Lecture 1: Introduction

1 Spectral method v. Finite-Difference v. Finite-Element

In this section we solve following heat equation using three different numerical methods: Finite Difference, Finite Element, and Spectral method.

$$\begin{cases} u_t = u_{xx}, \\ u(x,0) = u_0(x), \\ u(x+2\pi) = u(x). \end{cases} \tag{1}$$

- *Finite difference method.* We take N equidistant grid points on the interval $[0, 2\pi]$: $0 = x_0 < x_1 < \cdots < x_N = 2\pi$, $x_{j+1} - x_j = h = 2\pi/N$. Let $u_j(t) = u(x_j, t)$. By using approximation

$$u_{xx}(x_j) \approx \frac{u_{j+1} - 2\,u_j + u_{j-1}}{h^2}$$

we can discretize the heat equation (1) as

$$\begin{cases} u_j'(t) = \dfrac{u_{j+1} - 2\,u_j + u_{j-1}}{h^2}, & j = 0, 1, ..., N-1, \\ u_j(0) = u_0(x_j), & j = 0, 1, ..., N-1, \\ u_{-1} = u_{N-1}, \ u_N = u_0. \end{cases} \tag{2}$$

The above equation is linear ODE system. Let's write it into vector form

$$U'(t) = D\,U(t), \quad U(0) = U^0, \tag{3}$$

where $U(t) = (u_0(t), u_1(t), ..., u_{N-1}(t))^T$, $U^0 = (u_0(0), u_1(0), ..., u_{N-1}(0))^T$, and D (a Toeplitz matrix) is given by

$$D = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & & & 1 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 1 & & & 1 & -2 \end{pmatrix}. \tag{4}$$

The linear ODE system can be discretized by Euler method, Runge-Kutta method or linear Multistep methods. For example, the implicit Euler method of (3) is given by (U^n is an approximation of $U(n\,\delta t)$)

$$\frac{U^{n+1} - U^n}{\delta t} = D\,U^{n+1} \quad \Longleftrightarrow \quad \left(\frac{1}{\delta t} I - D\right) U^{n+1} = \frac{1}{\delta t} U^n, \tag{5}$$

where I is a $N \times N$ identity matrix.

- *Finite element method.* In the finite element method, we approximate the the equation in the weak form. To make things simply, we do the time discretization first. By using implicit Euler method, we get

$$\frac{u^{n+1}(x) - u^n(x)}{\delta t} = u_{xx}^{n+1}(x),$$

or

$$\alpha\,u^{n+1}(x) - u_{xx}^{n+1}(x) = \alpha\,u^n(x), \quad \alpha = 1/\delta t, \ n = 0, 1, ..., M, \quad u^0(x) = u_0(x). \tag{6}$$

Let $\Omega = [0,\ 2\pi]$, $H_p^1(\Omega)$ is the Hilbert space contains all periodic functions defined on $[0, 2\pi]$ with square integrable weak derivative. The the weak form of equation (6) is:

$$\text{Find } u^{n+1} \in H_p^1(\Omega), \ s.t. \quad \alpha(u^{n+1}, v) + (u_x^{n+1}, v_x) = \alpha(\,u^n, v), \quad \forall\ v \in H_p^1(\Omega),$$

where $(u,v)=\int_\Omega u\,v\,\mathrm{d}x$ is the inner product endowed with $H_p^1(\Omega)$. The Galerkin approximation is to use a finite dimensional subspace $V_N \subseteq H_p^1(\Omega)$ to approximate $H_p^1(\Omega)$, i.e.

$$\text{Find } u_N^{n+1}\in V_N,\ s.t.\quad \alpha(u_N^{n+1},v_N)+\left(\frac{\mathrm{d}}{\mathrm{d}x}u_N^{n+1},\frac{\mathrm{d}}{\mathrm{d}x}v_N\right)=\alpha(u_N^n,v_N),\quad \forall\ v_N\in V_N. \tag{7}$$

The finite element method uses piecewise polynomial functions to approximate functions, i.e. V_N is composed of piecewise polynomials. Each grid interval is an element. The most commonly used finite element is the piecewise linear element. i.e.

$$V_N=\{\, f\in C_p^0(\Omega):\ f(x)\in P_1 \text{ for } x\in\Omega_j=[x_j,x_{j+1}], j=0,...,N-1.\,\},$$

here $C_p^0(\Omega)$ is composed of periodic continous function defined on Ω. Every function in V_N can be expressed as

$$u(x)=\sum_{j=0}^{N-1} u(x_j)\varphi_j(x), \tag{8}$$

where

$$\varphi_j(x)=\begin{cases} 1-\dfrac{|x-x_j|}{h}, & x\in[x_{j-1},x_{j+1}],\\ 0, & \text{otherwise.}\end{cases}$$

is the basis function in V_n with the property

$$\varphi_j(x_i)=\delta_{ij}:=\begin{cases}1, & i=j,\\ 0, & \text{otherwise.}\end{cases}$$

Here δ_{ij} is the Kronecker delta.
Then by substituting $u(x)$ for u_N, and substituting φ_k for v_N in equation (7), we get

$$\alpha\sum_{j=0}^{N-1} u^{n+1}(x_j)(\varphi_j,\varphi_k)+\sum_{j=0}^{N-1} u^{n+1}(x_j)(\varphi_j',\varphi_k')=\alpha\sum_{j=0}^{N-1} u^n(x_j)(\varphi_j,\varphi_k),\quad \forall\varphi_k\in V_N.$$

Let $U^n=(u^n(x_0),\ ...,\ u^n(x_{N-1}))^T$, $m_{k,j}=(\varphi_j,\ \varphi_k)$, $s_{k,j}=(\varphi_j',\ \varphi_k')$, M and S are matrix formed by taking $m_{k,j}$ and $s_{k,j}$ as components, respectively(M and S are called mass matrix and stiff matrix respectively). Then the above equation can be written into vector form

$$\alpha MU^{n+1}+SU^{n+1}=\alpha MU^n.$$

By direct calculation, we find that both M and S are Toeplitz matrix, and their components are given by

$$m_{k,j}=\begin{cases}\dfrac{2}{3}h, & k=j,\\ \dfrac{1}{6}h, & k=j\pm1,\\ 0, & \text{otherwise,}\end{cases}\qquad s_{k,j}=\begin{cases}\dfrac{2}{h}, & k=j,\\ -\dfrac{1}{h}, & k=j\pm1,\\ 0, & \text{otherwise.}\end{cases}$$

- *Fourier spectral method.* We use Fourier series to approximate the solution of (1).

$$u(x,t)\approx u_N(x,t)=\sum_{j=0}^{N-1}\hat{u}_j(t)\,e^{ijx}. \tag{9}$$

By expanding u in equation (1) by (9) and taking the inner product of the resulting equation with test function e^{ikx}, we get

$$\sum_{j=0}^{N-1}\hat{u}_j'(t)\,\delta_{kj}\,2\pi=\sum_{j=0}^{N-1}\hat{u}_j(t)(-kj)2\pi\,\delta_{kj},\quad \forall k=0,...,N-1.$$

or

$$\hat{u}_k'(t) = -k^2 \hat{u}_k(t), \quad \forall k = 0, ..., N-1.$$

This is a decoupled linear ODE system. Its solution is given by

$$\hat{u}_k(t) = \hat{u}_k(0)\, e^{-k^2 t},$$

where

$$\hat{u}_k(0) = \frac{(u_0(x), e^{ikx})}{(e^{ikx}, e^{ikx})} = \frac{1}{2\pi}\int_0^{2\pi} u_0(x) e^{-ikx}\, \mathrm{d}x.$$

The solution at time t is given by

$$u_N(x,t) = \sum_{k=0}^{N-1} \hat{u}_k(0) e^{-k^2 t}\, e^{ikx}.$$

The solution error is

$$\begin{aligned} |u(x,t) - u_N(x,t)| &= \sum_{k=N}^{\infty} \left|\hat{u}_k(0) e^{-k^2 t}\, e^{ikx}\right| \\ &\leqslant \max_{k \geqslant N} |\hat{u}_k(0)| \sum_{k=N}^{\infty} e^{-k^2 t} \\ &< \max_{k \geqslant N} |\hat{u}_k(0)| \int_N^{\infty} e^{-x^2 t}\, \mathrm{d}x \\ &= \max_{k \geqslant N} |\hat{u}_k(0)| \sqrt{\frac{\pi}{4t}} \mathrm{erfc}\big(\sqrt{t} N\big) \end{aligned}$$

Here erfc is the complementary error function defined as

$$\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-y^2} \mathrm{d}y.$$

By using the property that $\mathrm{erfc}(x) \sim \frac{e^{-x^2}}{\sqrt{\pi} x}$ for $x \gg 1$, we get

$$|u(x,t) - u_N(x,t)| \leqslant \max_{k \geqslant N} |\hat{u}_k(0)| \times \frac{e^{-N^2 t}}{2Nt}.$$

Homework: Write a matlab program to verify the convergence rates of three different numerical schemes discussed above.

2 Weighted Residual Methods(WRM)

Consider following general problem

$$\boldsymbol{L}u(x) := \alpha\, u - \mathcal{L}u = f, \quad x \in \Omega, \tag{10}$$

where α is a given coefficient function, $\mathcal{L}$ is a linear elliptic operator. The starting point of the WRM is to approximate the solution u of (10) by a finite sum

$$u(x) \approx u_N(x) = \sum_{k=0}^{N} a_k\, \phi_k(x), \tag{11}$$

where $\{\phi_k\}$ are the trial (or basis) functions, and the expansion coefficients $\{a_k\}$ are to be determined. Substituting u_N for u in (10) leads to the residual:

$$R_N(x) = \boldsymbol{L}u_N - f(x), \quad x \in \Omega.$$

The notion of the WRM is to force the residual to zero by requiring

$$(R_N, \psi_j)_\omega := \int_\Omega R_N(x) \psi_j(x) \omega(x)\, \mathrm{d}x = 0, \quad 0 \leqslant j \leqslant N. \tag{12}$$

The choice of trial/test functions is one of the main features that distinguishes spectral methods from finite-element and finite-difference methods. In the latter two methods, the trial/test functions are local in character with finite regularities. In constrast, spectral methods emply globally smooth functions as trial/test functions. The most commonly used trial/test functions are trigonometric functions or orthogonal polynomials (typically, the eigenfunctions of singular Sturm-Liouville problems), which include

- $\phi_k(x) = e^{ikx}$ Fourier spectral method, $\omega(x) = 1, \Omega = [0, 2\pi]$
- $\phi_k(x) = T_k(x)$ Chebyshev spectral method, $\omega(x) = 1/\sqrt{1-x^2}$, $\Omega = [-1, 1]$
- $\phi_k(x) = L_k(x)$ Legendre spectral method, $\omega(x) = 1$, $\Omega = [-1, 1]$
- $\phi_k(x) = \mathcal{L}_k(x)$ Laguerre spectral method, $\omega(x) = e^{-x}$, $\Omega = [0, +\infty)$
- $\phi_k(x) = H_k(x)$ Hermite spectral method, $\omega(x) = e^{-x^2}$, $\Omega = [-\infty, +\infty]$

Here, T_k, L_k, $\mathcal{L}_k$ and H_k are the Chebyshev, Legendre, Laguerre and Hermite polynomials of degree k, respectrively.
The choice of test functions distinguishes the following formulations:

- *Galerkin*. The test functions are the same as the trial ones (i.e. $\phi_k = \psi_k$).
- *Petrov-Galerkin*. The test functions are different from the trial ones.
- *Collocation*. The test functions $\{\psi_k\}$ are delta functions, i.e. $\psi_k(x_j) = \delta_{jk}$ and the trial basis function are Lagrange polynomials which also satisfies $\phi_k(x_j) = \delta_{jk}$.

In the Galerkin method, there is a modified version called *Galerkin with Numerical Integration* (G-NI), in which the integration in (12) is replaced by a numerical integration, This is a compromise for nonlinear problem or linear problem with variable coefficients, in which the exact integration is hard to calculate.
Usually, the classical spectral method only refer to the Galerkin spectral method. Spectral methods involves collocation or numerical integration are sometimes called *pseudo-spectral* methods in the literature.

3 Basis of Error Analysis

The error analysis of spectral method usually use variational method, i.e. weak form of the PDE. For equation (10) with certain boundary condition usually can be reformulated into a weak form

$$\text{Find } u \in X \text{ such that } \quad a(u, v) = F(v), \quad \forall v \in Y. \tag{13}$$

where X is the space of trial functions, Y is the space of test functions, and F is a linear functional on Y. The expression $a(u, v)$ defines a bilinear form on $X \times Y$. It is conventional to assume that X and Y are Hilbert spaces.
There are well-established theory for equation (13) and related Galerkin/Petrov-Galerkin approximations[see Shen, Tang, and Wang's Spectral Method book, chapter 1.4]. Here we only discuss the symmetric one where $X = Y$.

Theorem 1. *(Lax-Milgram Lemma) Consider the problem*

$$\begin{cases} \text{Find } u \in X \text{ such that} \\ a(u, v) = F(v), \quad \forall v \in X, \end{cases} \tag{14}$$

where X is a Hilbert space, F: $X \to \mathbb{R}$ is a linear functional in X'. $a(\cdot, \cdot)$ is a bilinear form satisfies following continuous and coercive properties:

$$|a(u, v)| \leqslant C \|u\|_X \|v\|_X, \quad \forall\, u, v \in X, \tag{15}$$

$$a(u, v) \geqslant \alpha \|u\|_X^2, \quad \forall\, v \in X. \tag{16}$$

Then the variational problem (14) has a unique solution. Moreover, we have

$$\|u\| \leqslant \frac{1}{\alpha}\|F\|_{X'}.$$

Using the Lax-Milgram Theorem, it is not hard to deduce following theorem, which states that the Galerkin solution error is bounded by the best approximation error.

Theorem 2. *(Cea Lemma) Assume that* $X_N \subseteq X$ *and*

$$\forall\ v \in X, \quad \inf_{v_N \in X_N} \|v - v_N\|_X \to 0 \text{ as } N \to \infty, \tag{17}$$

then the following Galerkin approximation to (14)

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a(u_N, v_N) = F(v_N), \quad \forall\ v_N \in X_N \end{cases} \tag{18}$$

admits a unique solution $u_N \in X_N$ *such that*

$$\|u\|_X \leqslant \frac{1}{\alpha}\|F\|_{X'}. \tag{19}$$

Moreover, if u *is the solution of (14), we have*

$$\|u - u_N\|_X \leqslant \frac{C}{\alpha} \inf_{v_N \in X_N} \|u - v_N\|_X. \tag{20}$$

Proof. Since X_N is a subspace of X, applying the Lax-Milgram lemma to (18) leads to the existence and uniqueness of u_N and the stability result (19). Now take $v = v_N$ in equation (14), and substracting (18) from the resulting equation, we obtain the error equation

$$a(u - u_N, v_N) = 0, \quad \forall\ v_N \in X_N,$$

with together with equation (15) and (16), implies

$$\begin{aligned} \alpha\|u - u_N\|_X^2 &\leqslant a(u - u_N, u - u_N) = a(u - u_N, u - v_N) \\ &\leqslant C\|u - u_N\|_X \|u - v_N\|_X, \quad \forall v_N \in X_N, \end{aligned}$$

from which (20) follows.

□

Galerkin method with numerical integration.

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that} \\ a_N(u_N, v_N) = F_N(v_N), \quad \forall v_N \in X_N, \end{cases} \tag{21}$$

with X_N satisfies (17), and $a_N(\cdot,\cdot)$ and $F_N(\cdot)$ are suitable approximations to $a(\cdot,\cdot)$ and $F(\cdot)$, respectively.

Theorem 3. *(First Strang Lemma). Under the assumptions of the Lax-Milgram lemma, suppose further that the discrete forms* $F_N(\cdot)$ *and* $a_N(\cdot,\cdot)$ *satisfy the same properties in the subspace* $X_N \subset X$, *and* $\exists \alpha_* > 0$, *independent of* N, *such that*

$$a_N(v, v) \geqslant \alpha_* \|v\|_X^2, \quad \forall v \in X_N. \tag{22}$$

then, the problem (21) admits a unique solution $u_N \in X_N$, *satisfying*

$$\|u_N\|_X \leqslant \frac{1}{\alpha_*} \sup_{0 \neq v_N \in X_N} \frac{|F_N(v_N)|}{\|v_N\|_X}. \tag{23}$$

moreover, if u is the solution of (14), we have

$$\|u-u_N\|_X \leqslant \inf_{w_N\in X_N}\left\{\left(1+\frac{C}{\alpha_*}\right)\|u-w_N\|_X+\frac{1}{\alpha_*}\sup_{0\neq v_N\in X_N}\frac{|a(w_N,v_N)-a_N(w_N,v_N)|}{\|v_N\|_X}\right\} \\ +\frac{1}{\alpha_*}\sup_{0\neq v_N\in X_N}\frac{|F(v_N)-F_N(v_N)|}{\|v_N\|_X}. \tag{24}$$

Here, the constant C is given in (15).

Proof. The existence-unique and stability of (23) follow from the Lax-Milgram lemma. The proof of (24) is slightly different from that of (20). For any $w_N \in X_N$, let $e_N = u_N - w_N$. Using (22), (14) and (21) leads to

$$\alpha^*\|e_N\|_X^2 \leqslant a_N(e_N,e_N) = a(u-w_N,e_N)+a(w_N,e_N)-a_N(w_N,e_N)+F_N(e_N)-F(e_N).$$

Since the result is trivial for $e_N = 0$, we derive from (15) that for $e_N \neq 0$,

$$\begin{aligned}\alpha^*\|e_N\|_X &\leqslant C\|u-w_N\|_X+\frac{|a(w_N,e_N)-a_N(w_N,e_N)|}{\|e_N\|_X}+\frac{|F(e_N)-F_N(e_N)|}{\|e_N\|_X}\\ &\leqslant C\|u-w_N\|_X+\sup_{0\neq v_N\in X_N}\frac{|a(w_N,v_N)-a_N(w_N,v_N)|}{\|v_N\|_X}+\sup_{0\neq v_N\in X_N}\frac{|F(v_N)-F_N(v_N)|}{\|v_N\|_X},\end{aligned}$$

which, together with the triangle inequality, yields

$$\|u-u_N\|_X \leqslant \|u-w_N\|_X+\|e_N\|_X.$$

Finally, taking the infimum over $w_N \in X_N$ leads to the desired result. □

4 Summary

	FD and/or FD	Spectral method
Basis	local	global
convergence rate	low order	high order
to increase accuracy	refine grid	increase number of bases
computational efficiency	sparse system	sparse for particular PDEs only