FINITE ELEMENT METHODS FOR PARABOLIC EQUATIONS

LONG CHEN

As a model problem of general parabolic equations, we shall consider the following heat equation and study corresponding finite element methods

$$\left\{\begin{array}{rcll} u_t - \Delta u &=& f & \text{in} \quad \Omega \times (0,T), \\ u &=& 0 & \text{on} \quad \partial\Omega \times (0,T), \\ u(\cdot,0) &=& u_0 & \text{in} \quad \Omega. \end{array}\right. \tag{1}$$

Here $u = u(x,t)$ is a function of spatial variable $x \in \Omega \subset \mathbb{R}^n$ and time variable $t \in (0,T)$. The ending time T could be $+\infty$. The Laplace operator Δ is taking with respect to the spatial variable. For the simplicity of exposition, we consider only homogenous Dirichlet boundary condition and comment on the adaptation to Neumann and other type of boundary conditions. Besides the boundary condition on $\partial\Omega$, we also need to assign the function value at time $t = 0$ which is called initial condition. For parabolic equations, the boundary $\partial\Omega \times (0,T) \cup \Omega \times \{t = 0\}$ is called the parabolic boundary. Therefore the initial condition can be also thought as a boundary condition of the space-time domain $\Omega \times (0,T)$.

1. VARIATIONAL FORMULATION AND ENERGY ESTIMATE

We multiply a test function $v \in H_0^1(\Omega)$ and apply the integration by part to obtain a variational formulation of the heat equation (1): given an $f \in L^2(\Omega) \times (0,T]$, for any $t > 0$, find $u(\cdot,t) \in H_0^1(\Omega), u_t \in L^2(\Omega)$ such that

$$(u_t, v) + a(u,v) = (f,v), \quad \text{for all } v \in H_0^1(\Omega). \tag{2}$$

where $a(u,v) = (\nabla u, \nabla v)$ and $(\cdot,\cdot)$ denotes the L^2-inner product.

We then refine the weak formulation (2). The right hand side could be generalized to $f \in H^{-1}(\Omega)$. Since Δ map $H_0^1(\Omega)$ to $H^{-1}(\Omega)$, we can treat $u_t(\cdot,t) \in H^{-1}(\Omega)$ for a fixed t. We thus introduce the Sobolev space for the time dependent functions

$$L^q(0,T;W^{k,p}(\Omega)) := \{u(x,t) \,|\, \|u\|_{L^q(0,T;W^{k,p}(\Omega))} := \left(\int_0^T \|u(\cdot,t)\|_{k,p}^q \,\mathrm{d}t\right)^{1/q} < \infty\}.$$

Our refined weak formulation will be: given $f \in L^2(0,T;H^{-1}(\Omega))$ and $u_0 \in H_0^1(\Omega)$, find $u \in L^2(0,T;H_0^1(\Omega))$ and $u_t \in L^2(0,T;H^{-1}(\Omega))$ such that

$$\left\{\begin{array}{rcll} \langle u_t, v\rangle + a(u,v) &=& \langle f, v\rangle, & \forall v \in H_0^1(\Omega), \text{ and } t \in (0,T) \ a.e. \\ u(\cdot,0) &=& u_0 & \end{array}\right. \tag{3}$$

where $\langle\cdot,\cdot\rangle$ is the duality pair of $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. We assume equation (3) is well posed. For the existence and uniqueness of the solution $[u, u_t]$, we refer to [3]. In most places, we shall still use the formulation (2) and assume $f, u_t \in L^2(\Omega)$ so that the duality pair is realized by the L^2 inner product $(\cdot,\cdot)$.

Remark 1.1. The topology for the time variable should also be treat in L^2 sense. But in (2) and (3) we still pose the equation point-wise (almost everywhere) in time. In particular, one has to justify the point value $u(\cdot,0)$ does make sense for an L^2 type function which can be proved by the regularity theory of the heat equation. □

To easy the stability analysis, we treat t as a parameter and the function $u = u(x,t)$ as a mapping

$$u : [0,T] \to H_0^1(\Omega),$$

defined as

$$u(t)(x) := u(x,t) \quad (x \in \Omega, 0 \le t \le T).$$

With a slight abuse of notation, we still use $u(t)$ to denote the map. The norm $\|u(t)\|$ or $\|u(t)\|_1$ is taken with respect to the spatial variable and thus becomes a function of time.

We then introduce the operator

$$\mathcal{L} : L^2(0,T;H_0^1(\Omega)) \to L^2(0,T;H^{-1}(\Omega)) \times L^2(\Omega)$$

as

$$\begin{aligned}(\mathcal{L}u)(\cdot,t) &= \partial_t u - \Delta u \ \text{ in } H^{-1}(\Omega), \ \text{for } t \in (0,T] \ a.e.\\ (\mathcal{L}u)(\cdot,0) &= u(\cdot,0).\end{aligned}$$

Then the equation (3) can be written as

$$\mathcal{L}u = [f, u_0].$$

Here we explicitly include the initial condition u_0. The spatial boundary condition is build into the space $H_0^1(\Omega)$. In most places, when it is clear from the context, we also use $\mathcal{L}$ for the differential operator only.

We shall prove several stability results of $\mathcal{L}$ which are known as energy estimates in [5].

Theorem 1.2 (Energy estimates for the heat equation)**.** *Suppose $[u, u_t]$ is the solution of (2) and $u_t \in L^2(0,T;L^2(\Omega))$, then for $t \in (0,T]$ $a.e.$*

$$\|u(t)\| \le \|u_0\| + \int_0^t \|f(s)\| \,\mathrm{d}s \tag{4}$$

$$\|u(t)\|^2 + \int_0^t |u(s)|_1^2 \,\mathrm{d}s \le \|u_0\|^2 + \int_0^t \|f(s)\|_{-1}^2 \,\mathrm{d}s, \tag{5}$$

$$|u(t)|_1^2 + \int_0^t \|u_t(s)\|^2 \,\mathrm{d}s \le |u_0|_1^2 + \int_0^t \|f(s)\|^2 \,\mathrm{d}s. \tag{6}$$

Proof. The solution is defined via the action of all test functions. The art of the energy estimate is to choose an appropriate test function to extract desirable information.

We first choose $v = u$ to obtain

$$(u_t, u) + a(u,u) = (f,u).$$

We manipulate these three terms as:

- $(u_t,u) = \displaystyle\int_\Omega \frac{1}{2}(u^2)_t = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 = \|u\|\frac{\mathrm{d}}{\mathrm{d}t}\|u\|$;
- $a(u,u) = |u|_1^2$;
- $|(f,u)| \le \|f\|\|u\|$ or $|(f,u)| \le \|f\|_{-1}|u|_1 \le \dfrac{1}{2}(\|f\|_{-1}^2 + |u|_1^2)$.

The inequality (4) is an easy consequence of the following inequality

$$\|u\|\frac{\mathrm{d}}{\mathrm{d}t}\|u\| \le \|f\|\|u\|.$$

From

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 + |u|_1^2 \le \frac{1}{2}(\|f\|_{-1}^2 + |u|_1^2),$$

we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u\|^2 + |u|_1^2 \le \|f\|_{-1}^2.$$

Integrating over $(0,t)$, we obtain (5).

The last energy estimate (6) can be proved similarly by choosing $v = u_t$ and left as an exercise. □

From (5), we can obtain the stability of the operator $\mathcal{L}$

$$\mathcal{L} : L^2(0,T;H_0^1(\Omega)) \to L^2(0,T;H^{-1}(\Omega)) \times L^2(\Omega)$$

as

$$\|u\|^2_{L^2(0,T;H_0^1(\Omega))} \le \|u_0\|^2 + \|f\|^2_{L^2(0,T;H^{-1}(\Omega))}.$$

Since the equation is posed a.e for t, we could also obtain the maximum-norm estimate in time. For example, (4) can be formulated as

$$\|u\|_{L^\infty(0,T;L^2(\Omega))} \le \|u_0\| + \|f\|_{L^1(0,T;L^2(\Omega))},$$

and (6) implies

$$\|u\|_{L^\infty(0,T;H_0^1(\Omega))} \le |u_0|_1 + \|f\|_{L^2(0,T;L^2(\Omega))}.$$

Exercise 1.3. Prove the energy estimate

$$\|u\|^2 \le e^{-\lambda t}\|u_0\|^2 + \int_0^t e^{-\lambda(t-s))}\|f\|_{-1}^2\,\mathrm{d}s, \tag{7}$$

where $\lambda = \lambda_{\min}(-\Delta) > 0$. The estimate (7) shows that the effect of the initial data is exponential decay.

2. Finite element methods: semi-discretization in space

2.1. **Semi-discretization in space.** Let $\{\mathcal{T}_h, h \to 0\}$ be a quasi-uniform family of triangulations of Ω and $\mathbb{V}_h \subset H_0^1(\Omega)$ be a finite element space based on $\mathcal{T}_h$. The semi-discretized finite element method is: given $f \in \mathbb{V}_h' \times (0,T], u_{0,h} \in \mathbb{V}_h \subset H_0^1(\Omega)$, find $u_h \in L^2(0,T;\mathbb{V}_h)$ such that

$$\left\{\begin{array}{rcll} (\partial_t u_h, v_h) + a(u_h, v_h) &=& \langle f, v_h\rangle, & \forall v_h \in \mathbb{V}_h, t \in \mathbb{R}^+. \\ u_h(\cdot,0) &=& u_{0,h} & \end{array}\right. \tag{8}$$

The scheme (8) is called semi-discretization since u_h is still a continuous (indeed differential) function of t. The initial condition u_0 is approximated by $u_{0,h} \in \mathbb{V}_h$.

Take $\mathbb{V}_h$ as the linear finite element space as an example. We can expand $u_h = \sum_{i=1}^N u_i(t)\varphi_i(x)$, where φ_i is the standard hat basis at the vertex x_i for $i = 1, \cdots, N$, the number of interior nodes, and the corresponding coefficient $u_i(t)$ now is a function of time t. The solution u_h can be computed by solving an ODE system

$$\boldsymbol{M}\dot{\boldsymbol{u}} + \boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}, \tag{9}$$

where $\boldsymbol{u} = (u_1, \cdots, u_N)^\intercal$ is the coefficient vector, $\boldsymbol{M}, \boldsymbol{A}$ are the mass matrix and the stiffness matrix, respectively, and $\boldsymbol{f} = (f_1, \cdots, f_N)^\intercal$ with $f_i = (f, \varphi_i)$.

When the linear finite element is used, one can use three vertices quadrature rule i.e.

$$\int_\tau g(x)\,\mathrm{d}x \approx \frac{1}{3}\sum_{i=1}^{3} g(x_i)|\tau|.$$

Then the mass matrix becomes diagonal $M = \operatorname{diag}(m_1, \cdots, m_N)$. This is known as the mass lumping. For 2-D uniform grids, $m_i = h^2$ and A is the five point stencil discretization of $-\Delta$. Therefore (9) can be interpret as a rescaled finite difference discretization at each vertex and the ODE system (9) can be solved efficiently by mature ODE solvers.

2.2. **Setting for the error analysis.** We shall apply our abstract error analysis developed in *Unified Error Analysis* to estimate the error $u - u_h$ in certain norms. The setting is

- $\mathscr{X} = L^2(0,T;H_0^1(\Omega))$, and $\|u\|_{\mathscr{X}} = \left(\int_0^T |u(t)|_1^2\,\mathrm{d}t\right)^{1/2}$
- $\mathscr{Y} = L^2(0,T;H^{-1}(\Omega))$, and $\|f\|_{\mathscr{Y}} = \left(\int_0^T \|f(t)\|_{-1}^2\,\mathrm{d}t\right)^{1/2}$
- $\mathscr{X}_h = L^2(0,T;\mathbb{V}_h)$, and $\|u_h\|_{\mathscr{X}_h} = \left(\int_0^T |u_h(t)|_1^2\,\mathrm{d}t\right)^{1/2}$
- $\mathscr{Y}_h = L^2(0,T;\mathbb{V}_h')$, and $\|f_h\|_{\mathscr{Y}_h} = \left(\int_0^T \|f_h(t)\|_{-1,h}^2\,\mathrm{d}t\right)^{1/2}$. Recall the dual norm
$$\|f_h\|_{-1,h} = \sup_{v_h\in\mathbb{V}_h} \frac{\langle f_h, v_h\rangle}{|v_h|_1}, \quad \text{for } f_h \in \mathbb{V}_h'.$$
- $I_h = R_h(t) : H_0^1(\Omega) \to \mathbb{V}_h$ is the Ritz-Galerkin projection, i.e., $R_h u \in \mathbb{V}_h$ such that
$$a(R_h u, v_h) = a(u, v_h), \quad \forall v_h \in \mathbb{V}_h.$$
- $\Pi_h = Q_h(t) : H^{-1}(\Omega) \to \mathbb{V}_h'$ is the projection
$$\langle Q_h f, v_h\rangle = \langle f, v_h\rangle, \quad \forall v_h \in \mathbb{V}_h.$$
- $P_h : \mathscr{X}_h \to \mathscr{X}$ is the natural inclusion
- $\mathcal{L} : \mathscr{X} \to \mathscr{Y} \times L^2(\Omega)$ is $\mathcal{L}u = \partial_t u - \Delta u, \mathcal{L}u(\cdot,0) = u(\cdot,0)$, and $\mathcal{L}_h = \mathcal{L}|_{\mathscr{X}_h} : \mathscr{X}_h \to \mathscr{Y}_h \times \mathbb{V}_h$.

We summarize the setting in the following diagram

$$\begin{array}{ccc} \mathscr{X} & \xrightarrow{L} & \mathscr{Y} \\ \downarrow{\scriptstyle R_h} & & \downarrow{\scriptstyle Q_h} \\ \mathscr{X}_h & \xrightarrow{L_h} & \mathscr{Y}_h \end{array}$$

which is not commutative and the difference is the consistency error $\|Q_h L u - L_h R_h u\|_{\mathscr{Y}_h}$.

The discrete equation we are solving is

$$\begin{aligned} \mathcal{L}_h u_h &= Q_h f, \quad \text{in } \mathbb{V}_h', \forall t \in (0,T] \ a.e. \\ u_h(\cdot,0) &= u_{0,h}. \end{aligned}$$

2.3. **Stability.** Adapt the proof of the energy estimate for $\mathcal{L}$, we can obtain similar stability results for $\mathcal{L}_h$. The proof is almost identical and thus skipped here.

Theorem 2.1 (Energy estimate for finite element discretization)**.** *Suppose u_h satisfy $\mathcal{L}_h u_h = f_h, u_h(\cdot,0) = u_{0,h}$, then*

$$\|u_h(t)\| \le \|u_{0,h}\| + \int_0^t \|f_h(s)\| \,\mathrm{d}s \tag{10}$$

$$\|u_h(t)\|^2 + \int_0^t |u_h(s)|_1^2 \,\mathrm{d}s \le \|u_{0,h}\|^2 + \int_0^t \|f_h(s)\|_{-1,h}^2 \,\mathrm{d}s, \tag{11}$$

$$|u_h(t)|_1^2 + \int_0^t \|\partial_t u_h(s)\|^2 \,\mathrm{d}s \le |u_{0,h}|_1^2 + \int_0^t \|f_h(s)\|^2 \,\mathrm{d}s. \tag{12}$$

Note that in the energy estimate (11), the dual norm $\|\cdot\|_{-1}$ is replaced by a weaker one $\|\cdot\|_{-1,h}$ since we can apply the inequality

$$\langle f_h, u_h\rangle \le \|f_h\|_{-1,h}|u_h|_1.$$

The weaker norm $\|\cdot\|_{-1,h}$ can be estimated by

$$\|f\|_{-1,h} \le \|f\|_{-1} \le C\|f\|.$$

2.4. **Consistency.** Recall that the consistency error is $\|Q_h L u - L_h R_h u\|_{\mathscr{Y}_h}$. The choice of $I_h = R_h$ simplifies the consistency error analysis.

Lemma 2.2 (Error equation)**.** *For the semi-discretization, we have the error equation*

$$\mathcal{L}_h(u_h - R_h u) = Q_h(I - R_h)u_t,\ t > 0,\ \text{in } \mathbb{V}_h', \tag{13}$$

$$(u_h - R_h u)(\cdot,0) = u_{0,h} - R_h u_0. \tag{14}$$

Proof. Let $A = -\Delta$. By our definition of consistency, the error equation is: for $t > 0$

$$\mathcal{L}_h(R_h u - u_h) = \mathcal{L}_h R_h u - Q_h \mathcal{L} u = \partial_t(R_h u - Q_h u) + (A R_h u - Q_h A u).$$

The desired result then follows by noting that $AR_h = Q_h A$ in $\mathbb{V}_h'$ and $Q_h R_h = R_h$. □

The error equation (13) holds in $\mathbb{V}_h'$ which is a weak topology. The motivation to choose $I_h = R_h$ is that in this weak topology

$$\langle AR_h u, v_h\rangle = (\nabla R_h u, \nabla v_h) = (\nabla u, \nabla v) = \langle Au, v_h\rangle = \langle Q_h A u, v_h\rangle,$$

or simply in operator form

$$AR_h = Q_h A.$$

This technique is firstly proposed by Wheeler [6].

Apply the stability to the error equation, we obtain the following estimate on the discrete error $R_h u - u_h$.

Theorem 2.3 (Stability of discrete error)**.** *The solution u_h of (8) satisfy the following error estimate*

$$\|R_h u - u_h\| \le \|u_{0,h} - R_h u_0\| + \int_0^t \|Q_h(I - R_h)u_t\| \,\mathrm{d}s \tag{15}$$

(16)

$$\|R_h u - u_h\|^2 + \int_0^t |(u_h - R_h u)|_1^2 \,\mathrm{d}s \le \|u_{0,h} - R_h u_0\|^2 + \int_0^t \|Q_h(I - R_h)u_t\|_{-1,h}^2 \,\mathrm{d}s,$$

$$|R_h u - u_h|_1^2 + \int_0^t \|\partial_t(R_h u - u_h)\|^2 \,\mathrm{d}s \le |u_{0,h} - R_h u_0|_1^2 + \int_0^t \|Q_h(I - R_h)u_t\|^2 \,\mathrm{d}s. \tag{17}$$

We then estimate the two terms $u_{0,h} - R_h u_0$ and $Q_h(I - R_h)u_t$ involved in these error estimates. We use the linear element as an example since it is the most commonly choice. The first issue is on the choice of $u_{0,h}$. An optimal one is obviously $u_{0,h} = R_h u_0$ so that no error coming from approximation of the initial condition. However, this choice requires the inversion of a stiffness matrix which is not cheap. A simple choice would be the nodal interpolation, i.e. $u_{0,h}(x_i) = u_0(x_i)$ or any other choice with optimal approximation property

$$\|u_{0,h} - R_h u_0\| \leq \|u_0 - u_{0,h}\| + \|u_0 - R_h u_0\| \lesssim h^2 \|u_0\|_2, \tag{18}$$

and similarly

$$|u_{0,h} - R_h u_0|_1 \lesssim h \|u_0\|_2.$$

According to (7) in Exercise 1.3, the effect of the initial boundary error will be exponentially decay to zero as t goes to infinity. So in practice, we can choose the simple nodal interpolation.

On the estimate of the second term, assume $u_t \in H^2(\Omega)$ and H^2-regularity result hold for Poisson equation (for example, the domain is smooth or convex), then

$$\|Q_h(I - R_h)u_t\| \leq \|(I - R_h)u_t\| \lesssim h^2 \|u_t\|_2. \tag{19}$$

The negative norm can be bounded by the L^2-norm as

$$\|Q_h(I - R_h)u_t\|_{-1,h} \leq \|Q_h(I - R_h)u_t\|_{-1} \leq C\|Q_h(I - R_h)u_t\| \lesssim h^2 \|u_t\|_2.$$

When using quadratic and above polynomial, we can prove a stronger estimate for the negative norm and will be discussed in Section 2.6.

2.5. **Convergence.** The convergence of the discrete error comes $R_h u - u_h$ from the stability and consistency.

Theorem 2.4 (Convergence of the discrete error)**.** *Suppose the solution u to (3) satisfying $u_t \in L^2(0,T;H^2(\Omega))$ and the H^2-regularity holds for the Poisson equation. Let u_h be the solution of (8) with $u_{0,h}$ satisfying (18). We then have*

$$|R_h u - u_h|_1 + \|R_h u - u_h\| \leq Ch^2 \left(\|u_0\|_2 + \int_0^t \|u_t\|_2 \, ds \right). \tag{20}$$

To estimate the true error $u - u_h$, we need the approximation error estimate of the projection R_h; see *Introduction to Finite Element Methods*

$$h^{-1}\|u - R_h u\| + |u - R_h u|_1 \leq Ch\|u\|_2.$$

Theorem 2.5 (Convergence of the discretization error)**.** *Suppose the solution u to (3) satisfying $u_t \in L^2(0,T;H^2(\Omega))$. Then the solution u_h of (8) with $u_{0,h}$ having optimal approximation property (18) satisfy the following optimal order error estimate:*

$$h^{-1}|u - u_h|_1 + \|u - u_h\| \leq Ch^2 \left(\|u_0\|_2 + \int_0^t \|u_t\|_2 \, ds \right). \tag{21}$$

2.6. ***Superconvergence and error estimate in the maximum norm.** The error $e_h = R_h u - u_h$ satisfies the evolution equation (13) with $e_h(0) = 0$ with the chose $u_{0,h} = R_h u_0$

$$\partial_t e_h + A_h e_h = \tau_h, \tag{22}$$

where $\tau_h := Q_h \partial_t (R_h u - u)$. Therefore

$$e_h(t) = \int_0^t \exp(-A_h(t-s))\tau_h ds.$$

Due to the smoothing effect of the semi-group $e^{-A_h t}$, we have the following estimate. Here we follow the work by Garcia-Archilla and Titi [4].

Lemma 2.6 (Smoothing property of the heat kernel)**.** *For $\tau_h \in \mathbb{V}_h$, we have*

$$\max_{0\le t\le T}\left\|\int_0^t \exp(-A_h(t-s))A_h\tau_h ds\right\| \le C|\log h| \max_{0\le t\le T}\|\tau_h\|. \tag{23}$$

Proof. Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimal and maximal eigenvalue of A_h. Then it is easy to check

$$\left\|e^{-A_h(t-s)}A_h\right\| \le \begin{cases} \lambda_{\max}e^{-\lambda_{\max}(t-s)} & \text{if } (t-s)\le \lambda_{\max}^{-1},\\ (t-s)^{-1} & \text{if } \lambda_{\max}^{-1}\le (t-s)\le \lambda_{\min}^{-1},\\ \lambda_{\min}e^{-\lambda_{\min}(t-s)} & \text{if } (t-s)\ge \lambda_{\min}^{-1}. \end{cases}$$

Note that $\lambda_{\min} = \mathcal{O}(1)$ and $\lambda_{\max} \le Ch^{-2}$. We get

$$\max_{0\le t\le T}\left\|\int_0^t e^{-A_h(t-s)}A_h\tau_h ds\right\| \le C|\log h| \max_{0\le t\le T}\|\tau_h\|.$$

□

Theorem 2.7 (Superconvergence in H^1-norm)**.** *Suppose the solution u to (3) satisfying $u_t \in L^\infty(0,T;H^2(\Omega))$. Let u_h be the solution of (8) with $u_{0,h} = R_h u_0$. Then*

$$\max_{0\le t\le T}|R_h u - u_h|_1 \le C|\log h|\, h^2 \max_{0\le t\le T}\|u_t\|_2. \tag{24}$$

When $u_t \in L^2(0,T;H^2(\Omega))$, then

$$|R_h u - u_h|_1 \le Ch^2\left(\int_0^t \|u_t\|_2^2\, ds\right)^{1/2}. \tag{25}$$

Proof. We multiply $A_h^{1/2}$ to (22) and apply Lemma 2.6 to get

$$\begin{aligned} |e_h(T)|_1 = \|A_h^{1/2}e_h(T)\| &= \left\|\int_0^T e^{-A_h(T-s)}A_h^{1/2}\tau_h ds\right\| \\ &\le \left\|\int_0^T e^{-A_h(T-s)}A_h\, \mathrm{d}s\right\| \max_{0\le t\le T}\|A_h^{-1/2}\tau_h\| \\ &\le C|\log h|h^2 \max_{0\le t\le T}\|u_t\|_2. \end{aligned}$$

In the last step, we have used the fact $\|A_h^{-1/2}\tau_h\| = \|\tau_h\|_{-1,h} \le \|\tau_h\|$.

To get (25), we use the energy estimate (17). □

Since the optimal convergent rate for $|u - R_h u|_1$ or $|u - u_h|_1$ is only first order, the second order error estimate (24) and (25) are called superconvergence.

To control the maximum norm, we use the discrete embedding result (for 2-D only)

$$\|R_h u - u_h\|_\infty \le C|\log h||R_h u - u_h|_1,$$

and the error estimate of R_h in the maximum norm

$$\|u - R_h u\|_\infty \le C|\log h|h^2\|u\|_{2,\infty},$$

to obtain the following result.

Theorem 2.8 (Maximum norm estimate for linear element in two dimensions)**.** *Suppose the solution u to (3) satisfying $u \in L^\infty(0,T;W^{2,\infty})$ and $u_t \in L^2(0,T;W^{2,\infty})$ or $u_t \in L^\infty(0,T;W^{2,\infty}(\Omega))$. Let u_h be the solution of (8) with $u_{0,h} = R_h u_0$. Then in two dimensions*

$$\|(u-u_h)(t)\|_\infty \leq C|\log h|h^2 \left[\|u\|_{2,\infty} + \left(\int_0^t \|u_t\|_2^2\, ds\right)^{1/2}\right], \tag{26}$$

$$\max_{0\leq t\leq T}\|(u-u_h)(t)\|_\infty \leq C|\log h|^2 h^2 \max_{0\leq t\leq T}\left[\|u(t)\|_{2,\infty} + \|u_t(t)\|_2\right]. \tag{27}$$

For high order elements, we could get superconvergence in L^2 norm. Let us define the order of the polynomial as the degree plus 1, which is the optimal order when measuring the approximation property in L^p norm. For example, the order of the linear polynomial is 2. When the order of polynomial r is bigger than 3 (,i.e., quadratic and above polynomial), we can prove a stronger estimate for the negative norm

$$\|u - R_h u\|_{-1} \leq Ch^{r+1}\|u\|_r. \tag{28}$$

Using the technique in Lemma 2.6 and Theorem 2.7, we have the following estimate on the L^2 norm.

Theorem 2.9 (Superconvergence in L^2-norm for high order elements)**.** *Suppose the solution u to (3) satisfying $u_t \in L^\infty(0,T;H^r(\Omega))$. Let u_h be the solution of (8) with $u_{0,h} = R_h u_0$. Then, for $r \geq 3$,*

$$\max_{0\leq t\leq T}\|(R_h u - u_h)(t)\| \leq C|\log h|h^{r+1}\max_{0\leq t\leq T}\|(u_t)(t)\|_r^2. \tag{29}$$

When $u_t \in L^2(0,T;H^r(\Omega))$, then

$$\|R_h u - u_h\| \leq Ch^{r+1}\left(\int_0^t \|u_t\|_r^2\, ds\right)^{1/2}. \tag{30}$$

Proof. From (22), we have

$$\begin{aligned}
\|e_h(T)\| &= \left\|\int_0^T e^{-A_h(T-s)}\tau_h ds\right\| \\
&\leq \left\|\int_0^T e^{-A_h(T-s)}A_h^{-1/2}\,\mathrm{d}s\right\| \max_{0\leq t\leq T}\|A_h^{-1/2}\tau_h\| \\
&\leq C|\log h|\|\tau_h\|_{-1}.
\end{aligned}$$

In the second step, we have used the estimate

$$\left\|\int_0^T e^{-A_h(T-s)}A_h^{-1/2}\,\mathrm{d}s\right\| \leq C|\log h|,$$

which can be proved by the estimate $\|e^{-A_h(T-s)}A_h^{-1/2}\| \leq (T-s)^{-1}$

To prove (30), we simply use the energy estimate (16) and (28. □

In 1-D, using the inverse inequality $\|R_h u - u_h\|_\infty \leq h^{-1/2}\|R_h u - u_h\|$, we could obtain the superconvergence in the maximum norm

$$\|R_h u - u_h\|_\infty \leq Ch^{r+1/2}\left(\int_0^t \|u_t\|_r^2\, ds\right)^{1/2}. \tag{31}$$

Again using the inverse inequality $\|u_h - R_h u\|_\infty \le Ch^{-1}\|u_h - R_h u\|$, the superconvergence of L^2 norm, and the maximum norm estimate of Ritz-Galerkin projection, for $r \ge 3$, $\|u - R_h u\|_\infty \le Ch^r \|u\|_{r,\infty}$, we can improve the maximum norm error estimate.

Theorem 2.10 (Maximum norm estimate for high order elements in two dimensions)**.** *Suppose the solution u to (3) satisfying $u \in L^\infty(0,T;W^{2,\infty})$ and $u_t \in L^2(0,T;W^{2,\infty})$ or $u_t \in L^\infty(0,T;W^{2,\infty}(\Omega))$. Let u_h be the solution of (8) with $u_{0,h} = R_h u_0$. Then in two dimensions and for $r \ge 3$*

$$\|(u - u_h)(t)\|_\infty \le Ch^r \left[\|u\|_{r,\infty} + \left(\int_0^t \|u_t\|_r^2 \, ds \right)^{1/2} \right]. \tag{32}$$

3. Finite element methods: semi-discretization in time

In this section, we consider the semi-discretization in time. We first discretize the time interval $(0,T)$ into a uniform grid with size $\delta t = T/N$ and denoted by $t^n = n\delta t$ for $n = 0, \dots N$.

3.1. Low order schemes.

A continuous function in time will be interpolated into a vector by $f^n := (I^n f)(\cdot, t^n) = f(\cdot, t^n)$. Recall that $A = -\Delta : H_0^1 \to H^{-1}$. Below we list three low order schemes in operator form.

- *Forward Euler Method*: $u^0 = u_0$
$$\frac{u^n - u^{n-1}}{\delta t} + Au^{n-1} = f^{n-1}.$$
- *Backward Euler Method*: $u^0 = u_0$
$$\frac{u^n - u^{n-1}}{\delta t} + Au^{n} = f^{n}.$$
- *Crank-Nicolson Method*: $u^0 = u_0$
$$\frac{u^n - u^{n-1}}{\delta t} + A(u^n + u^{n-1})/2 = f^{n-1/2}.$$

Note that these equations hold in $H^{-1}(\Omega)$ sense. Taking Crank-Nicolson as an example, the equation reads as

$$\frac{1}{\delta t}(u^n - u^{n-1}, v) + \frac{1}{2}(\nabla u^n + \nabla u^{n-1}, \nabla v) = (f^{n-1/2}, v) \quad \text{for all } v \in H_0^1. \tag{33}$$

We now study the stability of these schemes. We rewrite the Backward Euler method as

$$(I + \delta t A)u^n = u^{n-1} + \delta t f^n.$$

Since A is SPD, $\lambda_{\min}(I + \delta t A) \ge 1$ and consequently, $\lambda_{\max}((I + \delta t A)^{-1}) \le 1$. This implies the L^2 stability

$$\|u^n\| \le \|u^{n-1}\| + \delta t \|f^n\| \le \|u^0\| + \sum_{k=1}^n \delta t \|f^k\|. \tag{34}$$

The stability (35) is the discrete counter part of (4): discretize the integral $\int_0^{t_n} \|f\| \, \mathrm{d}s$ by a Riemann sum.

Similarly one can derive the L^2 stability for the C-N scheme

$$\|u^n\| \le \|u^0\| + \sum_{k=1}^n \delta t \|f^{k-1/2}\|. \tag{35}$$

The integral $\int_0^{t_n} \|f\| \, \mathrm{d}s$ is approximated by the middle point rule.

Remark 3.1. For C-N method, the right hand side can be also chosen as $(f^n + f^{n-1})/2$. It corresponds to the trapezoid quadrature rule. For nonlinear problem $A(u)$, it can be $A((u^n + u^{n-1})/2)$ or $(A(u^n) + A(u^{n-1}))/2$. Which one to chose is problem dependent.

The energy estimate can be adapted to the semi-discretization in time easily. For example, we chose $v = (u^n + u^{n-1})/2$ in (33) to get

$$\frac{1}{2}\|u^n\|^2 - \frac{1}{2}\|u^{n-1}\|^2 + \delta t |u^{n-1/2}|_1^2 = \delta t (f^{n-1/2}, u^{n-1/2}),$$

which implies the counter part of (5)

$$\|u^n\|^2 + \sum_{k=1}^n \delta t |u^{k-1/2}|_1^2 \leq \|u^0\|^2 + \sum_{k=1}^n \delta t \|f^{k-1/2}\|_{-1}^2. \tag{36}$$

Exercise 3.2. Study the stability of the forward and backward Euler method.

We then study the convergence. We use C-N as a typical example. We apply the discrete operator $\mathcal{L}^n$ to the error $I^n u - u^n$

$$\mathcal{L}^n(I^n u - u^n) = \mathcal{L}^n I^n u - I^n \mathcal{L} u = \frac{u(\cdot, t^n) - u(\cdot, t^{n-1})}{\delta t} - \partial_t u(\cdot, t_{n-1/2}).$$

The consistency error is

$$\left| \frac{u(\cdot, t^n) - u(\cdot, t^{n-1})}{\delta t} - \partial_t u(\cdot, t_{n-1/2}) \right| \leq C(\delta t)^2. \tag{37}$$

By the stability result, we then get

$$\|I^n u - u^n\| \leq C t_n \delta t^2.$$

From the consistency error estimate (37), one can easily see the backward and forward Euler methods are only first order in time.

3.2. **High order discretizations in time.** We assume that $U \in C(0, T; H_0^1(\Omega))$ is a continuous piecewise q-th degree polynomial in time, that is, on the time interval $J_n := (t^{n-1}, t^n)$,

$$U|_{J_n}(x, t) = \sum_{j=0}^q (t - t^{n-1})^j u_j(x), \quad u_j(x) \in H_0^1(\Omega).$$

We denote $\mathcal{P}_q(J_n)$ as the set of such q-th degree polynomials on J_n, and define an operator $Q_{q-1}^n : C(0, T; H_0^1(\Omega)) \to \mathcal{P}_{q-1}(J_n)$ satisfying

$$\int_{J_n} \langle Q_{q-1}^n u, p \rangle \, dt = \int_{J_n} \langle u, p \rangle \, dt, \quad \forall p \in \mathcal{P}_{q-1}(J_n).$$

Then, the semi-discretization in time is to seek $U(t) \in \mathcal{P}_q(J_n)$ such that

$$\mathcal{L}^n U(t) := \partial_t U(t) + Q_{q-1}^n A U(t) = Q_{q-1}^n f(t), \quad \forall t \in J_n, \tag{38}$$

which is equivalent to the Petrov-Galerkin formulation

$$\int_{J_n} \langle \partial_t U, v \rangle \, dt + \int_{J_n} \langle AU, v \rangle \, dt = \int_{J_n} \langle f, v \rangle \, dt, \quad \forall v \in \mathcal{P}_{q-1}(J_n). \tag{39}$$

Note that the self-adjoint operator A is commute with the L^2 projection, i.e., $Q_{q-1}^n A = A Q_{q-1}^n$.

The initial condition is given by $U(0) = u_0$, and $U(t^{n-1})$ is obtained from the problem (39) on the previous time interval J_{n-1} for $n \geq 2$. For example, if $q = 1$, the solution $U(t)$ is a piecewise linear function and the test function v is piecewise constant in time. More exactly,

$$\partial_t U|_{J_n} = \frac{u^n - u^{n-1}}{\delta t}.$$

Thus, if we use the midpoint $t^{n-1/2}$ to compute $Q^n_{q-1}U(t)$ and $Q^n_{q-1}f(t)$ in the problem (38), then we arrive at the Crank-Nicolson method. Moreover, using left and right end points yields the Forward and Backward methods, respectively. For high order discretization, inside one time interval J_n, one needs to solve a mass matrix equation to get u^n. The naive basis for polynomial $\{(t - t^{n-1})^j,\ j = 0, \ldots, q\}$ is not friendly to the implementation. Instead one can use quadrature points, e.g., Gauss-Legendre points or Radau points; see [2].

3.3. **Stability.** We shall adapt the energy estimate to the discretization. Let us choose $v = Q_{q-1}U$. Then, we control the following terms:

- $\displaystyle\int_{J_n} \langle \partial_t U, Q^n_{q-1}U\rangle \, dt = \int_{J_n} \langle \partial_t U, U\rangle \, dt = \frac{1}{2}\left(\|U(t^n)\|^2 - \|U(t^{n-1})\|^2\right);$
- $\displaystyle\int_{J_n} \langle Q^n_{q-1}AU, Q^n_{q-1}U\rangle \, dt = \int_{J_n} \langle AQ^n_{q-1}U, Q^n_{q-1}U\rangle \, dt = \int_{J_n} |Q^n_{q-1}U|^2_1 \, dt;$
- $\displaystyle\int_{J_n} \langle f, Q^n_{q-1}U\rangle \, dt \leq \frac{1}{2}\int_{J_n} \|f\|^2_{-1} dt + \frac{1}{2}\int_{J_n} |Q^n_{q-1}U|^2_1 \, dt.$

We then obtain the energy estimate as follows.

Theorem 3.3 (Energy estimate for semi-discretization in time)**.** *Suppose U satisfies $\partial_t U + Q^n_{q-1}AU = Q^n_{q-1}f$, for $n \geq 1$, then*

$$\|U(t)\|^2 + \int_0^t |Q^n_{q-1}U|^2_1 \ \mathrm{d}s \leq \|U(0)\|^2 + \int_0^t \|f\|^2_{-1} \ \mathrm{d}s. \tag{40}$$

3.4. **Consistency.** Let $I^n : C(0, T; H^1_0(\Omega)) \to \mathcal{P}_q(J_n)$ be the operator of the Lagrange interpolation such that $I^n u(t^{n-1}) = u(t^{n-1})$, $I^n u(t^n) = u(t^n)$, and

$$\int_{J_n} \langle I^n u, p\rangle \, \mathrm{d}t = \int_{J_n} \langle u, p\rangle \, \mathrm{d}t, \quad \forall p \in \mathcal{P}_{q-2}(J_n).$$

Note that we use moments instead of point values to define the Lagrange interpolation.

It is straight forward to verify that $I^n u = u$ if $u \in \mathcal{P}_q(J_n)$, and

$$\left(\int_{J_n} \|u - I^n u\|^2 dt\right)^{1/2} \lesssim (\delta t)^{(q+1)} \|u\|_{H^{q+1}(t^{n-1}, t^n; L^2(\Omega))}.$$

From the definition of the consistency error, we have

$$\mathcal{L}^n(U - I^n u) = (Q^n_{q-1}\partial_t u - \partial_t I^n u) + (Q^n_{q-1}Au - Q^n_{q-1}AI^n u).$$

For the first time, the integration by part in terms of t provides

$$\begin{aligned}\int_{J_n} \langle Q^n_{q-1}\partial_t u - \partial_t I^n u, v\rangle \, dt &= \int_{J_n} \langle \partial_t u - \partial_t I^n u, v\rangle \, dt \\ &= \langle u - I^n u, v\rangle\big|^{t^n}_{t^{n-1}} - \int_{J_n} \langle u - I^n u, \partial_t v\rangle \, dt \\ &= 0,\end{aligned}$$

for every $v \in \mathcal{P}_{q-1}(J_n)$. The first identity holds because $\partial_t I^n u \in \mathcal{P}_{q-1}(J_n)$ so that $\partial_t I^n u = Q^n_{q-1}\partial_t I^n u$. The third one comes from the definition of the interpolation operator I^n.

Hence, we have the error equation

$$\mathcal{L}^n(U - I^n u) = Q^n_{q-1}Au - Q^n_{q-1}AI^n u =: \tau_n.$$

And

$$\begin{aligned}\int_{J_n} \|\tau^n\|^2\, dt &= \int_{J_n} \|Q^n_{q-1}(I - I^n)Au\|^2 \,\mathrm{d}t \\ &\leq \int_{J_n} \|(I - I^n)Au\|^2 \,\mathrm{d}t \\ &\lesssim (\delta t)^{2(q+1)} \|Au\|^2_{H^{q+1}(t^{n-1},t^n;L^2(\Omega))}.\end{aligned}$$

3.5. **Convergence.** Let $e^n = U - I^n u$. Clearly, $e^n \in \mathcal{P}_q(J_n)$ and $e^1(0) = U(0) - (I^1 u)(0) = u_0 - u(0) = 0$. Then, from the consistency result (40), we get

$$\|e^N(T)\|^2 \leq \sum_{n=1}^N \int_{J_n} \|\tau^n\|^2 dt \lesssim (\delta t)^{2(q+1)} \|Au\|^2_{H^{q+1}(0,T;L^2(\Omega))}.$$

Therefore, we obtain the following convergence rate

Theorem 3.4 (Energy estimate for semi-discretization in time)**.** *Suppose U satisfies $\partial_t U + Q^n_{q-1}AU = Q^n_{q-1}f$, for $n \geq 1$, then for any $t \in (0, T)$*

$$\|u(t) - U(t)\| \lesssim (\delta t)^{(q+1)} \|Au\|_{H^{q+1}(0,t;L^2(\Omega))}.$$

Refined analysis including a posteriori error analysis and superconvergence at nodal points can be found in [1].

REFERENCES

[1] G. Akrivis, C. Makridakis, and R. H. Nochetto. Galerkin and Runge–Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numerische Mathematik*, 118: 429-456, 2011. 12

[2] A.K. Aziz and P. Monk. Continuous finite elements in space and time for the heat equation. *Mathematics of Computation*, 52(186): 255-274, 1989. 11

[3] L. C. Evans. *Partial Differential Equations*. American Mathematical Society, 1998. 1

[4] B. GarcÍa-Archilla and E. S. Titi. Postprocessing the galerkin method: The finite-element case. *SIAM J. Numer. Anal.*, 37(2):470–499, 2000. 7

[5] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006. 2

[6] M. F. Wheeler. A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations. *SIAM J. Numer. Anal.*, 10:723–759, 1973. 5