The 2D Heat Equation

— Analysis and Numerical Approaches

Jiajun Wang and Jiongyi Wang

Last Updated on: May 23, 2022

Contents

1 Formulation and Primary Exploration of Problem

1.1 Physical Interpretation

We want to study how heat is conducted in a metal rod of length L over time.One end of the metal rod is at and the other end is at. The length of the metal rod is much greater than its cross-sectional radius, so we can think of heat conduction as a function of x and t.

Assuming the specific heat capacity of the metal rod is known, if we can find a function of temperature, we can know how the heat diffuses.

The rod is assumed to be adiabatic along its length, so it can only absorb or dissipate heat through the ends. This means that the temperature distribution depends only on the following three factors:

$\boxed{1.}$ Initial temperature distribution, $T(x,0)$this is called initial condition.

$\boxed{2.}$ The temperature at both ends of the metal rod $T(0,t)$, $T(L,t)$is called boundary conditions.

$\boxed{3.}$ The law of heat transfer from one point to another in the metal rod. The heat equation is a mathematical representation of this physical law.

We assume that the initial boundary value of the solved heat equation $T(0,t) = T(L,t) = 0$

The problem of solving partial differential equations for a specific set of initial and boundary conditions is called an initial boundary value problem.

The heat equation can be derived from the conservation of energy: the time rate of change of the heat stored at a point on the metal rod is equal to the net heat flux into that point. This process fits the continuity equation. If Q is the heat at various points and V is the vector field of heat flow, then:

$$\frac{\partial Q}{\partial t} + \nabla V = 0$$

According to the second law of thermodynamics, if two identical bodies are in thermal contact, one being hotter than the other, then heat must flow from the hotter body to the cooler body at a rate proportional to the temperature difference. Therefore, V is proportional to the negative gradient of temperature, so: $V = -kT$ where k is the thermal conductivity of the metal. In one dimension, $Q = \rho c T$, k, ρ, and c are the thermal conductivity, density, and specific heat capacity of the metal, respectively. Substituting into the expressions for V and Q yields the heat equation:

$$\frac{\partial T}{\partial t} - \frac{k}{\rho c}\frac{\partial^2 T}{\partial x^2} = 0$$

Now we give a formal formulation for the heat equation:

Let $\Omega \subset \mathbb{R}^d$ be an open set with boundry $\Gamma := \partial\Omega$, set $\Omega_T = \Omega \times]0,T[$, $\Gamma_T := \Gamma \times]0,T[$, Γ_T is called the *lateral* boundary of the cylinder Ω_T.

Consider the heat equation with L-periodic boundary condition:

$$\left\{\begin{array}{rll} \partial_t u - k\Delta u & = f(x,t) & \text{in } \Omega_T \\ u(x,t) & = u(x+L,t) & \text{on } \Gamma_T \\ u(x,0) & = u_0(x) & \text{on } \Omega \times \{t=0\} \end{array}\right. \tag{1.1}$$

N.B. where $k > 0$ is a ”diffusion coefficient”. However, since the constant can be scaled out by defining a rescaled time $\tau = t/k$ to get

$$u_\tau - \Delta u = f$$

Hence we could simplify the formulation as

$$\begin{cases} \partial_t u - \Delta u &= f(x,t) & \text{in } \Omega_T \\ u(x,t) &= u(x+L,t) & \text{on } \Gamma_T \\ u(x,0) &= u_0(x) & \text{on } \Omega \times \{t=0\} \end{cases} \tag{1.2}$$

In high dimensional case, $\Delta u = \sum_1^d \frac{\partial^2 u}{\partial x_i^2} = \operatorname{div}(\operatorname{grad} u)$. Finally, in PDE context, we often note $\frac{\partial u}{\partial t}$ as u_t and $u_{xx} := \frac{\partial^2 u}{\partial x^2} = \Delta u$ if no ambiguity occurs.

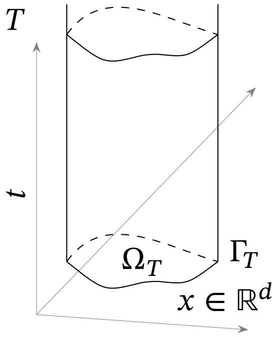


Figure 1: Region Ω_T

The Modeling Context: An Excursion to IC and BCs

$\boxed{1.}$ **Dirichlet**: The temperature is fixed at the boundary

$\boxed{2.}$ **Neumann**: The end is insulated (no heat enters or escapes).

$\boxed{3.}$ **Robin**: Some heat enters or escapes, with an amount proportional to the temperature:

$$\alpha u = -\beta \frac{\partial u}{\partial n}$$

where $\frac{\partial u}{\partial n} := \nabla u \cdot \vec{n}$. For the area Ω whether heat enters or escapes the system depends on the boundary $\Gamma = \partial\Omega$. The heat flux $-\beta\frac{\partial u}{\partial n}$ is to the right if it is positive, so at the point a belongs to boundary , heat enters the system when $\alpha > 0$ and leaves when $\alpha < 0$.

$\boxed{4.}$ **Periodic**: The temperature is periodic at the boundary.

The same interpretations apply when the equation is describing diffusion of some other quantity . Non-homogeneous boundary conditions can be imposed, for instance

$$u(z,t) = g(z,t), \quad z \in \Gamma_T$$

which might be used to model the ambient temperature increasing with time.

Remark 1 (Robin's BC)**:** All BCs could be expressed as Robin's condition

$$\alpha u(a) + \beta \frac{\partial u}{\partial n}(a) = g, \quad a \in \partial\Omega$$

Dirichlet's condition means that $\beta = 0$ and Neumann's condition means that $\alpha = 0$.

Remark 2: At each point of Γ, the BCs could be different.

1.2 Fundamental Solution for Heat Equation

We observe that the heat equation involves one derivative with respect to the time variable t ,and two derivatives with respect to the space variable $x = (x_1, \ldots, x_d)$. At the same time ,we can observe that if u is a solution of the heat equation , $u(\lambda x, \lambda^2 t)$ for $\lambda \in \mathbb{R}$ is also a solution. This scaling indicates the ratio $\frac{|r|^2}{t}$ $(r := x)$ is very important for the heat equation. We need to find a solution as $u(x,t) = v(\frac{|x|^2}{t})(t > 0, x \in \mathbb{R})$.

But here we want an idea for which we can more easily find a solution [cf. Eva10, p.45] $u(x,t) = \frac{1}{t^a} v(\frac{x}{t^b})$ and we have x in $\mathbb{R}, t > 0$.

The equation 1.3 is easily obtained by a simple calculation.

$$\lambda^a u(\lambda^b, \lambda t) = \frac{\lambda^a}{(\lambda t)^a} v(\frac{\lambda^b x}{(\lambda t)^b}) = \frac{1}{t^a} v(\frac{x}{t^b}) = u(x,t) \tag{1.3}$$

Let $\lambda = \frac{1}{t}$ and $y = \lambda^b x$, we can get $v(y) = u(y, 1)$. We substitute $u(x,t)$ into the first expression of the heat equation and get

$$a t^{-a-1} v(y) + b t^{-a-1} y \cdot Dv(y) + t^{-a-2b} \Delta v(y) = 0$$

To simplify the equation, we assume $v(y) = w(|y|)$, and $w : \mathbb{R} \to \mathbb{R}$. Then the above formula becomes:

$$a t^{-a-1} v(y) + b t^{-a-1} y \cdot Dv(y) + t^{-a-2b} \Delta v(y) = aw + \frac{1}{2} r w' + w'' + \frac{d-1}{r} w' = 0$$

if we let a= $\frac{d}{2}$,we will get the following equation:

$$\frac{d}{2} w + \frac{1}{2} r w' + w'' + \frac{d-1}{r} w' = (r^{d-1} w' + \frac{1}{2} r^d w)' = 0$$

which implies that $r^{d-1} w' + \frac{1}{2} r^d w = C$ and we assume that $\lim_{r \to 0} w = 0, \lim_{r \to 0} w' = 0$, and we can derive $C = 0$ and $w' = -\frac{1}{2} r w$. Thus, we have $w = \beta e^{-\frac{r^2}{4t}}$ for some constant β .

Finally, with $u(x,t) = \frac{1}{t^{\frac{d}{2}}} v(\frac{x}{t^{\frac{1}{2}}})$, we deduce $u(x,t) = \frac{\beta}{t^{\frac{d}{2}}} e^{-\frac{|x|^2}{4t}}$

So we define the solution of the heat equation as

✠ **Definition 1.1** (fundamental solution)**:** *The function*

$$\Phi(x,t) = \begin{cases} \frac{1}{(4\pi t)^{d/2}} e^{-\frac{|x|^2}{4t}} & x \in \mathbb{R}^d, t > 0 \\ 0 & x \in \mathbb{R}^d, t < 0 \end{cases} \tag{1.4}$$

is called the fundamental solution of the heat equation.

The choice of the normalizing constant $(4\pi)^{-\frac{d}{2}}$ is dictated by

$$\begin{aligned} \int_{\mathbb{R}^d} \Phi(x,t) dx &= \frac{1}{(4\pi t)^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{-\frac{|x|^2}{4t}} dx \\ &= \frac{1}{\pi^{\frac{d}{2}}} \int_{\mathbb{R}^d} e^{-|z|^2} dz = \frac{1}{\pi^{\frac{d}{2}}} \prod_{i=1}^{d} \int_{-\infty}^{+\infty} e^{-z_i^2} dz_i \\ &= 1 \end{aligned}$$

To proceed a verification of the correctness of this result, we need to calculate $\frac{\partial \Phi(x,t)}{\partial t}$ and $\frac{\partial^2 \Phi(x,t)}{\partial x_i^2}$ separately, and then substitute the results into the heat equation for verification. The result follows:

$$\begin{aligned} \frac{\partial \Phi(x,t)}{\partial t} &= \frac{1}{(4\pi)^{\frac{d}{2}}}((-\frac{d}{2})(t^{-\frac{d}{2}-1})e^{-\frac{|x|^2}{4t}} + \frac{1}{t^{\frac{d}{2}}} e^{-\frac{|x|^2}{4t}} \frac{|x|^2}{4t}) \\ &= \frac{e^{-\frac{|x|^2}{4t}}}{(4\pi)^{\frac{d}{2}}}((-\frac{d}{2})\frac{1}{t^{\frac{d}{2}+1}} + \frac{|x|^2}{4t^{\frac{d}{2}+2}}) \end{aligned} \tag{1.5a}$$

$$\frac{\partial \Phi(x,t)}{\partial x_i} = \frac{e^{-\frac{|x|^2}{4t}}}{(4\pi t)^{\frac{d}{2}}}(-\frac{1}{2t}x_i) \tag{1.5b}$$

$$\frac{\partial^2 \Phi(x,t)}{\partial x_i^2} = \frac{e^{-\frac{|x|^2}{4t}}}{(4\pi t)^{\frac{d}{2}}}(\frac{1}{4t^2}x_i^2) - \frac{1}{2t}\frac{e^{-\frac{|x|^2}{4t}}}{(4\pi t)^{\frac{d}{2}}} \tag{1.5c}$$

$$\begin{aligned} \sum_{i=0}^{d} \frac{\partial^2 \Phi(x,t)}{\partial x_i^2} &= \frac{e^{-\frac{|x|^2}{4t}}}{(4\pi t)^{\frac{d}{2}}}(\frac{1}{4t^2}|x|^2) - \frac{d}{2t}\frac{e^{-\frac{|x|^2}{4t}}}{(4\pi t)^{\frac{d}{2}}} \\ &= \frac{e^{-\frac{|x|^2}{4t}}}{(4\pi)^{\frac{d}{2}}}((-\frac{d}{2})\frac{1}{t^{\frac{d}{2}+1}} + \frac{|x|^2}{4t^{\frac{d}{2}+2}}) \end{aligned} \tag{1.5d}$$

By substituting them into the heat equation, we can find that

$$\frac{\partial u}{\partial t}(x,t) + \frac{\partial^2 u}{\partial x^2}(x,t) = 0$$

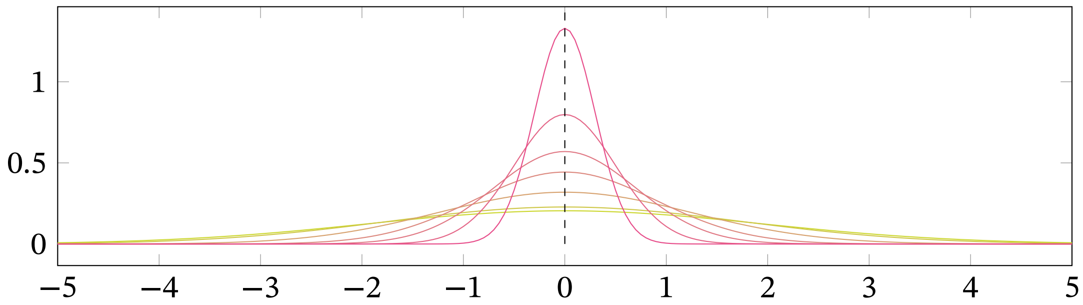


Figure 2: evolution of temperature on a rod along the time

1.3 Spectrum Theory and Spectral Analysis

In this part, for analysis preliminaries, [Fol99, ch.05-06] offers a panorama on the theory of L^p spaces.

✠ **Definition 1.2** (compact operator): *Given E, F two normed vector spaces, an operator $T \in \mathcal{L}(E;F)$ is said to be copmact if the image of unit ball in E under T, i.e. $T(B_E)$ is relatively compact (in the sense of closure) in F.*

We call an operator T is of finite rank, if $\dim \operatorname{Im} T < \infty$.

1.3.1 Riesz-Fredholm theory

▶ **Lemma 1.3** (Riesz): *Let E be a normed vector space (not necessary complete), $M \subsetneq E$ a proper closed linear subspace, then $\forall \varepsilon > 0$, $\exists u \in E$ s.t. $\|u\| = 1$ and $d(u, M) \geq 1 - \varepsilon$.*

▶ **Theorem 1.4:** *Let E be a normed vector space with compact unit ball B_E, then E is finite-dimensional.*

▶ **Theorem 1.5** (Fredholm alternative): *Let $T \in \mathcal{L}(E)$ be a compact operator, then*

- $\operatorname{Ker}(I - T)$ *is finite-dimensional.*
- $\operatorname{Im}(I - T)$ *is closed. More precisely,* $\operatorname{Im}(I - T) = \operatorname{Ker}(I - T')^{\perp}$
- $\operatorname{Ker}(I - T) = \{0\} \iff \operatorname{Im}(I - T) = E$
- $\dim \operatorname{Ker}(I - T) = \dim \operatorname{Ker}(I - T')$

▷ *Proof:* Admitted. [cf. Bre11, p.160-162] □

✠ **Definition 1.6** (resolvent set, spectrum and eigenvalue): *Let $T \in \mathcal{L}(E)$, the resolvent set, denoted by $\rho(T)$, is defined by*

$$\rho(T) := \{\lambda \in \mathbb{C}; (T - \lambda I) : E \to E \textit{ is bijective }\}$$

The spectrum, denoted by $\operatorname{Spec}(T)$, *is the complement of the resolvent set, i.e.,* $\operatorname{Spec}(T) = \mathbb{C} \backslash \rho(T)$. *A complex number λ is said to be an eigenvalue of T if* $\operatorname{Ker}(T - \lambda I) \neq \{0\}$. *The set of eigenvalues of T is denoted by $EV(T)$. The space* $\operatorname{Ker}(T - \lambda I)$ *is called the eigenspace of T, the elemet in it is called eigenvector.*

Remark 3: If E is a Banach space, the open mapping theorem tells us, the bijectivity of T equals that $T^{-1} \in \mathcal{L}^{-1}(E)$. Actually we have following consequnce :

▶ **Proposition 1.7:**

- *If $T \in \mathcal{L}(E)$ and $\|I\text{-}T\| < 1$ where I is the identity operator, then T is invertible, the series* $\lim_{n\to\infty}\sum_{n=0}^{\infty}(I-T)^n = T^{-1}$ *in $\mathcal{L}(E)$.*
- *The set of invertible operators in $\mathcal{L}(E)$, denoted as $GL(E)$, is an open set in $\mathcal{L}(E)$, and $GL(E) \to GL(E); T \mapsto T^{-1}$ is continuous. More precisely, if $S \in GL(E)$ and $\|T\text{-}S\| < \left\|T^{\text{-}1}\right\|^{-1}$, then $S \in GL(E)$.*

▶ **Theorem 1.8** (Gelfand)**:**

- *We have $\|T^n\|^{1/n} \xrightarrow{n\to\infty} \inf_n \|T^n\|^{1/n}$, we call this limite, denoted by $r(T)$, the spectral radius of T. Moreover, $r(T) \leq \|T\|$, and $\forall\lambda \in \mathrm{Spec}(T), |\lambda| \leq r(T)$. In particular, $\mathrm{Spec}(T)$ is a compact set in $\mathbb{C}$.*
- *For all $T \in \mathcal{L}(E)$, we have $\mathrm{Spec}(T) \neq \varnothing$. Moreover*

$$r(T) = \max_{\lambda\in\mathrm{Spec}(T)} \{|\lambda|\}$$

▷ *Proof:* [cf. Lax02, p195-197] □

✠ **Definition 1.9** (adjoint, self-adjoint)**:** *Let $A : \mathrm{Dom}(A) \subset E \to F$ be an unbounded linear operator that is densely defined. We shall introduce an unbounded operator $A' : \mathrm{Dom}(A') \subset F' \to E'$ as follows:*

$$\mathrm{Dom}(A') := \{v \in F' : \exists c \geq 0 \; s.t. \; |\langle v, Au\rangle| \geq c\,\|u\|, \quad \forall u \in \mathrm{Dom}(A)\}$$

$${}_{F'}\langle v, Au\rangle_F = {}_{E'}\langle A'v, u\rangle_E, \quad \forall u \in \mathrm{Dom}(A), \forall v \in \mathrm{Dom}(A')$$

A bounded operator T is said to be self-adjoint if $T' = T$.

▶ **Theorem 1.10:** *Suppose that H is a separable Hilbert space, T is a compact self-adjoint operator, then there exists a Hilbert basis composed of eigenvectors of T.*

Our last statement is a fundamental result. It asserts that every compact self-adjoint operator may be diagonalized in some suitable basis.

1.3.2 Eigenfunctions and spectral decomposition

Now, we have sufficient tools to proceed the spectral analysis of heat equation. (More generally, the spectral analysis could be applied to other types of PDE [cf. Bre11, ch.08-09; Lax02, ch.33-36])

✠ **Definition 1.11** (Sobolev spaces, distribution)**:**

$$W^{m,p}(\Omega) := \left\{u \in L^p(\Omega) : \partial^\alpha u \in L^p(\Omega), \forall \alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}_+^d \text{ and } 1 \leq |\alpha| \leq m\right\}$$

For index $\alpha \in \mathbb{R}_+^d$, we note $|\alpha| = \sum_1^d \alpha_i$. The norm $\|\cdot\|_{W^{m,p}(\Omega)}$ defined by

$$\|u\|_{W^{m,p}(\Omega)} = \left(\sum_{|\alpha|=0}^{m} \|\partial^\alpha u\|_{L^p(\Omega)}^p\right)^{\frac{1}{p}}$$

makes the Sobolev space $W^{m,p}(\Omega)$ complete.

We simply note $H^m(\Omega) := W^{m,2}(\Omega)$, since it's a Hilbert space.

At last, we define $H_0^m(\Omega) := \overline{\mathcal{D}(\Omega)}^{H^m(\Omega)}$, that means the closure of $\mathcal{D}(\Omega)$ in $H^m(\Omega)$.

Where $\mathcal{D}(\Omega) = C_c^\infty(\Omega)$ called the set of test functions. Its dual space $\mathcal{D}'(\Omega)$ is called distribution.

$H_0^1(\Omega)$ need not to inherit the norm from $H^1(\Omega)$, there is an equivalent norm inducted by the inner product:

$$\langle v, u\rangle_{H_0^1(\Omega)} = \langle \nabla u, \nabla v\rangle_{L^2(\Omega)}$$

The notion of distriburion generalized the notion of function. We could find that the Dirac mass at $x = a$: $\delta(x-a)$ is not a function, however, it's a distriburion. Important example: $L^1_{\text{loc}}(\Omega)$ is a distribution. For more details, [cf. Gos20, ch.03-05].

▶ **Theorem 1.12** (the spectrum of Laplacian operator)**:** *Suppose $T : L^2(\Omega) \to L^2(\Omega); f \mapsto u$, where u is the weak solution (i.e. solution in $H_0^1(\Omega)$) of*

$$\begin{cases} -\Delta u_f &= f \qquad \text{in } \Omega \\ \quad u_f &= 0 \qquad \text{on } \partial\Omega \end{cases} \tag{1.6}$$

Then T is compact and self-adjoint. Moreover T is positive defined.

▷ *Proof:* u_f is characterized by $u_f \in H_0^1(\Omega), \forall \varphi \in H_0^1(\Omega), \int_\Omega \nabla u_f \nabla \varphi = \int_\Omega f\varphi$ and $S : f \mapsto u_f, L^2(\Omega) \to H_0^1(\Omega)$ is linear continuous.

$$\begin{tikzcd} f \arrow[r, "T"] \arrow[dr, "S"'] & u \\ & u_f \arrow[u, "i"'] \end{tikzcd}$$

$T = i \circ S$, where S is linear continuous and $i : H_0^1(\Omega) \to L^2(\Omega)$ is a canonical injection which is compact (it's a result of Reillich-Kondrachov's Theorem, [cf. Bre11, ch.9.3, p.285], we don't discuss the interpolation of Sobolev space here)

To prove that T self-adjoint and positive is relatively easy, it's direct consequnce of properties of $L^2(\Omega)$. □

Thus, by theorem 1.10, there exists a Hilbert basis in $L^2(\Omega)$ consists of the eigenvalues of T: assume that $\Omega \subset \mathbb{R}^d$ is a bounded open set, then there exist a Hilbert basis $\{e_n\}_n$ of $L^2(\Omega)$ s.t. $e_n \in H_0^1(\Omega) \cap C^\infty(\Omega), \forall n$, $Te_n = \lambda_n e_n$ and a sequence $\{\lambda_n\}_n$ of real numbers with $\|T\| \geq \lambda_n > 0, \forall n$ and $\lambda_n \to 0$

Now return to Laplacian operator. For all $n \geq 1$, we define $\mu_n := \frac{1}{\lambda_n}$ s.t.

$$-\Delta e_n = \mu_n e_n \quad \text{in } \Omega$$

The sequence $\{\mu_n\}$ is increasing to $+\infty$ and is called the eigenvalues of the$-\Delta$ (with Dirichlet boundary condition), the $\{e_n\}_n$ are the associated eigenfunctions. We say that $\{\lambda_n\}_n$ are the eigenvalues of Δ.

▶ **Corollary 1.13:** *Let $\{e_n\}$ be the eigenfunctions of Laplacian, $\left\{\sqrt{\lambda_n} e_n\right\}_n$ is a Hilbert basis for $H_0^1(\Omega)$ equipped with the inner product*

$$\langle v, w\rangle_{H_0^1(\Omega)} = \int_\Omega \nabla v \nabla w$$

.

▷ *Proof:*

$$\begin{aligned}\left\langle \nabla\sqrt{\lambda_n}e_n, \nabla\sqrt{\lambda_m}e_m\right\rangle_{L^2(\Omega)} &= \frac{1}{\lambda_n}\left\langle\sqrt{\lambda_n}\nabla Te_n, \nabla\sqrt{\lambda_m}e_m\right\rangle_{L^2(\Omega)} \\ &= \frac{1}{\sqrt{\lambda_n}}\sqrt{\lambda_n\lambda_m}\int_\Omega e_n e_m \\ &= \sqrt{\frac{\lambda_m}{\lambda_n}}\langle e_n, e_m\rangle_{L^2(\Omega)} = \sqrt{\frac{\lambda_m}{\lambda_n}}\delta_n^m\end{aligned}$$

Then $\left\langle \nabla\sqrt{\lambda_n}e_n, \nabla\sqrt{\lambda_m}e_m\right\rangle = \sqrt{\frac{\lambda_m}{\lambda_n}}\delta_n^m \implies \left\{\sqrt{\lambda_n}e_n\right\}_n$ is an orthonomal family in $H_0^1(\Omega)$.

Lastly, we should prove $\left\{\sqrt{\lambda_n}e_n\right\}_n$ is a maximal family, i.e. it spans a dense party of $H_0^1(\Omega)$.

Let $v \in H_0^1(\Omega)$ s.t. $\forall n, \left\langle\sqrt{\lambda_n}\nabla e_n, \nabla v\right\rangle_{L^2(\Omega)} = 0 \implies \frac{1}{\sqrt{\lambda_n}}\langle\nabla Te_n, \nabla v\rangle_{L^2(\Omega)} = 0 \implies \frac{1}{\sqrt{\lambda_n}}\langle e_n, v\rangle_{L^2(\Omega)} = 0$.

Since $\{e_n\}_n$ is a Hilbert basis of $L^2(\Omega)$, then $v = 0$ in $L^2(\Omega)$, naturally $v = 0$ in $H_0^1(\Omega)$.

We have hereby proved that $\left(\overline{\text{span}(\{e_n\})}^{H_0^1(\Omega)}\right)^{\perp} = \{0\}$ in $H_0^1(\Omega)$, then $\left\{\sqrt{\lambda_n}e_n\right\}_n$ is a maximal family.

□

However, our problem is more complicated than the case above, that's because the restriction of initial condition (IC) and boundary conditions (BCs) may not allow the solution belong to $H_0^m(\Omega)$, which implies only a zero Dirichlet BC. Thus, the choice of a proper working space is the first and the crucial matter to think about.

Separation of Variables and Superpostition Principle: A History

When Joseph Fourier contemplate the heat equation

As we did before, sometimes we note u_{xx} as Δu, because the Laplacian operator is a linear operator on u with respect to x — actually, the heat equat also belongs to linear PDE, which refers that

$$u_t = -L[u] + f(x,t) \tag{1.7}$$

A linear, **homogeneous** PDE obeys the superposition principle: u_1, u_2 are solutions $\implies$ $c_1u_1 + c_2u_2$ is a solution for all scalars $c_1, c_2 \in \mathbb{R}$. The concepts of linearity and homogeneity also apply to boundary conditions, in which case the variables are evaluated at specific points.

For example, here lists the linear operators for some basic linear PDEs, i.e. heat equation, wave equation, Poisson's equation.

$$L_h = \partial_t - \partial_{xx}, \quad L_w = \partial_{tt} - \partial_{xx}, \quad L_P = \nabla^2$$

an example of non-linear PDE:

$$u_t + uu_x = u_{xx}$$

However, L_h is generally not a compact and self-adjoint operator, which means we should consider rather the Laplacian operator Δ and the basis is a series of functions involves only x.

To have an eigenfunction of the operator L, we must prescribe Ω and associated boundary conditions. Generally, we could sketch out the proper working space V as

$$V = \{f \in H^m(\Omega) : f \text{ satisfies BCs}\} \tag{1.8}$$

The eigenfunctions we need is to solve the eigenvalue problem

$$L[\phi] = \lambda\phi, \quad \phi \in V$$

By theorem 1.10, there is a sequence of eigenfunctions $\{\phi_n\}$ with eigenvalues $\{\lambda_n\}$ that form an orthogonal basis for the space V.

Now at each fixed time t, the function $u(x,t)$, regarded as a function of x, lies inside the space V . It follows that there are coefficients $a_n(t)$ such that

$$u(x,t) = \sum_{n=0}^{\infty} a_n(t)\phi_n(x) \tag{1.9}$$

For each t, $\{a_n(t)\}$ is the set of coefficients for expressing $u(x,t)$ in terms of the basis $\{\phi_n\}$.

Our objective now is to determine the functions $a_n(t)$. We write the heat source f in terms of eigenfunctions:

$$f(x,t) = \sum_{n=0}^{\infty} f_n(t)\phi_n(x) \tag{1.10}$$

Now substitute this equation 1.10 and the eigenfunction expansion for u (equation 1.9) into the PDE $u_t = -L[u] + f(x,t)$ to obtain

$$\underbrace{\sum_{n=0}^{\infty} a_n'(t)\phi_n(x)}_{u_t} = \underbrace{-\sum_{n=0}^{\infty} a_n(t)\lambda_n\phi_n(x)}_{-L[u]} + \underbrace{\sum_{n=0}^{\infty} f_n(t)\phi_n(x)}_{f} \tag{1.11}$$

In detail, the second term was found using the eigenfunction property and linearity of L:

$$\begin{aligned} L[u] &= L\left[\sum_{n=0}^{\infty} a_n(t)\phi_n(x)\right] \\ &= \sum_{n=0}^{\infty} a_n(t)L[\phi_n(x)] \\ &= \sum_{n=0}^{\infty} a_n(t)\lambda_n\phi_n(x) \end{aligned}$$

Note that since $a_n(t)$is only a function of t, it is constant as far as L is concerned so by linearity, $L[a_n(t)\phi_n(x)] = a_n(t)L[\phi_n(x)]$, where we have used the fact that L is linear to move.

Now we gather equation 1.11 together under the basis $\{\phi_n\}$:

$$0 = \sum_{n=0}^{\infty} \underbrace{[a_n'(t) + \lambda_n a_n(t) - f_n(t)]}_{\Psi_n(t)}\phi_n(x) = \sum_{n=0}^{\infty} \Psi_n(t)\phi_n(x)$$

Since the $\{\phi_n(x)\}$'s are a basis, the coefficient of each basis function must be zero at all times t (otherwise, the $\{\phi_n(x)\}$'s would be linearly dependent at some t), i.e. $\Psi_n(t) \equiv 0$. It follows that for each n,

$$a_n'(t) + \lambda_n a_n(t) = f_n(t), \quad \forall n > 0, t \in]0, T] \tag{1.12}$$

This equation is a first-order linear ODE for $a_n(t)$ that is easy to solve.

To complete our problem, the last missing piece is the initial condition, i.e. $a_n(0)$.

Recall we set $u(x, 0) = u_0(x)$. Write the IC in terms of eigenfunctions:

$$u_0(x) = \sum_{n=0}^{\infty} u_{0,n}\phi_n(x)$$

For the solution 1.9 satisfies the IC, we need the constants γ_n

$$\underbrace{\sum_{n=0}^{\infty} a_n(0)\phi_n(x)}_{u(x,0)} = \underbrace{\sum_{n=0}^{\infty} u_{0,n}\phi_n(x)}_{u_0(x)}$$

Again, since the $\{\phi_n(x)\}$'s are absis, the two sums must be equal term-by-term, so

$$a_n(0) = u_{0,n}, \quad \forall n$$

Finally, this condition and equation 1.12 lets us solve for a unique $a_n(t)$ (as we get a first order IC problem), which completes the process.

Sturm–Liouville Eigenvalue Problems

We have found the method of separation of variables to be quite successful in solving some homogeneous partial differential equations with homogeneous boundary conditions.

In all examples we have analyzed so far, the boundary value problem that determines the needed eigenvalues (separation constants) has involved the simple ordinary differential equation

$$\phi''(x) + \lambda\phi(x) = 0 \tag{1.13}$$

A more complete (and complicated) version is

$$\frac{d}{dx}(p\frac{d\phi}{dx}) + q\phi + \lambda\sigma\phi = 0 \tag{1.14}$$

Explicit solutions of this equation determined the eigenvalues λ from the homogeneous boundary conditions.

▶ **Proposition 1.14:** *For any regular Sturm-Liouville problem, all of the following theorems are valid:*

$\boxed{1.}$ *All the eigenvalues λ are real.*

$\boxed{2.}$ *There exist an infinite number of distinct eigenvalues:*

$$\lambda_1 < \lambda_2 < \cdots < \lambda_n < \cdots \to +\infty$$

$\boxed{3.}$ *Corresponding to each eigenvalue λ_n, there is an eigenfunction, denoted $\phi_n(x)$ (which is unique to within an arbitrary multiplicative constant)*

$\boxed{4.}$ *The eigenfunctions $\phi_n(x)$ form a "complete" set, meaning that any piecewise smooth function $f(x)$ can be represented by a generalized Fourier series of the eigenfunctions:*

$$f(x) \sim \sum_{n=0}^{\infty} a_n \phi_n(x)$$

$\boxed{5.}$ *Eigenfunctions belonging to different eigenvalues are orthogonal relative to the weight function $\sigma(x)$. In other words,*

$$\int_\Omega \phi_m(x)\phi_n(x)\sigma(x)dx = 0, \quad \lambda_m \neq \lambda_n$$

$\boxed{6.}$ *Any eigenvalue can be related to its eigenfunction by the Rayleigh quotient*

$$\lambda = \frac{-\int_\Gamma p\phi\frac{\partial\phi}{\partial n}d\Gamma + \int_\Omega [p(\frac{d\phi}{dx})^2 - q\phi^2)]}{\int_\Omega \phi^2\sigma}$$

1.4 Maximum Principle

▶ **Theorem 1.15** (Strong maximum principle): *Assume that $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ solve the heat equation in Ω_T, then*

$$\max_{\overline{\Omega}_T} u = \max_{\partial\Omega_T} u$$

▶ **Theorem 1.16:** *Assume that $u_0 \in L^2(\Omega)$, u is the solution of the equation 1.2, then we have, for all $(x,t) \in \Omega_T$*

$$\min\left\{0, \inf_\Omega u_0\right\} \leq u(x,t) \leq \max\left\{0, \sup_\Omega u_0\right\}$$

▷ *Proof:* Instead of a classical mean value formula [cf. Eva10, ch.02.3.2-02.3.3], we use Stampacchia's truncation method. Set

$$K = \max\left\{0, \sup_\Omega u_0\right\}$$

and assume that $K < +\infty$. Fix a function G s.t.

- $|G'(s)| \leq M, \quad \forall s \in \mathbb{R}$

- G is strictly increasing on $]0,\infty[$
- $G(s) = 0, \quad \forall s \leq 0$

and let

$$H(s) = \int_0^s G(r)\,\mathrm{d}r, \quad s \in \mathbb{R}$$

Easy to check the function φ defined by

$$\varphi(t) = \int_\Omega H(u(x,t) - K)\,\mathrm{d}x$$

has the following properties:

- $\varphi \in C([0,\infty[;\mathbb{R})$, $\varphi(0) = 0$, $\varphi \geq 0$ on $[0,\infty[$
- $\varphi \in C^1(]0,\infty[;\mathbb{R})$
-

$$\begin{aligned}\varphi'(t) &= \int_\Omega G(u(x,t) - K)\partial_t u(x,t)\,\mathrm{d}x \\ &= \int_\Omega G(u(x,t) - K)\Delta u(x,t)\,\mathrm{d}x \\ &= -\int_\Omega G'(u-K)\,|\nabla u|^2\,\mathrm{d}x \\ &\leq 0\end{aligned}$$

since $G(u(x,t) - K) \in H_0^1(\Omega)$ for every $t > 0$, it follows that $\varphi \equiv 0$ and thus, $\forall t > 0$, $u(x,t) \leq K$ a.e. on Ω. □

▶ **Theorem 1.17:** *Assume $u \in C(\overline{\Omega} \times [0,T])$, u is of class C^1 in t and of class C^2 in $x \in \Omega \times]0,T[$, $\partial_t u - \partial_{xx} u \leq 0 \in \Omega \times]0,T[$, then*

$$\max_{\overline{\Omega}\times[0,T]} u = \max_{\mathfrak{P}} u$$

where $\mathfrak{P} = (\overline{\Omega} \times \{0\}) \cup (\Gamma \times]0,T[)$ is called the ***parabolic boundary*** *of the cylinder $\Omega \times]0,T[$*

2 Numerical Approaches

2.1 Discrete and Fast Fourier Transform (DFT & FFT)

Consider a bounded open area $\Omega \subset \mathbb{R}^d$:

$$\begin{cases} u_t - u_{xx} &= 0 \quad \forall x \in \Omega \subset \mathbb{R}^d \\ u|_{\partial\Omega} &= 0 \\ u(x,0) &= u_0 \end{cases} \tag{2.1}$$

The existence and uniqueness of the solution is guaranteed by the theorem below:

▶ **Theorem 2.1:** *For all $u_0 \in L^2(\Omega)$, there exists a unique solution*

$$u \in C^0(\mathbb{R}^+; L^2(\Omega)) \cap C^0(\mathbb{R}^+_*; L^2(\Omega)) \cap C^1(\mathbb{R}^+; L^2(\Omega))$$

of the equation 2.1.

In section 1.3.2 we have exhibited a complete procedure to solve heat equation in using eigenfunction's method, however, it's impossible to calculate a basis and corresponding coefficients presented by infinite series in computer. Generally we calculate a sufficient large number of finite items to approach the infinite series.

Let's start with a concrete example in 1D spacetime .

Suppose we have a mug of water and place a tea bag in it,then let it sit. We want to know how long it will take for the tea to diffuse through so that it is uniformly mixed. Assume the mug is one dimensional with height 2 and let $c(x,t)$ be the concentration of tea at height x and time t.

$$\partial_t u - k u_{xx} = 0 \quad (x,t) \text{ in }]0,1[\times]0,\infty[\tag{2.2}$$

where k > 0 is the diffusivity of the tea. Since tea cannot leave the cup, the flux at the top and bottom must be zero, so impose Neumann boundary conditions

$$u_x(0,t) = u_x(1,t) = 0$$

Suppose the teabag starts at the bottom of the cup $(x = 0)$, and the initial concentration is

$$u(x,0) = \begin{cases} 1, & x < 1/2 \\ 0, & x > 1/2 \end{cases}$$

Eigenfunctions:

$$u_t = -2tL[u] + f(x,t)$$

where

$$L = \partial_{xx}, \quad f(x,t) = e^{-t^2} \sin x$$

Note that we cannot put t in the definition of L, since L must only involve derivatives in x. The eigenvalues/functions are

$$\lambda_n = n^2, \quad \phi_n(x) = \sin(nx), \quad n = 1, 2, \ldots$$

[cf. Sha03; Sch01, ch.08; Mal08, ch.03.3]

because $\{e_k(t)\}_{k\in\mathbb{Z}} = \left\{e^{2i\pi kt}\right\}_{k\in\mathbb{Z}}$ is a Hilbert basis.

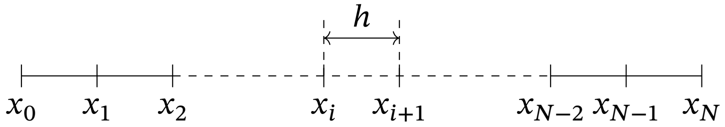


Figure 3: Equilong discretization in one dimension

The subsection 1.3 inspires us

$$\widehat{f}(k) = \int_0^\tau f(x) e^{-2i\pi kx}\, \mathrm{d}x \xrightarrow{\text{discretization}} U_k = \frac{1}{N} \sum_{j=0}^{N-1} f(\frac{j}{N}) e^{-2i\pi k \frac{j}{N}} \tag{2.3}$$

2.1.1 FFT: interlacement

2.2 Finite Difference Method (FDM) for 1D Heat Equation

2.2.1 Explicit Euler's scheme

First, we determine the heat equation we need to consider

$$\begin{cases} \frac{\partial u}{\partial t}(x,t) + \frac{\partial^2 u}{\partial x^2}(x,t) & = f(x,t), \quad \text{in } \Omega \\ u(0,t) = u(L,t) & = g(t) \quad\quad \text{on } (0,T) \\ u(x,0) & = u_0(x) \quad\quad \text{in } \Omega \end{cases} \tag{2.4}$$

The general idea of finite difference methods for evolution equations is that we replace the continuous variables in time and space with discrete points. In the case of an evolution equation, we need a spacetime grid. Let us thus be given two positive integers N and M. We set $h = \delta x = \frac{1}{N+1}$ and $X_n = nh$ for $n = 0, 1, 2, 3, \cdots, N+1$.We set $k = \delta t = \frac{T}{M+1}$ and $t_j = jk$ for $j = 0, 1, 2, 3, \cdots, M+1$. The parameter h is called the space grid step and the parameter k the time grid step, or time step. The grid points are the points (x_n, t_j).we will let N and M go to infinity, or h and k go to 0.

The u_n^j for the above values of n and j, and it is hoped that u_n^j will be an approximation of $u(x_n, t_j)$, that should become better and better as N and M are increased. The boundary condition can be enforced exactly by requiring that

$$u_0^j = u_{N+1}^j = g(j) = 0$$

And we note that $j = 0, 1, 2, 3, \cdots, M+1$. By the initial condition, we can acquire that

$$u_n^0 = u_0(x_n)$$

And we note that $n = 1, \cdots, N$.

In this way, we will find that all the points of the boundary are known.The only values that are left unknown at this stage are thus u_n^j for $n = 1, 2, 3, \cdots, N$ and $j = 1, 2, 3, \cdots, M+1$. We define that

$$U^j = \begin{pmatrix} u_1^j \\ u_2^j \\ \vdots \\ u_N^j \end{pmatrix} \in \mathbb{R}^N$$

Next, we need to calculate the remaining unknown points through the points of the known boundary. In the explicit Euler three point scheme, we can use the forward differential quotient approximation

$$\frac{\partial u}{\partial t}(x_n, t_j) \approx \frac{u(x_n, t_{j+1}) - u(x_n, t_j)}{k} \tag{2.5}$$

and for the second order space derivative,Using the same approximation, we can get:

$$\frac{\partial^2 u}{\partial x^2}(x_n, t_j) \approx \frac{\frac{u(x_{n+1},t_j) - u(x_n,t_j)}{h} - \frac{u(x_n,t_j) - u(x_{n-1},t_j)}{h}}{h} = \frac{u(x_{n+1}, t_j) - 2u(x_n, t_j) + u(x_{n-1}, t_j)}{h^2} \tag{2.6}$$

We substitute (3) and (4) into (1), we can get

$$\begin{cases} \frac{u_n^{j+1}-u_n^j}{k} - \frac{u_{n+1}^j-2u_n^j+u_{n-1}^j}{h^2} & = f_n^j \quad \text{for } n=1,2,3,\cdots,N, j=0,1,2,3,\cdots,M, \\ u_n^0 & = u_0(x_n) \quad \text{for } n=1,...,N \\ u_0^j = u_{N+1}^j & = g^j \quad \text{for } j=0,\cdots,M+1 \end{cases} \tag{2.7}$$

So,we can rewrite the first N equations of the scheme in vector form as

$$\frac{U^{j+1}-U^j}{k} + A_h U^j = F^j \quad \text{for } j=0,1,2,3,\dots,M. \tag{2.8}$$

where A_h is the same $N \times N$ tridiagonal matrix[1]

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

And we define that F_j and the discrete initial condition are two vectors of $\mathbb{R}^N$.

$$F^j = \begin{pmatrix} f_1^j \\ f_2^j \\ \vdots \\ f_N^j \end{pmatrix} \in \mathbb{R}^N \qquad U_0 = \begin{pmatrix} u_0(x_1) \\ u_0(x_2) \\ \vdots \\ u_0(x_N) \end{pmatrix} \in \mathbb{R}^N$$

So,we can get the numerical scheme :

$$\begin{cases} U^{j+1} = (I-kA_h)U^j + kF^j \text{ for } j=0,\cdots,M \\ U^0 = U_0 \end{cases} \tag{2.9}$$

2.2.2 Implicit Euler's scheme and Leapfrog method

For a full disciption of the schemes, [cf. Luc16, ch.08.2].

The first example is the implicit or backward Euler three point scheme, which is associated with the backward differential quotient approximation of the time derivative

$$\frac{\partial u}{\partial t}(x_n,t_j) \approx \frac{u(x_n,t_j)-u(x_n,t_{j-1})}{k} \tag{2.10}$$

In vector form, this scheme reads

$$\frac{U^j-U^{j-1}}{k} + A_h U^j = F^j \quad \text{for } j=1,2,3,\dots,M+1. \tag{2.11}$$

[1] Since the Dirichlet BC is fixed, which indicates the insulation of rod, the A_h holds thus the same during iteration, otherwise A_h evolves along with each step of iteration.

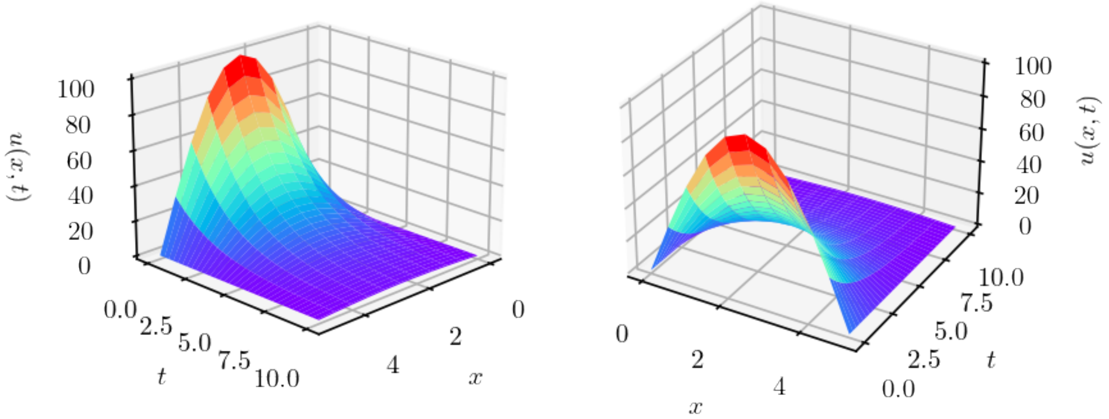


Figure 4: The numerical simulation for the 1D diffusion equation

This scheme is called implicit, because the above formula is not a simple recurrence relation. Indeed, U^{j+1} appears as the solution of an equation once U^j is known. It is not a priori clear that this equation is solvable. In this particular case, we have

$$\begin{cases} U^j &= (I + kA_h)^{-1}(U^{j-1} + kF^j) \quad \text{for } j = 1, 2, 3, \dots, M+1 \\ U^0 &= U_0 \end{cases} \tag{2.12}$$

since it is not hard to see that the matrix $I+kA_h$ is symmetric, positive definite, hence invertible.

In practical terms, the implementation of the backward Euler method entails the solution of a linear system at each time step, whereas the explicit method is simply a matrix-vector product and vector addition at each time step. The implicit method is thus more computationally intensive than the explicit method, but it has other benefits as we will see later.

The second example is the leapfrog or Richardson method, which is associated with the central differential quotient approximation of the time derivative

$$\frac{\partial u}{\partial t}(x_n, t_j) \approx \frac{u(x_n, t_{j+1}) - u(x_n, t_{j-1})}{2k} \tag{2.13}$$

Like the explicit Euler three point method we substitute this into the heat equation to get the equations for U^{j+1}, $U^j$and U^{j-1}:

$$\frac{U^{j+1} - U^{j-1}}{2k} + A_h U^j = F^j \quad \text{for } j = 1, \dots, M. \tag{2.14}$$

Thus,we find that this method is an explicit two-step method since U^{j+1} is explicitly given in terms of U^j and U^{j-1}. Simplify the equation above, we can get :

$$\begin{cases} U^{j+1} & = U^{j-1} - 2kA_h U^j + 2kF^j \quad \text{for} j = 0, \cdots, M \\ U^0 & = U_0, U^1 = U_1 \end{cases} \tag{2.15}$$

In particular, since this is a two-step method, we must somehow be ascribed to U^1 in order to initialize the recurrence, in addition to U^1.

2.2.3 Crank-Nicolson method

$$\begin{cases} \frac{u_n^{j+1}-u_n^j}{k} & +\frac{1}{2}\left(-\frac{u_{n+1}^{j+1}-2u_n^{j+1}+u_{n-1}^{j+1}}{h^2} + \frac{u_{n+1}^{j+1}-u_{n-1}^{j+1}}{h}\right) \\ & +\frac{1}{2}\left(-\frac{u_{n+1}^{j}-2u_n^{j}+u_{n-1}^{j}}{h^2} + \frac{u_{n+1}^{j}-u_{n-1}^{j}}{h}\right) = \frac{1}{2}f(x_n, t_{j+1}) + \frac{1}{2}f(x_n, t_j), \quad n \in \{1, \dots, N\}, j \in \{1, \dots, J\} \\ u_n^0 = u_0(x_n), & n \in \{1, \dots, N\} \\ u_0^j = u_{N+1}^j = g^j, & j \in \{0, \dots, J\} \end{cases} \tag{2.16}$$

From 2.16, we could establish directly the relationship

$$\frac{1}{k}(U^{j+1} - U^j) + \frac{1}{2}(\frac{1}{h^2}AU^{j+1} + \frac{1}{h}BU^{j+1}) + \frac{1}{2}(\frac{1}{h^2}AU^j + \frac{1}{h}BU^j) = \frac{1}{2}(F^{j+1} + F^j) \tag{2.17}$$

with

$$A = \begin{pmatrix} 2 & -1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 2 & -1 & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 2 \end{pmatrix} \quad B = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & \dots & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 & -1 & 0 \end{pmatrix}$$

We reformulate the method 2.16 in a condensed way:

$$\begin{cases} \left(I + \frac{k}{2}C\right)U^{j+1} = \left(I - \frac{k}{2}C\right)U^j + \frac{k}{2}(F^{j+1} + F^j), \quad j \in \{1, \dots, J\} \\ U^0 = U_0 \end{cases} \tag{2.18}$$

then A, B, C satisfied $C = \frac{1}{h^2}A + \frac{1}{h}B$, we shall write C explicitly:

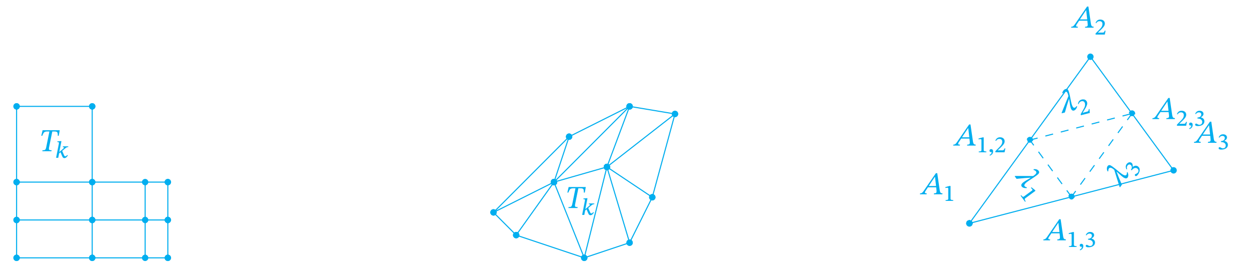
$$C = \frac{1}{h^2}\begin{pmatrix} 2 & -1+h & 0 & 0 & 0 & \dots & 0 \\ -1-h & 2 & -1+h & 0 & 0 & \dots & 0 \\ 0 & -1-h & 2 & -1+h & 0 & \dots & 0 \\ 0 & 0 & -1-h & 2 & -1+h & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 2 & -1+h \end{pmatrix}$$

Given $N+2, M+2 \in \mathbb{N}^*$ the number of knots incluing end points, we take space step size $h = 1/(N+1)$ and set inner point $x_i = ih$, similally temporal step size is set as $k = 1/(M+1)$. The unknown values are $u_1^j, u_2^j, \dots, u_n^j, \dots, u_N^j$. For $n = 1, \cdots, N$ and $j = 1, \cdots, M+1$, the corresponding unknown vector is $U^j = (u_1^j, \dots, u_N^j)^T$. N.B. u_0^j, u_{N+1}^j are defined by BCs.

2.3 Finite Element Method (FEM) Approximation for 1D Heat Equation

[NOT FINISHED]

However, in higher dimensional case, the design of mesh turns to be complicated. We list the following cases in figure 5, each of them has advantages and disadvantages, we don't expand the details here.



(a) Rectangular Q_1 Finite Elements

(b) Triangular P_1 Lagrange Elements

(c) Triangular P_2 Lagrange Elements

Figure 5: Different 2D FEM schemes

2.4 FDM for 2D Heat Equation

In constructing numerical approximations to solutions of the heat equation in a bounded domain Ω of the plane, approximately, with a lattice and replaces the second partial derivatives with **centered differences**:

$$u_{xx} \approx \frac{u_W - 2u_O + u_E}{h_x^2}$$

$$u_{yy} \approx \frac{u_N - 2u_O + u_S}{h_y^2}$$

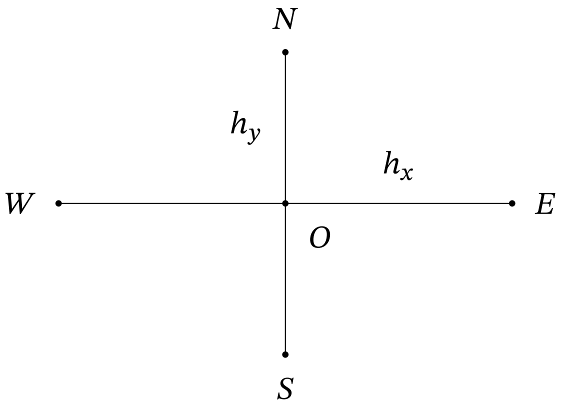


Figure 6: 2D centered difference approximation

and h_x (h_y) is the horizonal (with respect to vertical) mesh spacing.

For simplicity, we will use the same mesh spacing $h_x = h_y = h$ for both two dimensions. Let us consider the following scheme : given IC $u_0(x, y)$, we define iteratively the sequence $u_j(x, y)$ by

$$\begin{aligned}
&\frac{u_{j+1}(x,y) - u_j(x,y)}{k} - \Delta u(x,y) = f(x,y,j)\\
\Longrightarrow\ &\frac{u_{j+1}(x,y) - u_j(x,y)}{k} - \frac{u_j(x+h,y) + u_j(x-h,y) + u_j(x,y+h) + u_j(x,y-h) - 4u_n(x,y)}{h^2}\\
&= f(x,y,j)\\
\Longrightarrow\ &u_{j+1} - u_j = \frac{k}{h^2}\left[u_j(x+h,y) + u_j(x-h,y) + u_j(x,y+h) + u_j(x,y-h) - 4u_n(x,y)\right] + kf(x,y,j)
\end{aligned}$$

We shall write it

$$u_{j+1} = u_j + A * u_j + kf_j, \quad \text{with} \quad A = \begin{bmatrix} 0 & s & 0 \\ s & -4s & s \\ 0 & s & 0 \end{bmatrix} \text{ and } s = \frac{k}{h^2}$$

✠ **Definition 2.2** (Multidimensional discrete convolution)**:**

$$\begin{aligned}
y(n_1, n_2, \dots n_M) &= x(n_1, n_2, \dots, n_M) \overbrace{* \cdots *}^{M\ times} h(n_1, n_2, \dots, n_M)\\
&= \sum_{k_1=-\infty}^{\infty} \sum_{k_2=-\infty}^{\infty} \cdots \sum_{k_M=-\infty}^{\infty} h(k_1, k_2, \dots, k_M) x(n_1 - k_1, n_2 - k_2, \dots, n_M - k_M)
\end{aligned}$$

2.4.1 The Explicit Euler Three Point Scheme

The two-dimensional heat conduction equation has the form

$$\begin{cases} \frac{\partial u}{\partial t}(x,y,t) + \frac{\partial^2 u}{\partial x^2}(x,y,t) + \frac{\partial^2 u}{\partial y^2}(x,y,t) = f(x,y,t), \\ u(x,y,0) = u_0(x,y) \quad \text{in } \Omega \times \{t = 0\} \end{cases} \tag{2.20}$$

Euler's method only deals with the time partial derivatives, and then discretizes the second-order spatial derivatives. Use $u_{i,j}$ to represent the temperature value of row i (y direction) and column j (x direction), the temperature field at time t is stored in a matrix of size (N_y+1, N_x+1), N_x and N_y represent the number of segments in the direction x and y , the equal length of each segment is $\Delta x, \Delta y$.(For variable t, the letter representation is the same as before)

We redefine the matrix U:

$$U = \begin{pmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,N_x} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,N_x} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_y,0} & \cdots & \cdots & u_{N_y,N_x} \end{pmatrix}$$

Each point uses the central difference format for the x and y directions, we can get:

$$\begin{cases} \frac{\partial^2 u_{i,j}}{\partial x^2} = \frac{1}{\Delta x^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}), \\ \frac{\partial^2 u_{i,j}}{\partial y^2} = \frac{1}{\Delta y^2}(u_{i+1,j} - 2u_{i,j} + u_{i-1,j}), \end{cases} \tag{2.21}$$

Assemble the second-order partial derivatives into a matrix form with the same dimensions as U

$$\frac{\partial^2}{\partial x^2}\begin{pmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,N_x} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,N_x} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_y,0} & \cdots & \cdots & u_{N_y,N_x} \end{pmatrix} =$$

$$\frac{1}{\Delta x^2}\begin{pmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,N_x} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,N_x} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_y,0} & \cdots & \cdots & u_{N_y,N_x} \end{pmatrix}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{pmatrix}$$

$$\frac{\partial^2}{\partial y^2}\begin{pmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,N_x} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,N_x} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_y,0} & \cdots & \cdots & u_{N_y,N_x} \end{pmatrix} =$$

$$\frac{1}{\Delta y^2}\begin{pmatrix} -2 & 1 & 0 & \cdots & 0 \\ 1 & -2 & 1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 1 & -2 & 1 \\ 0 & \cdots & 0 & 1 & -2 \end{pmatrix}\begin{pmatrix} u_{0,0} & u_{0,1} & \cdots & u_{0,N_x} \\ u_{1,0} & u_{1,1} & \cdots & u_{1,N_x} \\ \vdots & \vdots & \ddots & \vdots \\ u_{N_y,0} & \cdots & \cdots & u_{N_y,N_x} \end{pmatrix}$$

Written in matrix product form, denoted as:

$$\begin{cases} \frac{\partial^2 U}{\partial x^2} = \frac{1}{\Delta x^2}UA, \\ \frac{\partial^2 U}{\partial y^2} = \frac{1}{\Delta y^2}BU, \end{cases} \tag{2.22}$$

A and$B$ are called difference coefficient matrices, they are different in size, A is (N_x+1, N_x+1), B is (N_y+1, N_y+1). So the iterative formula becomes:

$$\begin{cases} U^{n+1} = U^n + k(\frac{1}{\Delta x^2}U^nA + \frac{1}{\Delta y^2}BU^n) + kF^n \ for \ n = 0, \cdots, M \\ U^0 = V_0(X, Y) \end{cases} \tag{2.23}$$

It should be noted here that the boundary point of the second-order partial derivative matrix obtained by the formula is invalid, because the first row and the last row of the difference coefficient matrix are incomplete, which is equivalent to the temperature of the boundary point is always 0, so it needs to be given separately boundary conditions to determine the temperature value of the matrix boundary.

3 Analysis of Algorithms

There is no uniqueness of a general form for a given scheme. Indeed, given a general form, we can obtain another one by multiplying everything by an arbitrary function of h and k, different

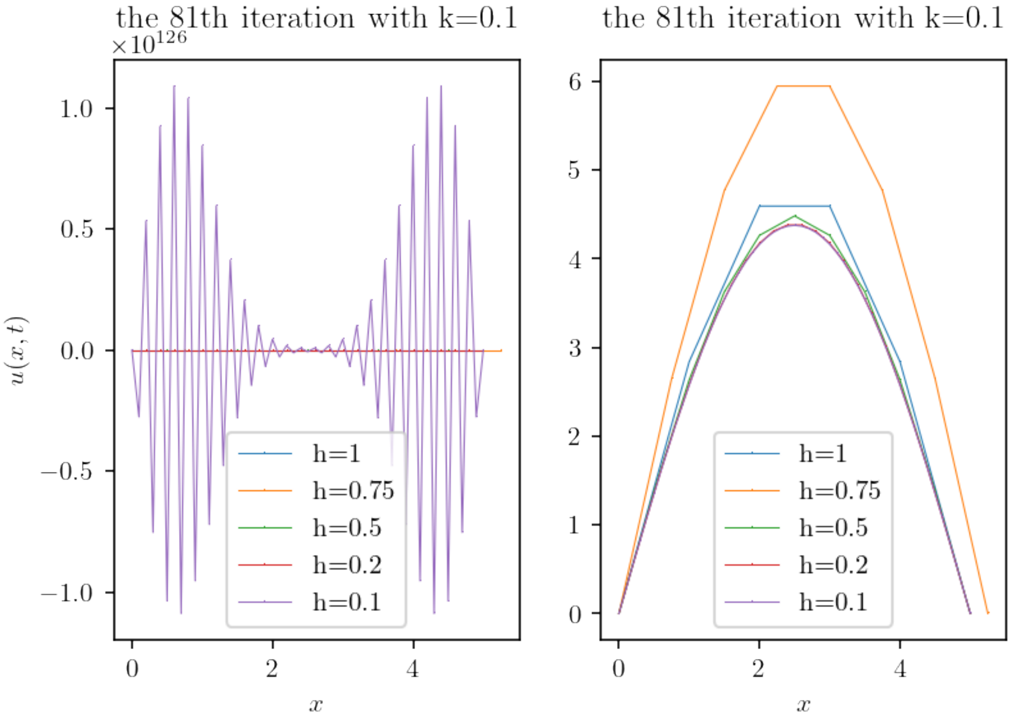


Figure 7: Approximation by explicit method (left) and implicit method (right)

general forms give rise to different properties of schemes. In practice, we use different scheme in consideration of various reasons, such as the accuracy of the scheme, the computational cost, the stability of the scheme, the convergence rate of the scheme, etc.

These reasons constitute the main factors that determine the choice of a particular scheme. Here we discuss the most important two : consistency and stability.

Before continuing, let's observe an interesting (also confusing) phenomenon that arises when we approximate $u(x,t)$ by different schemes in using various h and k values in figure 7.

Here we controlled the time step k unchanged and we randomly picked a time section during iterative procedure:

```python
import random
t_section = random.randint(1,time_knots-1)
```

Listing 1: randomly choiced time knot during iterative procedure

The result is somewhat surprising. It seems that the explict method failled to converge to the solution, while the implicit method converged to the solution. This is because the explicit method is not stable, and the implicit method is not consistent.

3.1 Consistency

✠ **Definition 3.1** (truncation error)**:** *In a word, truncation error is an error between closed form solution and numerical output caused by approximating a mathematical process.*

In Fourier analysis we have our classical truncation error (inducted from Bessel's inequality and Parseval's equality) defined by

$$\varepsilon_{\text{Fourier}}[M] = \left\| f\text{-}\widetilde{f}_M \right\|^2 = \sum_{k>M/2} \left| \left\langle f(\xi), e^{2i\pi k\xi} \right\rangle \right|^2$$

✠ **Definition 3.2:** *Let u be a function defined on Ω. We define the space grid sampling operator S_h by*

$$S_h(u^j) = \begin{pmatrix} u(x_1, t_j) \\ u(x_2, t_j) \\ \vdots \\ u(x_N, t_j) \end{pmatrix} \in \mathbb{R}^N$$

We assume that u is the solution of the heat equation, we define the truncation error of the present finite difference method to be the sequence of vectors

$$\varepsilon_{h,k}(u)^j = \frac{S_h(u^{j+1}) - S_h(u^j)}{k} + A_h S_h(u^j) - F^j \tag{3.1}$$

To get the truncation error, we simply take the finite difference format and replace Corresponding grid sampling of discrete unknowns and solutions heat equation. If we fictionally apply Numerical scheme with one time step starting from the exact sample value at t_j, If the data at $S_h(u^j)$ is accurate, we note the data at t_{j+1} as $\widetilde{U}$ by the the present finite difference method :

$$\widetilde{U}^{j+1} = S_h(u^j) - kA_h S_h(u^j) + kF^j \tag{3.2}$$

We can combine the two equations above to get an error :

$$S_h(u_{t_{j+1}}) - \widetilde{U}^{j+1} = k\varepsilon_{h,k}(u)^j \tag{3.3}$$

In order to analyze the convergence of finite difference methods, we need to introduce the function spaces:

$$C^{m,n}(\overline{Q}) = \{\forall t \in [0,T], u_t \in C^m([0,1]), \forall x \in [0,1], u_x \in C^n([0,T])\} \tag{3.4}$$

with all derivatives uniformly bounded on $\overline{Q}$.

For the time being, we equip the space $\mathbb{R}^N$ with the infinity norm as before, except that the dependence on h is here made explicit,

$$\|U\|_{\infty,h} = \max_{1\le n\le N} |U_n| \tag{3.5}$$

where $U_1, \cdots, U_N$ are the components of $U \in \mathbb{R}^N$. It follows that

▶ **Proposition 3.3:** *Assume that $u \in C^{4,2}(\overline{Q})$. Then we have*

$$\max_{0\le j\le M} \left|\varepsilon_{h,k}(u)^j\right|_{\infty,h} \le C(h^2 + k)$$

where the constant C depends only on u.

▷ *Proof:* We use Taylor-Lagrange expansions. First noticed the fact that u_x is of class C^2. Therefore, for all n and j, there exists $\theta_n^j \in [t_j, t_{j+1}]$ such that

$$\begin{aligned} u(x_n, t_j+1) &= u(x_n,t_j) + k\frac{\partial u}{\partial t}(x_n,t_j) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_n,\theta_n^j) \\ \frac{u(x_n,t_j+1)-u(x_n,t_j)}{k} &= \frac{\partial u}{\partial t}(x_n,t_j) + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_n,\theta_n^j) \end{aligned}$$

Similarly, u_t is of class C^4. Therefore, for all n and j, there exists $\gamma_n^{j,+} \in (x_n, x_{n+1})$ and $\gamma_n^{j,-} \in (x_{n-1}, x_n)$, we can get the equation:

$$\begin{aligned} u(x_{n+1},t_j) &= u(x_n,t_j) + h\frac{\partial u}{\partial x}(x_n,t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_n,t_j) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_n,t_j) + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4}(\gamma_n^{j,+},t_j) \\ u(x_{n-1},t_j) &= u(x_n,t_j) - h\frac{\partial u}{\partial x}(x_n,t_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_n,t_j) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_n,t_j) + \frac{h^4}{24}\frac{\partial^4 u}{\partial x^4}(\gamma_n^{j,-},t_j) \\ \frac{u(x_{n+1},t_j)-2u(x_n,t_j)+u(x_n,t_j)}{h^2} &= \frac{\partial^2 u}{\partial x^2}(x_n,t_j) + \frac{h^2}{12}\frac{\partial^4 u}{\partial x^4}(\gamma_n^j,t_j) \end{aligned}$$

where $\gamma_n^j \in (x_{n-1}, x_{x+1})$. Taking into account the boundary conditions $u(x_0,t_j) = u(x_{N+1},t_j) = 0$, we can get that,

$$\varepsilon_{h,k}(u)^j = S_h((\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2})_{t_j}) - F^j + R^j \tag{3.6}$$

with

$$R_n^j = \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_n,\theta_n^j) - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^2}(\gamma_n^j,t_j) \tag{3.7}$$

So, we will find that if u is the solution of the heat equation, the following formula will be zero.

$$S_h((\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2})_{t_j}) - F^j = 0 \tag{3.8}$$

in this case we have

$$\varepsilon_{h,k}(u)^j = R^j \tag{3.9}$$

Moreover,

$$|R_n^j| \leq \max(\frac{1}{2}\max_{\overline{Q}}|\frac{\partial^2 u}{\partial t^2}|, \frac{1}{12}\max_{\overline{Q}}|\frac{\partial^4 u}{\partial x^4}|)(k+h^2) \tag{3.10}$$

□

✠ **Definition 3.4** (consistency)**:** *We say that the scheme is consistent for the family of norms* $\|\cdot\|_N$ *if*

$$\max_j \left\|\varepsilon_{h,k}(u)^j\right\|_N \xrightarrow{(h,k)\to(0,0)} 0 \tag{3.11}$$

And we say it's of order p *in space and* q *in time for the family of norms* $\|\cdot\|_N$ *if*

$$\max_j \left\|\varepsilon_{h,k}(u)^j\right\|_N \leq C(h^p + k^q) \tag{3.12}$$

Where C *is a constant that depends only on* u*.*

Consistency means that the scheme is trying its best to locally approximate the right numerical problem in the norm $\|\cdot\|_N$.

3.2 Truncation Error of 2D Explicit Euler Scheme

We will show here the format of the truncation error in the 2D case

✠ **Definition 3.5:** *Let u be a function defined on $[0,1]\times[0,1]$. We define the space grid sampling operator S_h by*

$$S_{\Delta x,\Delta y}(u^j) = \begin{pmatrix} u(x_0,y_0,t_j) & u(x_1,y_0,t_j) & \cdots & u(x_{N_x},y_0,t_j) \\ u(x_0,y_1,t_j) & u(x_1,y_1,t_j) & \cdots & u(x_{N_x},y_1,t_j) \\ \vdots & \vdots & \ddots & \vdots \\ u(x_0,y_{N_y},t_j) & \cdots & \cdots & u(x_{N_x},y_{N_y},t_j) \end{pmatrix} \tag{3.13}$$

$$\varepsilon_{\Delta x,\Delta y,k}(u)^j = \frac{S_{\Delta x.\Delta y}(u^{j+1}) - S_{\Delta x,\Delta y}(u^j)}{k} - S_{\Delta x,\Delta y}(u^j)A - BS_{\Delta x,\Delta y}(u^j) - F^j \tag{3.14}$$

Same as before, we can use the fact that u_x is of class C^2. Therefore, for all n and j, there exists $\theta_n^j \in [t_j, t_{j+1}]$ such that

$$\begin{aligned} u(x_n,y_m,t_{j+1}) &= u(x_n,y_m,t_j) + k\frac{\partial u}{\partial t}(x_n,y_m,t_j) + \frac{k^2}{2}\frac{\partial^2 u}{\partial t^2}(x_n,y_n,\theta_n^j) \\ \frac{u(x_n,y_m,t_j+1) - u(x_n,y_m,t_j)}{k} &= \frac{\partial u}{\partial t}(x_n,y_m,t_j) + \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_n,y_m,\theta_n^j) \end{aligned}$$

Therefore, for all n,m and j, there exists $\gamma_n^{j,+} \in (x_n, x_{n+1})$ and $\gamma_n^{j,-} \in (x_{n-1}, x_n)$,there exists $\xi_m^{j,+} \in (y_m, y_{m+1})$ and $\xi_m^{j,-} \in (y_{m-1}, y_m)$, we can get the equation:

$$\begin{aligned} u(x_{n+1},y_m,t_j) &= u(x_n,y_m,t_j) + \Delta x\frac{\partial u}{\partial x}(x_n,y_m,t_j) + \frac{(\Delta x)^2}{2}\frac{\partial^2 u}{\partial x^2}(x_n,y_m,t_j) \\ &\quad + \frac{(\Delta x)^3}{6}\frac{\partial^3 u}{\partial x^3}(x_n,y_m,t_j) + \frac{(\Delta x)^4}{24}\frac{\partial^4 u}{\partial x^4}(\gamma_n^{j,+},y_m,t_j) \\ u(x_{n-1},y_m,t_j) &= u(x_n,y_m,t_j) - \Delta x\frac{\partial u}{\partial x}(x_n,y_m,t_j) + \frac{(\Delta x)^2}{2}\frac{\partial^2 u}{\partial x^2}(x_n,y_m,t_j) \\ &\quad - \frac{(\Delta x)^3}{6}\frac{\partial^3 u}{\partial x^3}(x_n,y_m,t_j) + \frac{(\Delta x)^4}{24}\frac{\partial^4 u}{\partial x^4}(\gamma_n^{j,-},y_m,t_j) \end{aligned}$$

$$\begin{aligned} \frac{u(x_{n+1},y_m,t_j) - 2u(x_n,y_m,t_j) + u(x_n,y_m,t_j)}{(\Delta x)^2} =& \frac{\partial^2 u}{\partial x^2}(x_n,y_m,t_j) \\ &+ \frac{(\Delta x)^2}{12}\frac{\partial^4 u}{\partial x^4}(\gamma_n^j,y_m,t_j) \end{aligned}$$

$$
\begin{aligned}
u(x_n, y_{m+1}, t_j) &= u(x_n, y_m, t_j) + \Delta y \frac{\partial u}{\partial y}(x_n, y_m, t_j) + \frac{(\Delta y)^2}{2}\frac{\partial^2 u}{\partial y^2}(x_n, y_m, t_j) \\
&\quad + \frac{(\Delta y)^3}{6}\frac{\partial^3 u}{\partial y^3}(x_n, y_m, t_j) + \frac{(\Delta y)^4}{24}\frac{\partial^4 u}{\partial y^4}(x_n, \xi_m^{j,+}, t_j) \\
u(x_n, y_{m-1}, t_j) &= u(x_n, y_m, t_j) - \Delta y \frac{\partial u}{\partial y}(x_n, y_m, t_j) + \frac{(\Delta y)^2}{2}\frac{\partial^2 u}{\partial y^2}(x_n, y_m, t_j) \\
&\quad - \frac{(\Delta y)^3}{6}\frac{\partial^3 u}{\partial y^3}(x_n, y_m, t_j) + \frac{(\Delta y)^4}{24}\frac{\partial^4 u}{\partial y^4}(x_n, \xi_m^{j,-}, t_j)
\end{aligned}
$$

$$
\begin{aligned}
\frac{u(x_n, y_{m+1}, t_j) - 2u(x_n, y_m, t_j) + u(x_n, y_{m-1} t_j)}{(\Delta y)^2} &= \frac{\partial^2 u}{\partial y^2}(x_n, y_m, t_j) \\
&\quad + \frac{(\Delta y)^2}{12}\frac{\partial^4 u}{\partial x^4}(\xi_n^j, y_m, t_j)
\end{aligned}
$$

We repeat the calculation of the one-dimensional case, we can get:

$$
R_{n,m}^j = \frac{k}{2}\frac{\partial^2 u}{\partial t^2}(x_n, y_m, \theta_n^j) - \frac{h^2}{12}\frac{\partial^4 u}{\partial x^2}(\gamma_n^j, y_m, t_j) - \frac{h^2}{12}\frac{\partial^4 u}{\partial y^2}(x_n, \xi_m^j, t_j) \tag{3.15}
$$

Moreover,

$$
|R_{n,m}^j| \le \max\left(\frac{1}{2}\max_{\overline{Q}}|\frac{\partial^2 u}{\partial t^2}|, \frac{1}{12}\max_{\overline{Q}}|\frac{\partial^4 u}{\partial x^4}|, \frac{1}{12}\max_{\overline{Q}}|\frac{\partial^4 u}{\partial y^4}|\right)(k + (\Delta x)^2 + (\Delta y)^2) \tag{3.16}
$$

Thus we can get the following equation:

$$
\lim_{(\Delta x, \Delta y, k)\to(0,0,0)} \max_{0\le j\le M} \left\|\varepsilon_{\Delta x, \Delta y, k}(u)^j\right\|_{\infty,h} \le \lim_{(\Delta x, \Delta y, k)\to(0,0,0)} C((\Delta x)^2 + (\Delta y)^2 + k) = 0 \tag{3.17}
$$

So explicit Euler method is consistent in 2D case.

3.3 Stability

3.4 The Stability of the Explicit Euler Scheme

✠ **Definition 3.6** (stability): *if$(U^j)_{0\le j\le M}$ is a solution to heat equation. We say that the scheme (3.10) is stable if there exists a constant C such that*

$$
\max_{0\le j\le M} ||U^j||_\infty \le C(||U^0||_\infty + T\max_{0\le j\le M}||F^j||_\infty) \tag{3.18}
$$

▶ **Theorem 3.7** (CFL condition for the stability of the explicit Euler scheme): *Let (U^j), $0 \le j \le M$ be the solution of heat equation associated with a uniform mesh of parameters $\Delta t, \Delta x$. If the CFL (Courant-Friedricks-Levy) condition:*

$$
c = v\frac{\Delta t}{(\Delta x)^2} < 1/2 \tag{3.19}
$$

is satisfied, then the scheme is stable and

$$
\max_{0\le j\le M} ||U^j||_\infty \le ||U^0||_\infty + T\max_{0\le j\le M}||F^j||_\infty \tag{3.20}
$$

▷ *Proof:* In the one-dimensional case, we get the schema($v = 1$)

$$\begin{cases} U^{j+1} = (I - \Delta t A_h)U^j + \Delta t F^j \ for\ j = 0, \cdots, M \\ U^0 = U_0 \end{cases} \tag{3.21}$$

so,we can get the equation:

$$U^{j+1} = (I - cA)U^n + (\Delta t)F^j \tag{3.22}$$

$$||U^{j+1}||_\infty \leq ||(I - cA)U^j||_\infty + (\Delta t)||F^j||_\infty \tag{3.23}$$

$$\leq ||(I - cA)||_\infty ||U^j||_\infty + (\Delta t) \max_{0 \leq j \leq M} ||F^j||_\infty \tag{3.24}$$

If A is the matrix mentioned earlier:

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \tag{3.25}$$

and $c = \frac{\Delta t}{(\Delta x)^2} > 0$, so

$$||(I - cA)||_\infty = |1 - 2c| + 2c$$

For this equation, I have the following interpretation:

$$I - cA = \begin{pmatrix} 1 - 2c & c & 0 & \cdots & 0 \\ c & 1 - 2c & c & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & c & 1 - 2c & c \\ 0 & \cdots & 0 & c & 1 - 2c \end{pmatrix} \tag{3.26}$$

Here we define the norm of the matrix:

$$||A||_\infty = \max_{i=1,\dots,N} \sum_{j=1}^{N} |a_{ij}| \tag{3.27}$$

so,

$$||(I - cA)||_\infty = \max(|1 - 2c| + c, |1 - 2c| + 2c) = |1 - 2c| + 2c \tag{3.28}$$

if $c \leq \frac{1}{2}$, we have$|1 - 2c| + 2c = 1$ So,

$$||U^{j+1}||_\infty \leq ||U^j||_\infty + (\Delta t) \max_{0 \leq j \leq M} ||F^j||_\infty$$

$$||U^j||_\infty \leq ||U^0||_\infty + j(\Delta t) \max_{0 \leq j \leq M} ||F^j||_\infty$$

$$\leq ||U^0||_\infty + T \max_{0 \leq j \leq M} ||F^j||_\infty$$

With the help of lax theorem, the explicit Euler scheme is convergent.

□

3.5 The Stability of the Implicit Euler Scheme

Let $\lambda := \frac{v\Delta t}{\Delta x^2}$, the reminder could be expressed as

$$
\begin{aligned}
\varepsilon =& -\lambda u(x_{n+1}, t_j) + (1+2\lambda)u(x_n, t_j) - \lambda u(x_{n-1}, t_j) - u(x_n, t_{j-1}) - \Delta t f(x_n, t_j) \\
=& -\lambda \left[u(x_n, t_j) + \Delta x \frac{\partial u}{\partial x}(x_n, t_j) + \frac{(\Delta x)^2}{2!} \frac{\partial^2 u}{\partial x^2}(x_n, t_j) + \frac{(\Delta x)^3}{3!} \frac{\partial^3 u}{\partial x^3}(x_n, t_j) + \frac{(\Delta x)^4}{4!} \frac{\partial^4 u}{\partial x^4}(\theta_1, t_j) \right] \\
& + (1+2\lambda)u(x_n, t_j) - \lambda \left[u(x_n, t_j) + \Delta x \frac{\partial u}{\partial x}(x_n, t_j) + \frac{(\Delta x)^2}{2!} \frac{\partial^2 u}{\partial x^2}(x_n, t_j) + \frac{(\Delta x)^3}{3!} \frac{\partial^3 u}{\partial x^3}(x_n, t_j) + \frac{(\Delta x)^4}{4!} \frac{\partial^4 u}{\partial x^4}(\theta_2, t_j) \right] \\
& - \left[u(x_n, t_j) - \Delta x \frac{\partial u}{\partial x}(x_n, t_j) + \frac{(\Delta x)^2}{2!} \frac{\partial^2 u}{\partial x^2}(x_n, \xi) \right]
\end{aligned}
$$

Where

$$
\theta_1 \in [x_n, x_{n+1}], \quad \theta_2 \in [x_{n-1}, x_n], \quad \xi \in [t_{j-1}, t_j]
$$

Since $\frac{\partial u}{\partial t} = v \frac{\partial^2}{\partial x^2}$, we have thus

$$
\varepsilon = -v \frac{\partial^4}{\partial x^4} \left[\frac{(\Delta x)^4}{4!} \frac{\partial^4 u}{\partial x^4}(\theta_1, t_j) + \frac{(\Delta x)^4}{4!} \frac{\partial^4 u}{\partial x^4}(\theta_2, t_j) \right] - \frac{(\Delta t)^2}{2!} \frac{\partial^2 u}{\partial x^2}(x_n, \xi)
$$

The result follows:

$$
|\varepsilon| = \Delta t \cdot O((\Delta x)^2 + \Delta t) \tag{3.29}
$$

References

[Bre11] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer New York, 2011.

[Çin11] Erhan Çinlar. *Probability and Stochastics*. Springer New York, Feb. 2011. 558 pp.

[Eva10] Lawrence Evans. *Partial Differential Equations*. 2nd ed. American Mathematical Society, Mar. 2010.

[Fol99] Gerald B. Folland. *Real Analysis*. 2nd ed. John Wiley & Sons, Mar. 1999. 406 pp.

[Gos20] François Gosle. *Distributions, analyse de Fourier, équations aux dérivées partielles*. Ecole Polytechnique, Dec. 2020.

[Hab12] Richard Haberman. *Applied partial differential equations with fourier series and boundary value problems*. 5th ed. Addison-Wesley, 2012.

[Hul17] John C. Hull. *Options, Futures, and Other Derivatives*. 10th ed. PEARSON, Jan. 2017. 896 pp.

[Lax02] Peter D. Lax. *Functional Analysis*. John Wiley & Sons, Mar. 2002. 604 pp.

[Lax07] Peter D. Lax. *Linear Algebra and Its Applications*. 2nd ed. John Wiley & Sons, Aug. 2007. 394 pp.

[Luc16] Hervé Le Dret; Brigitte Lucquin. *Partial Differential Equations: Modeling, Analysis and Numerical Approximation*. Springer International Publishing, 2016.

[Mal08] Stephane Mallat. *A Wavelet Tour of Signal Processing*. 3rd ed. Elsevier Science Publishing Co Inc, Dec. 2008. 832 pp.

[Nef00] Salih N. Neftci. *An Introduction to the Mathematics of Financial Derivatives*. 3rd ed. Elsevier Science & Techn., June 2000. 527 pp.

[QSS06] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*. 2, illustrated. Texts in Applied Mathematics. Springer Berlin Heidelberg, 2006.

[Sch01] Michelle Schatzman. *Analyse numérique : une approche mathématique*. 2nd ed. Paris: Dunod, 2001.

[Sha03] Elias M. Stein; Rami Shakarchi. *Fourier Analysis: An Introduction*. PRINCETON UNIV PR, Apr. 2003. 328 pp.

[Shr10] Steven Shreve. *Stochastic Calculus for Finance II*. Springer New York, Dec. 2010. 572 pp.

[Shr98] Ioannis Karatzas; Steven E. Shreve. *Brownian Motion and Stochastic Calculus*. 2nd ed. Springer New York, 1998.

[Zil21] Matthieu Bonnivard; Adina Ciomaga; Alessandro Zilio. “Méthodes numériques pour les EDO et les EDP”. Notes de cours M1 Mathématiques. 2021.