and $\varphi_j$ 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 $\xi$ 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

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

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

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

$\because u \in C^4[0,1]$ 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 $\xi_{i1}$ and $\xi_{i2}$ ($\Rightarrow \xi_i \in (x_i - h, x_i + h)$) such that

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

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

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

# Finite difference method for problem (D)

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

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:

1. Finite difference method: $hf(x_1)$.
2. Finite element method:

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

# Computer project

Consider the following one-dimensional convection-diffusion problem:

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

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

1. Replace $u''(x_i) \approx \dfrac{U_{i+1} - 2U_i + U_{i-1}}{h^2}$ and $u'(x_i) \approx \dfrac{U_{i+1} - U_{i-1}}{2h}$ and consider $(\varepsilon, h) = (0.01, 0.1)$, $(\varepsilon, h) = (0.01, 0.01)$. Plot $u_h$.
2. Replace $u''(x_i) \approx \dfrac{U_{i+1} - 2U_i + U_{i-1}}{h^2}$ and $u'(x_i) \approx \dfrac{U_i - U_{i-1}}{h}$ (upwinding) and consider $(\varepsilon, h) = (0.01, 0.1)$, $(\varepsilon, h) = (0.01, 0.01)$. Plot $u_h$.