AM213B Numerical Methods for the Solution of Differential Equations

Lecture 09

List of topics in this lecture

- Numerical solution of the heat equation: initial boundary value problem (IBVP), IVP, numerical grid for (x, t), discretization of PDE, FTCS method, BTCS method
- Local truncation error, consistency, order of accuracy, stability
- Stability condition of FTCS; BTCS is unconditionally stable
- Function norm, vector norm, norm of numerical solution

We now discuss numerical solution of PDEs.

Numerical solution of the heat equation

IBVP (initial boundary value problem) of the heat equation:

$$\begin{cases} u_t = u_{xx} \\ u(x,t_0) = f(x) \\ u(a,t) = g_1(x), \quad u(b,t) = g_2(x) \end{cases}$$

Numerical discretization:

Numerical grid:

$$\Delta x = \frac{b-a}{N+1}, \qquad x_i = a + i\,\Delta x$$

$x_0 = a$, $\quad x_{N+1} = b$, $\qquad$ internal points = $\{x_i,\ 1 \le i \le N\}$

$t_n = t_0 + n\Delta t$

Notation:

$u(x_i, t_n)$: exact solution at (x_i, t_n)

u_i^n: numerical approximation of $u(x_i, t_n)$

Discretization of derivatives:

$$u_{xx}\Big|_{(x_i,t_n)} \approx \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}$$

$$u_t\Big|_{(x_i,t_n)} \approx \frac{u_i^{n+1} - u_i^n}{\Delta t}$$

Discretization of PDE:

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{(\Delta x)^2}$$

This is called the FTCS method (Forward-time, Central-space method).

Let $r = \dfrac{\Delta t}{(\Delta x)^2}$. We write out the FTSC method in terms of *r*, along with the initial and boundary conditions.

The FTCS method for IBVP:

$$u_i^{n+1} = u_i^n + r\left(u_{i+1}^n - 2u_i^n + u_{i-1}^n\right), \qquad 1 \le i \le N, \qquad r = \frac{\Delta t}{(\Delta x)^2}$$

$$u_i^0 = f(x_i), \qquad 1 \le i \le N$$

$$u_0^n = g_1(t_n), \quad u_{N+1}^n = g_2(t_n), \qquad n \ge 0$$

Sometimes, for theoretical simplicity and convenience, we also consider initial value problems over the infinite domain. But keep in mind that in computer implementation and in real applications, only IBVPs make practical sense.

IVP (initial value problem) of the heat equation:

$$\begin{cases} u_t = u_{xx} \\ u(x,t_0) = f(x) \end{cases}$$

The FTCS method for IVP:

$$u_i^{n+1} = u_i^n + r(u_{i+1}^n - 2u_i^n + u_{i-1}^n), \qquad -\infty < i < +\infty, \qquad r = \frac{\Delta t}{(\Delta x)^2}$$

$$u_i^0 = f(x_i), \qquad -\infty < i < +\infty$$

Next, we introduce

Local truncation error

Consistency, order of accuracy

Stability

Global error

Convergence

Local truncation error:

When we substitute an exact solution into the numerical method, the residual term is called the local truncation error (LTE) and is denoted by $e_i^n\left(\Delta x, \Delta t\right)$.

Consistency:

Suppose a numerical method satisfies

$$\lim_{\substack{\Delta x\to 0\\ \Delta t\to 0}} \frac{e_i^n(\Delta x, \Delta t)}{\Delta t} = 0$$

Then we say the method is consistent with the PDE.

Order of accuracy

The behavior of $\dfrac{e_i^n(\Delta x, \Delta t)}{\Delta t}$ describes the order of accuracy.

For example, if $\dfrac{e_i^n(\Delta x, \Delta t)}{\Delta t} = O\left(\Delta t + (\Delta x)^2\right)$, then we say the method is first order in time and second order in space.

Example: The FTCS method is consistent with the heat equation.

The FTCS method:

$$u_i^{n+1} - u_i^n - r(u_{i+1}^n - 2u_i^n + u_{i-1}^n) = 0\,, \qquad r = \frac{\Delta t}{(\Delta x)^2}$$

Its local truncation error is defined as

$$e_i^n(\Delta x, \Delta t) = u(x_i, t_{n+1}) - u(x_i, t_n) - r\left(u(x_{i+1}, t_n) - 2u(x_i, t_n) + u(x_{i-1}, t_n)\right)$$

Expanding everything around (x_i, t_n) gives us

$$u(x_i, t_{n+1}) - u(x_i, t_n) = u_t \Delta t + \frac{1}{2} u_{tt} (\Delta t)^2 + o\left((\Delta t)^2\right)$$

$$u(x_{i+1}, t_n) - 2u(x_i, t_n) + u(x_{i-1}, t_n) = 2\cdot\frac{1}{2!} u_{xx} (\Delta x)^2 + 2\cdot\frac{1}{4!} u_{xxxx} (\Delta x)^4 + o\left((\Delta x)^4\right)$$

Using $u_t = u_{xx}$, $u_{tt} = u_{xxxx}$ and $r = \dfrac{\Delta t}{(\Delta x)^2}$, we write the local truncation error as:

$$e_i^n(\Delta x, \Delta t) = \Delta t\left[u_{xx} + \frac{1}{2}u_{xxxx}\Delta t + o\left(\Delta t\right)\right] - \Delta t\left[u_{xx} + \frac{1}{12}u_{xxxx}(\Delta x)^2 + o\left((\Delta x)^2\right)\right]$$

$$= \Delta t\, u_{xxxx}\left[\frac{1}{2}\Delta t - \frac{1}{12}(\Delta x)^2 + \cdots\right]$$

$$\frac{e_i^n(\Delta x, \Delta t)}{\Delta t} = u_{xxxx}\left[\frac{1}{2}\Delta t - \frac{1}{12}(\Delta x)^2 + \cdots\right] \to 0 \quad \text{as } (\Delta x, \Delta t) \to 0$$

The FTCS is first order in time and second order in space.

The BTCS method (Backward-time, Central-space) for IVP

We discretize u_t and u_{xx} at (x_i, t_{n+1}).

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \frac{u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}}{(\Delta x)^2}$$

We write out the BTCS method for the IVP

$$u_i^{n+1} = u_i^n + r\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right), \qquad -\infty < i < +\infty, \qquad r = \frac{\Delta t}{(\Delta x)^2}$$

$$u_i^0 = f(x_i), \qquad -\infty < i < +\infty$$

Its local truncation error has the expression:

$$e_i^n(\Delta x, \Delta t) = \Delta t\, u_{xxxx}\left[-\frac{1}{2}\Delta t - \frac{1}{12}(\Delta x)^2 + \cdots\right]$$

(The derivation is based on expanding everything around (x_i, t_{n+1}).)

The BTCS is first order in time and second order in space.

Linear operator notation:

We write a general linear numerical method as a linear operator.

$$u^{n+1} = L_{num}(u^n)$$

where

L_{num} = linear operator representing the method

$u^n = \left\{ u_i^n \ ,\ 1 \le i \le N \right\}$ in an IBVP

Or $u^n = \left\{ u_i^n \ ,\ -\infty < i < \infty \right\}$ in an IVP

Remark:

Here we are solving a linear PDE and "a linear numerical method" means it is linear in u^n. That is, L_{num} is a linear operator. This is different from the context of "linear multistep method" where we are solving non-linear ODEs.

With the operator notation, the local truncation error is

$$\underbrace{\left\{ e_i^n(\Delta x, \Delta t) \right\}}_{\text{Vector of LTE at } t_{n+1}} = \underbrace{\left\{ u(x_i, t_{n+1}) \right\}}_{\substack{\text{Vector of} \\ \text{exact } u \text{ at } t_{n+1}}} - L_{num} \underbrace{\left\{ u(x_i, t_n) \right\}}_{\substack{\text{Vector of} \\ \text{exact } u \text{ at } t_n}}$$

This is also the error of one step.

After advancing n steps in time, the numerical solution is

$$u^n = (L_{num})^n (u^0)$$

We want that the ratio $\dfrac{||u^n||}{||u^0||}$ stay bounded for all u^0 and for all $(n\Delta t) \le T$.

This is equivalent to that the norm $\left\| (L_{num})^n \right\|$ remain bounded for all $(n\Delta t) \le T$, which leads to the definition of stability below.

Definition: (stability of linear numerical methods)

Consider a linear numerical method, $u^{n+1} = L_{\text{num}}(u^n)$.

Suppose for any $T > 0$, there exists a constant C_T such that

$$\left\| (L_{num})^n \right\| \le C_T \quad \text{for all } \ n\Delta t \le T$$

Then we say the numerical method L_{num} is stable.

Remark:

When the numerical operator L_{num} is linear, the three items below are equivalent.

- $\left\| (L_{num})^n u_0 - (L_{num})^n v_0 \right\| \le C_T \left\| (u_0 - v_0) \right\|$ for all u_0, v_0 and for all $n\Delta t \le T$
- $\left\| (L_{num})^n u_0 \right\| \le C_T \left\| u_0 \right\|$ for all u_0 and for all $n\Delta t \le T$
- $\left\| (L_{num})^n \right\| \le C_T$ for all $n\Delta t \le T$

Theorem:

If $||L_{\text{num}}|| \le 1+C\,\Delta t$, then the method L_{num} is stable.

Proof:

$$\left\|\left(L_{num}\right)^n\right\| \le \left\|L_{num}\right\|^n \le (1+C\Delta t)^n$$

$$\le \exp(C\,n\Delta t) = \exp(C\,T) \qquad \text{for all } n\Delta t \le T$$

Remark:

$||L_{\text{num}}|| \le 1+C\,\Delta t$ is easier to check than $||(L_{\text{num}})^n|| \le C_T$.

$||L_{\text{num}}|| \le 1+C\,\Delta t$ is a sufficient condition for $||(L_{\text{num}})^n|| \le C_T$.

For multistep methods, $||L_{\text{num}}|| \le 1+C\,\Delta t$ is not a necessary condition for $||(L_{\text{num}})^n|| \le C_T$.

We will almost exclusively discuss only single step method.

Stability of the FTCS method (IVP)

$$u_i^{n+1} = u_i^n + r(u_{i+1}^n - 2u_i^n + u_{i-1}^n), \qquad -\infty < i < +\infty, \qquad r = \frac{\Delta t}{(\Delta x)^2}$$

We write the method as

$$u_i^{n+1} = r\,u_{i+1}^n + (1-2r)u_i^n + r\,u_{i-1}^n$$

We consider the infinity norm:

$$\left\|\vec{u}\right\|_\infty \equiv \sup_{-\infty<i<\infty} \left|u_i\right|$$

This is called the supremum of u (the least upper bound).

We fix $r = \dfrac{\Delta t}{(\Delta x)^2}$ while both $\Delta x \to 0$ and $\Delta t \to 0$.

Theorem:

$$\text{FTCS method is } \begin{cases} \text{stable} & \text{if } r \le \dfrac{1}{2} \\ \text{unstable} & \text{if } r > \dfrac{1}{2} \end{cases}$$

Proof:

$\boxed{\text{For } r \le \frac{1}{2}}$, all three coefficients, r, $(1-2r)$ and r, are <u>non-negative</u>.

$$\left|u_i^{n+1}\right| = \left|r u_{i+1}^n + (1-2r)u_i^n + r u_{i-1}^n\right| \le \left|r u_{i+1}^n\right| + \left|(1-2r)u_i^n\right| + \left|r u_{i-1}^n\right|$$

$$= r\left|u_{i+1}^n\right| + (1-2r)\left|u_i^n\right| + r\left|u_{i-1}^n\right|$$

$$\le r\left\|u^n\right\|_\infty + (1-2r)\left\|u^n\right\|_\infty + r\left\|u^n\right\|_\infty = \left\|u^n\right\|_\infty$$

Since this is true at all values of i, we conclude

$$\left\|u^{n+1}\right\|_\infty \le \left\|u^n\right\|_\infty \qquad \text{for all } u^n$$

==> It is stable.

$\boxed{\text{For } r > \frac{1}{2}}$, we notice that coefficients r, $(1-2r)$, and r alternate in sign.

$r > 0$, $(1-2r) < 0$, $r > 0$

Consider a solution of the special form:

$$u_i^n = \rho^n (-1)^i$$

where ρ is called the magnification factor.

Substituting into the FTCS method, we obtain

$$\rho^{n+1}(-1)^i = r\rho^n(-1)^{i+1} + (1-2r)\rho^n(-1)^i + r\rho^n(-1)^{i-1}$$

==> $\rho = -r + (1-2r) - r = 1-4r$

For $r > 1/2$, the magnification factor $\rho = (1-4r)$ is negative and satisfies

$$|\rho| = 4r - 1 > 1$$

==> $u_i^n = \rho^n(-1)^i$ grows unbounded as $n \to \infty$.

==> It is unstable.

End of proof

Stability of the BTCS method (IVP)

$$u_i^{n+1} = u_i^n + r\left(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1}\right), \qquad -\infty < i < +\infty, \qquad r = \frac{\Delta t}{(\Delta x)^2} \tag{E01}$$

Due to the infinite domain of IVP, the BTCS method (E01) needs an additional constraint to ensure that u^{n+1} is uniquely defined given u^n.

$$\boxed{\text{Constraint: } \left|u_i^{n+1}\right| \text{ is bounded as } |i| \to \infty} \tag{C01}$$

Claim:

Without constraint (C01), u^{n+1} in (E01) is not uniquely defined!

Demonstration:

Consider the special case of $\left\{u_i^n \equiv 0\ ,\ -\infty < i < \infty\right\}$.

It is straightforward to verify that $\left\{u_i^{n+1} \equiv 0\ ,\ -\infty < i < \infty\right\}$ is a solution of (E01).

Besides this trivial solution, we look for a solution of the form $u_i^{n+1} = \alpha^i$.

Substituting $u_i^n \equiv 0$ and $u_i^{n+1} = \alpha^i$ into (E01), we get

$$\alpha^i = 0 + r\,\alpha^i\left(\alpha - 2 + \frac{1}{\alpha}\right)$$

$$\Longleftrightarrow \quad \alpha - 2 + \frac{1}{\alpha} = \frac{1}{r}$$

$$\Longleftrightarrow \quad \alpha^2 - \left(2 + \frac{1}{r}\right)\alpha + 1 = 0$$

The quadratic equation has two roots: $|\alpha_1| > 1$ and $|\alpha_2| < 1$.

$u_i^{n+1} = (\alpha_1)^i$ is a solution of (E01) satisfying $\lim\limits_{i\to-\infty}\left|u_i^{n+1}\right| = 0$, $\lim\limits_{i\to+\infty}\left|u_i^{n+1}\right| = \infty$.

$u_i^{n+1} = (\alpha_2)^i$ is a solution of (E01) satisfying $\lim\limits_{i\to-\infty}\left|u_i^{n+1}\right| = \infty$, $\lim\limits_{i\to+\infty}\left|u_i^{n+1}\right| = 0$.

Therefore, without imposing constraint (C01), u^{n+1} in (E01) is not uniquely defined.

Theorem:

The BTCS method is unconditionally stable. That is, it is stable for any $r > 0$.

Proof:

The BTCS:

$$u_i^{n+1} = u_i^n + r(u_{i+1}^{n+1} - 2u_i^{n+1} + u_{i-1}^{n+1})$$

We rewrite it as

$$u_i^n = (1+2r)u_i^{n+1} - r(u_{i+1}^{n+1} + u_{i-1}^{n+1})$$

With constraint (C01), we know that $\sup\limits_i |u_i^{n+1}|$ is finite.

From the definition of sup, for any $\varepsilon > 0$, there exists index i_0 such that

$$|u_{i_0}^{n+1}| \ge \left(\sup_i |u_i^{n+1}| - \varepsilon\right)$$

At index i_0, the BTCS gives us

$$\left|u_{i_0}^n\right| = \left|(1+2r)u_{i_0}^{n+1} - r(u_{i_0+1}^{n+1} + u_{i_0-1}^{n+1})\right| \ge (1+2r)\left|u_{i_0}^{n+1}\right| - r\left|u_{i_0+1}^{n+1} + u_{i_0-1}^{n+1}\right|$$

$$\ge (1+2r)\left(\sup_i |u_i^{n+1}| - \varepsilon\right) - 2r \cdot \sup_i |u_i^{n+1}|$$

$$= \sup_i |u_i^{n+1}| - (1+2r)\varepsilon$$

Using $\sup_i |u_i^n| \ge |u_{i_0}^n|$, we have

$$\sup_i \left|u_i^n\right| \ge \sup_i \left|u_i^{n+1}\right| - (1+2r)\varepsilon$$

Since this is true for any ε > 0, we conclude

$$\sup_i \left|u_i^n\right| \ge \sup_i \left|u_i^{n+1}\right|$$

That is,

$$\left\|u^{n+1}\right\|_\infty \le \left\|u^n\right\|_\infty \qquad \text{for all } u^n$$

==> It is stable for any $r > 0$.

Function norm, vector norm, norm of numerical solution

We clarify the definition of norm for numerical solution.

We first look at the new situation we are facing in numerical solution of PDEs.

Comparison of ODE solutions and PDE solutions:

ODE: $u_n \approx u(t_n)$, a vector of fixed size;

size does not increase with numerical resolution.

PDE: $u^n = \left\{u_i^n\ ,\ 1 \le i \le N\right\} \approx \left\{u(x_i, t_n)\ ,\ 1 \le i \le N\right\}$, a vector of size N;

size increases with numerical resolution.

For a PDE, the numerical solution at t_n is both

- a discrete vector and
- an approximation to a continuous function.

This new situation demands that the norm of numerical solution u^n should have features of both the vector norm and the function norm.

Consider a continuous function $u(x)$ over $[a, b]$ and a discrete version of $u(x)$:

$$u(x), \qquad a \le x \le b$$
$$\vec{u} = \left\{ u_i = u(x_i), \quad 1 \le i \le N \right\}$$

Function norm:

$$\left\| u \right\|_p = \left(\int_a^b \left| u(x) \right|^p dx \right)^{\frac{1}{p}} = \text{finite}$$

Vector norm:

$$\left\| \vec{u} \right\|_p = \left(\sum_{i=1}^{N} \left| u_i \right|^p \right)^{\frac{1}{p}}$$

$$= \left(\frac{1}{\Delta x} \sum_{i=1}^{N} \left| u(x_i) \right|^p \Delta x \right)^{\frac{1}{p}} = \left(\frac{1}{\Delta x} \right)^{\frac{1}{p}} \left(\sum_{i=1}^{N} \left| u(x_i) \right|^p \Delta x \right)^{\frac{1}{p}}$$

$$\approx \left(\frac{1}{\Delta x} \right)^{\frac{1}{p}} \left(\int_a^b \left| u(x) \right|^p dx \right)^{\frac{1}{p}} \to \infty \qquad \text{as } \Delta x \to 0$$

We adopt the norm below for numerical solutions of PDEs.

Norm of numerical solution:

$$\boxed{\left\| \vec{u} \right\|_p = \left(\sum_{i=1}^{N} \left| u_i \right|^p \Delta x \right)^{\frac{1}{p}}}$$