J. Austral. Math. Soc. Ser. B 35(1993), 87–102

A FINITE ELEMENT METHOD FOR A DIFFUSION EQUATION WITH CONSTRAINED ENERGY AND NONLINEAR BOUNDARY CONDITIONS

AMIYA K. PANI[1]

(Received 22 July 1990; revised 17 July 1992)

Abstract

A finite element Galerkin method for a diffusion equation with constrained energy and nonlinear boundary condition is analysed and optimal error estimates in L^2 and L^∞–norms are derived. These results improve upon previously derived estimates by Cannon *et al.* [4].

1. Introduction

In this paper, we consider a finite element Galerkin approximation for the following diffusion equation with constrained energy and nonlinear boundary conditions:

Problem A. Find $u = u(x,t)$ such that

$$u_t = u_{xx}, \qquad (x,t) \in I \times (0,T], \quad I = (0,1), \tag{1}$$

$$u(x,0) = u_0(x), \qquad x \in I, \tag{2}$$

$$u_x(1,t) = g(t, u(1,t)), \qquad 0 \le t \le T, \tag{3}$$

$$M(t) = \int_0^b u(x,t)\,dx, \qquad 0 < b < 1. \tag{4}$$

Problem A models various physical phenomena. For example, in a heat

[1]Department of Mathematics, Indian Institute of Technology, Powai, Bombay–400076, India.

conduction problem, u denotes the temperature distribution in an horizontal bar of unit length (say) and $M(t)$ represents an average temperature of the region $0 < x < b$ at time t. We note that (2) is the initial temperature in the bar and the condition (3) at the end of the bar may be of Fourier–Boltzman type (i.e. heat transfer through radiation is allowed at the end $x = 1$).

For the case $g = g(t)$ and for sufficiently smooth u_0, g and M, the existence of a smooth solution u is proved using an equivalent Volterra integral formulation and fixed point arguments, (cf. Cannon [1]). In case the function $g = g(t, u(1, t))$ satisfies (even locally) Lipschitz continuity condition with respect to u, their analysis can be easily extended to prove the existence of a unique local solution.

Recently, Cannon *et al.* [4] have discussed both continuous and discrete time Galerkin approximations for Problem A with $g = g(t)$, and have obtained only *a priori* error estimates in $L^\infty(0, T; L^2(I))$, which is not optimal. In the present paper, we not only consider Problem A with nonlinear boundary conditions, but also obtain improved error estimates in $L^\infty(0, T; L^2(I))$–norm. Further, we derive optimal rates of convergence in $L^\infty(0, T; L^\infty(I))$ and in $L^\infty(0, T; H^1(I))$–norms. Finally, the discrete-time Galerkin method is analysed by using a dual quadrature rule for the integral in the temporal variable t in the right-hand side of (8). For a similar analysis, one may refer to Sloan and Thomée [5].

Here we make several assumptions on the smoothness of u, the initial datum u_0 and on the boundary function g.

Assumptions B. (i) There is a unique smooth solution u to the Problem A such that for $r \geq 1$

$$u \in W^{2,\infty}(0, T; L^\infty) \cap W^{1,2}(0, T; W^{r+1,\infty})$$

or

$$u \in W^{4,\infty}(0, T; L^\infty) \cap W^{1,2}(0, T; W^{r+1,\infty}).$$

Further, the solution u is bounded by some constant (say) K_1 in the above-mentioned spaces.

(ii) For simplicity, we assume that the boundary function g is smooth and bounded with derivatives by a constant (say) K_1.

(iii) The initial fuction u_0 is smooth as required by the analysis.

REMARK. Since we prove error estimates in $L^\infty(L^\infty)$–norm, the present analysis will still work for locally Lipschitz continous g (Thomée and Wahlbin [6] or Thomée [7]).

In the sequel, we shall use the following spaces. The space $W^{m,p}(I)$, $1 \leq p \leq \infty$ is the usual Sopbolev space and its norm is written as $\|.\|_{m,p}$. For $p = 2$, we write H^m instead of $W^{m,2}$ and denote its norm by $\|.\|_m$. By $L^p(H^m)$, we mean the spaces $L^p(0, T; H^m(I))$.

The paper is organised as follows: In Section 2, a weak formulation and Galerkin procedures are derived for Problem A. Section 3 is devoted to an auxiliary projection, and related approximations. For continuous-time Galerkin approximation, optimal estimates in $L^\infty(L^\infty)$, $L^\infty(L^2)$ and $L^\infty(H^1)$–norms are derived in Section 4. Further, in Section 5, a priori error estimates for discrete in time Galerkin methods are presented.

2. Weak formulation and Galerkin approximations

For a weak formulation of Problem A, we multiply both the sides of (1) by $v \in H^1$ and integrate by parts with respect to x to obtain

$$(u_t, v) + (u_x, v_x) = g(t, u(1))v(1) - u_x(0, t)v(0). \tag{5}$$

For the term u_x at $x = 0$, we differentiate (4) with respect to t and apply (1) to have

$$u_x(0, t) = u_x(b, t) - M(t). \tag{6}$$

As in Cannon *et al.* [4], we use a representation formula for the solution u to have a viable form for $u_x(b, t)$, namely,

$$\begin{aligned} u_x(b, t) = {} & 2\theta(b, t)u_0(0) - \{\theta(b-1, t) + \theta(b+1, t)\}\, u_0(1) \\ & + \int_0^t \{\theta(b-\xi, t) + \theta(b+\xi, t)\}\, u_{0,x}(\xi)\, d\xi \\ & -2\int_0^t \{\theta_{xx}(b, t-\tau)u(0, \tau) - \theta_{xx}(b-1, t-\tau)u(1, \tau)\}\, d\tau, \end{aligned} \tag{7}$$

where $\theta(x, t) = \sum_{-\infty}^{\infty} \kappa(x + 2j, t)$ and

$$\kappa(x, t) = \frac{1}{2\sqrt{\pi t}} \exp(-\frac{x^2}{4t}), \text{ for } t > 0.$$

For simplicity, write $\mathcal{D}(t)$ as the sum of the first three terms in the right-hand side of (7). On substitution of (6) and (7) in (5), we obtain

$$(u_t, v) + (u_x, v_x) = g(t, u(1))v(1) + (M'(t) - \mathcal{D}(t))v(0)$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

$$-2v(0)\int_0^t \{\theta_{xx}(b, t-\tau)u(0, \tau) - \theta_{xx}(b-1, t-\tau)u(1, \tau)\}\, d\tau, \quad (8)$$

with $u(x, 0) = u_0(x)$.

Galerkin Approximations. Let $S_h, 0 < h \leq 1$, be a family of finite dimensional subspaces of H^1 with the following properties:

(I) APPROXIMATION PROPERTY: For $v \in W^{m,p}(I)$, ($p = 2$ or ∞) there is a constant K_0 such that for $j = 0, 1$ and $1 \leq m \leq r + 1$

$$\inf_{\phi \in S_h} \|v - \phi\|_{W^{j,p}} \leq K_0 h^{m-j} \|v\|_{W^{m,p}}.$$

(II) INVERSE PROPERTY: For $\phi \in S_h$

$$\|\phi\|_{L^\infty} \leq K_0 h^{-1/2} \|\phi\|_{L^2}.$$

Now a continuous-time Galerkin approximation is defined as:
'Find $U : [0, t] \mapsto S_h$ such that

$$\begin{aligned}(U_t, V) + (U_x, V_x) &= g(t, U(1))V(1) + (M'(t) - \mathcal{D}(t))V(0) \\ &\quad - 2V(0)\int_0^t \{\theta_{xx}(b, t-\tau)U(0, \tau) - \theta_{xx}(b-1, t-\tau)U(1, \tau)\}\, d\tau, \\ &\qquad V \in S_h,\ t > 0 \qquad (9) \\ U(x, 0) &= \mathcal{Q}_h u_0(x),\end{aligned}$$

where $\mathcal{Q}_h$ is an appropriate projection of u_0 on to S_h to be defined later.'

Note that (9) is a system of nonlinear integrodifferential equations. So for a Lipschitz continuous g and a given initial function, the solution U exists at least locally. Since an *a priori* estimate in $L^\infty(0, T; L^\infty)$ is followed as a consequence of the error estimates, the solution U can be bounded independent of h for bounded u and hence U is uniquely continued up to T.

For a fully discrete scheme related to (9), we discretise the temporal variable in the following way. Let $N \in \mathcal{Z}, k = \Delta t = T/N, t_n = nk$, for $n = 0, 1, \ldots, N$ and $t_{n+\frac{1}{2}} = (n + \frac{1}{2})k$. Further let $\phi_n = \phi(t_n)$,

$$\bar{\partial}_t \phi_n = \frac{\phi_n - \phi_{n-1}}{k} \qquad \text{and} \qquad \phi_{n+1/2} = \frac{1}{2}(\phi_{n+1} + \phi_n).$$

For a backward Euler's method, we now apply a quadrature rule for the integral in the right-hand side of (9). For the integral $\int_0^{t_n} \phi(s)\, ds$, we consider

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

the following quadrature rule

$$\int_0^{t_n} \phi(s)\,ds \simeq \sum_{j=0}^{n-1} \omega_{nj}\phi_j,$$

where ω_{nj} are quadrature weights. The simplest quadrature rule which is consistent with $O(k)$ accuracy for the backward Euler's scheme is the rectangle quadrature rule in which the weights $\omega_{nj} = k$, for $0 \le j \le n$. But as in Sloan and Thomée [5] we use a dual quadrature rule, i.e. on the larger part of the time interval we use a higher-order quadrature rule like the trapezoidal rule with large time steps and in the remaining part we use the rectangle rule with time step k. The effect is quite significant in the sense that this discretisation for the integral term, while being consistent with the backward Euler method, is increasingly sparse as the time step k converges to zero. In this way both the memory requirements and the computational effort can be greatly reduced. Let $l(n)$ denote the largest nonzero integer such that $lk_1 < nk$. The choice of k_1 will be made later. Now split the interval $[0, nk]$ in to $[0, lk_1] \cup [lk_1, nk]$ and approximate the integral over the first interval by a trapezoidal rule with step length k_1 and that in the remaining part by a rectangle rule with step length k. Then

$$\int_0^{t_n} \phi(s)\,ds \simeq k_1\left[\frac{1}{2}\phi(0) + \phi(k_1) + \ldots + \phi((l-1)k_1)\right] + \left(\frac{1}{2}k_1 + k\right)\phi(lk_1) + k\big[\phi(lk_1 + k) + \ldots + \phi((n-1)k)\big].$$

When $l = 0$, the trapezoidal rule is omitted. For sufficiently regular ϕ, the truncation error is $O(k_1^2 + k_1 k)$ as $t_n \le T < \infty$. The required consistency is achieved if $k_1 = O(k^{1/2})$. For a given j and sufficiently large n, the weights ω_{nj} are independent of n and are bounded by a quantity which is independent of n, that is $\sum_{j=0}^{n-1} \omega_{nj} \le (2T + 1)$. Therefore, the integrals appearing in the right-hand side of (9) can be replaced by

$$2\int_0^{t_n} \theta_{xx}(b, t_n - \tau)U(0, \tau)\,d\tau \simeq 2\sum_{j=0}^{n-1} \omega_{nj}\theta_{xx}(b, t_n - t_j)U_j(0),$$

and

$$2\int_0^{t_n} \theta_{xx}(b-1, t_n - \tau)U(1, \tau)\,d\tau \simeq 2\sum_{j=0}^{n-1} \omega_{nj}\theta_{xx}(b-1, t_n - t_j)U_j(1).$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

The fully-discrete Galerkin approximation U_n, $n = 1, 2, \ldots, N$ is defined as a solution in S_h of the following equation

$$(\bar{\partial}_t U_n, V) + (U_{n,x}, V_x) = g(t_n, U_n)V(1) - V(0)(M'(t_n) - \mathcal{D}(t_n)) - V(0)\sum_{j=0}^{n-1} \omega_{nj}\big[\theta_{xx}(b, t_n - t_j)U_j(0) - \theta_{xx}(b-1, t_n - t_j)U_j(1)\big]. \quad (10)$$

In order to achieve higher-order accuracy, we can formulate the following variant of the Crank–Nicolson scheme. Discretise the problem (9) at $t = t_{n-\frac{1}{2}}$ and use the following quadrature rule

$$\int_0^{t_{n-\frac{1}{2}}} \phi(s)\, ds \simeq \sum_{j=0}^{n-1} \omega_{nj}\phi_j.$$

Let $l(n)$ be the largest nonnegative integer such that $2lk_1 < nk$. Then split the interval into $[0, 2lk_1] \cup [2lk_1, (n-1)k] \cup [(n-1)k, (n-1/2)k]$. On the first interval with larger step length k_1, we use Simpson's rule which is of $O(k_1^4)$, and the trapezoidal rule for the second, while on the third use the rectangle rule with step length $k/2$.

Finally, we obtain

$$\int_0^{t_{n-\frac{1}{2}}} \phi(s)\, ds \simeq \frac{1}{3}[\phi(0) + 4\phi(k_1) + 2\phi(2k_1) + \ldots + \phi((2l-1)k_1)] + \left(\frac{1}{3}k_1 + \frac{1}{2}k\right)\phi(2lk_1) + k\big[\phi(2lk_1 + k) + \ldots + \phi((n-1)k)\big].$$

As nk is bounded above by $T < \infty$, the truncation error is of order $O(k_1^4 + k_1k^2 + k^2)$, for a smooth function ϕ. Thus to maintain second-order accuracy, k_1 should be of order $O(k^{1/2})$. In this way the discrete-time Galerkin approximation is defined by

$$(\bar{\partial}_t U_n, V) + (U_{n-\frac{1}{2},x}, V_x) = g(t_{n-\frac{1}{2}}, U_{n-\frac{1}{2}})V(1) - V(0)(M'(t_{n-\frac{1}{2}}) - \mathcal{D}(t_{n-\frac{1}{2}})) - 2V(0)\sum_{j=0}^{n-1} \omega_{nj}\big[\theta_{xx}(b, t_{n-\frac{1}{2}} - t_j)U_j(0) - \theta_{xx}(b-1, t_{n-\frac{1}{2}} - t_j)U_j(1)\big]. \quad (11)$$

This yields a system of nonlinear algebraic equations, and for given initial datum it has at least a unique solution for small k. Since nonlinearity occurs

due to the presence of the boundary function g in (11) at each time level, a linearised modification for g can be achieved by replacing $U_{n-\frac{1}{2}}$ through EU_n, where $EU_n = \frac{3}{2}U_{n-1} - \frac{1}{2}U_{n-2}$. Although it preserves the second-order accuracy, one needs to evaluate U_1. One possible way to achieve this is to use a predictor-corrector method (c.f. Thomée [7]).

3. Auxiliary projection and error estimates

Define an auxiliary projection $\tilde{u} \in S_h$ of u through the following form:

$$((u - \tilde{u})_x, V_x) + \lambda(u - \tilde{u}, V) = 0, \qquad V \in S_h \tag{12}$$

and for some fixed positive number λ. The existence of a unique $\tilde{u} \in S_h$ follows directly from the Lax–Milgram Lemma.

Let $\mathcal{L}$ be a one-dimensional elliptic operator associated with (12) and be defined by

$$\mathcal{L}\phi = -\phi_{xx} + \lambda\phi.$$

Now for $\psi \in H^s$, $s \geq 0$, the following problem:

$$\mathcal{L}\phi = \psi, \qquad \phi_x(0) = \phi_x(1) = 0,$$

has a unique solution $\phi \in H^{s+2}$ which satisfies the regularity condition

$$\|\phi\|_{s+2} \leq K_2\|\psi\|_s.$$

Let $\eta = u - \tilde{u}$ be the error involved in the auxiliary projection (12). The estimates of η and η_t in different norms are now quite standard. These estimates involve the use of the Aubin-Nitsche duality argument.

Here, we only state the estimates without proof (Thomée [7]).

LEMMA 3.1. *There is a constant* $K_3 = K_3(K_0, K_1, K_2; \lambda)$ *such that for* $p = 2$ *or* ∞

$$\|\eta\|_{W^{j,p}} + \|\eta_t\|_{W^{j,p}} \leq K_3 h^{m-j}, \qquad j = -1, 0, 1;\ 1 \leq m \leq r + 1.$$

Further, we need the following estimates of η at the end points $x = 0, 1$ for our future use.

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

LEMMA 3.2. *There is a constant K_3 (say) such that for $x = 0, 1$*

$$|\eta(x)| \leq K_3 h^{2(m-1)}, \qquad 1 \leq m \leq (r+1).$$

Here K_3 is a generic constant.

PROOF. We examine only the case $x = 0$. The other case follows similarly. Now consider an auxiliary fuction ϕ such that

$$\mathcal{L}\phi = 0, \qquad x \in I,$$

$$\phi_x(0) = 1, \qquad \phi_x(1) = 0.$$

Multiply by η and integrate to obtain

$$\begin{aligned} -\eta(0) &= (\eta_x, \phi_x) + \lambda(\eta, \phi) \\ &= (\eta_x, \phi_x - \chi_x) + \lambda(\eta, \phi - \chi), \qquad \chi \in S_h. \end{aligned}$$

Therefore,

$$|\eta(0)| \leq K(K_0, K_1; \lambda) h^{2(m-1)} \|u\|_m \|\phi\|_m$$

and this completes the proof.

4. Error estimates for continuous case

Define the projection $\mathcal{Q}_h$ by

$$((u_0 - \mathcal{Q}_h u_0)_x, V_x) + \lambda(u_0 - \mathcal{Q}_h u_0, V) = 0, \qquad V \in S_h.$$

Clearly $U(x, 0) = \tilde{u}(x, 0)$. For our error estimates we need the following inequalities:

(i) For $a, b \geq 0$ and $\epsilon > 0$,

$$ab \leq \frac{\epsilon^p a^p}{p} + \frac{b^q}{\epsilon^q q}, \qquad \frac{1}{p} + \frac{1}{q} = 1, \qquad 1 < p < \infty.$$

(ii) For $\phi \in H^1(I)$, $\sup_{0 \leq x \leq 1} |\phi(x)| \leq \|\phi\|_1$.

(iii) For $\phi \in H^1(I)$ and for $0 \leq x \leq 1$, $|\phi(x)| \leq \|\phi\| + \sqrt{2}\|\phi\|^{\frac{1}{2}} \|\phi\|_1^{\frac{1}{2}}$.

(iv) For a fuction ψ on the time interval [0,t],

$$\int_0^t \left(\int_0^\tau |\psi(s)|^2 \, ds \right) d\tau \leq t \int_0^t |\psi(s)|^2 \, ds.$$

Let $\zeta = U - \tilde{u}$ and $e = u - U$, then $e = \eta - \zeta$. Below we derive some estimates for ζ.

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

THEOREM 4.1. *There exists a constant K_4 such that*

$$\|\zeta\|_{L^\infty(L^2)} + \|\zeta\|_{L^2(H^1)} \le K_4 h^{r+1},$$

holds for $r \ge 2$. Here K_4 is a generic constant depending on $K_0, K_1, K_2, K_3, \lambda$ and also on the bounds of θ.

PROOF. From (8), (9) and (12), it follows that

$$\begin{aligned}(\zeta_t, V) + (\zeta_x, V_x) = (\eta_t, V) - \lambda(\eta, V) + V(1)[g(t, U(1)) - g(t, u(1))] \\ -2V(0)\int_0^t \{\theta_{xx}(b, t-\tau)(\zeta(0) - \eta(0)) \\ -\theta_{xx}(b-1, t-\tau)(\zeta(1) - \eta(1))\}\, d\tau. \quad (13)\end{aligned}$$

Choose $V = \zeta$ in (13) to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|\zeta\|^2 + \|\zeta_x\|^2 &\le (\|\eta_t\|_{-1} + \lambda\|\eta\|_{-1})\|\zeta\|_1 + K(K_1; \|\theta_{xx}(l, \cdot)\|_{L^\infty})|\zeta(0)| \\ &\quad \times \int_0^t (|\zeta(0)| + |\zeta(1)| + |\eta(0)| + |\eta(1)|)\, d\tau \\ &= I_1 + I_2,\end{aligned}$$

for $l = b, b-1$. For I_1, use of inequality (i) yields

$$|I_1| \le \epsilon\|\zeta\|_1^2 + (\|\eta_t\|_{-1} + \lambda^2\|\eta\|_{-1})/(2\epsilon).$$

To estimate I_2, apply the inequalities (i)–(iv) to obtain

$$\begin{aligned}|I_1| \le K(K_1, \lambda, T, \epsilon; \|\theta_{xx}(l, \cdot)\|_{L^\infty})(|\eta(1)|^2 + \int_0^t (|\eta(0)|^2 + |\eta(1)|^2)\, d\tau \\ + 3\epsilon\|\zeta\|^2 + K(K_1, T, \epsilon; \|\theta_{xx}(l, \cdot)\|_{L^\infty}) \int_0^t \|\zeta\|_1^2\, d\tau.\end{aligned}$$

These estimates now yield

$$\begin{aligned}\frac{d}{dt}\|\zeta\|^2 + 2\|\zeta\|_1^2 \le 8\epsilon\|\zeta\|_1^2 + K(K_1, \lambda, T; \epsilon; \|\theta_{xx}(l, \cdot)\|_{L^\infty}) \\ \times \Big[(\|\zeta\|^2 + \int_0^t \|\zeta\|_1^2\, d\tau)(\|\eta_t\|_{-1}^2 + \|\eta\|_{-1}^2 + |\eta(1)|^2) \\ + \int_0^t (|\eta(0)|^2 + |\eta(1)|^2)\, d\tau\Big].\end{aligned}$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

Integrating with respect to t and using inequality (iv), it follows that

$$\|\zeta(t)\|^2+(2-8\epsilon)\int_0^t\|\zeta(\tau)\|_1^2 \le K(K_1,\lambda,T,\epsilon;\|\theta_{xx}(l,\cdot)\|_{L^\infty})\int_0^t\|\zeta(\tau)\|^2\,d\tau$$
$$+K(K_1,\lambda,T,\epsilon;\|\theta_{xx}(l,\cdot)\|_{L^\infty})$$
$$\times\int_0^t(\|\eta_t\|_{-1}^2+\|\eta\|_{-1}^2+|\eta(0)|^2+|\eta(1)|^2)\,d\tau.$$

Choose $\epsilon = 1/8$ and apply Gronwall's Lemma to obtain

$$\|\zeta(t)\|^2+\int_0^t\|\zeta\|_1^2\,d\tau$$
$$\le K(K_1,T,\lambda;\|\theta_{xx}(l,\cdot)\|_{L^\infty})\int_0^t(\|\eta_t\|_{-1}^2+\|\eta\|_{-1}^2+|\eta(0)|^2+|\eta(1)|^2)\,d\tau.$$

Now the estimates in Lemma 3.1 and in Lemma 3.2 complete the proof.

From the inverse property and the Theorem 4.1, we obtain the following estimates for ζ.

COROLLARY 4.2. *The following estimates*

$$\|\zeta\|_{L^\infty(L^\infty)} \le K_4h^{r+1} \qquad \textit{and} \qquad \|\zeta\|_{L^\infty(H^{-1}(I))} \le K_4h^{r+2}$$

hold for $r \ge 2$. *Here* K_4 *is a generic constant.*

We now make use of the triangle inequality and Theorem 4.1 to obtain the following estimates for e.

THEOREM 4.3. *Let* u *and* U *be respectively the solutions of* (8) *and* (9). *Then the following estimates in the error* $e = u - U$

$$\|e\|_{L^\infty(L^2)}+\|e\|_{L^2(H^1)}+\|e\|_{L^\infty(L^\infty)} \le K_4h^{r+1},$$

and

$$\|e\|_{L^\infty(H^{-1})} \le K_4h^{r+2}$$

hold for $r \ge 2$.

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

REMARKS. In case only an $L^\infty(L^2)$ estimate for the error e is needed, one may even consider piecewise-linear finite element spaces i.e., $r \geq 1$. Note that in Theorem 4.1, one may use the estimate $\|\eta_t\|\|\zeta\|$ for the term (η_t, ζ) instead of the estimate $\|\eta_t\|_{-1}\|\zeta\|_1$.

Below we shall examine the rate of convergence for the error e in the H^1–norm.

THEOREM 4.4. *There is a constant, again say K_4, such that*

$$\|e_t\|_{L^2(L^2)} + \|e\|_{L^\infty(H^1)} \leq K_4 h^{r+1}$$

holds for $r \geq 1$.

PROOF. Choose $V = \zeta_t$ in (13) and rewrite the resulting equation as

$$\begin{aligned}\|\zeta_t\|^2 + \frac{1}{2}\frac{d}{dt}\|\zeta\|_1^2 \leq &\left[(\eta_t, \zeta_t) - \lambda(\eta, \zeta_t) + \frac{1}{2}\frac{d}{dt}\|\zeta\|^2\right] \\ &+ \zeta_t(1)(g(t, U(1)) - g(t, u(1)) \\ &- 2\zeta_t(0)\int_0^t \Big[\theta_{xx}(b, t-\tau)(\zeta(0)-\eta(0)) \\ &\qquad - \theta_{xx}(b-1, t-\tau)(\zeta(1)-\eta(1))\Big] d\tau \\ = &\ I_1 + I_2 + I_3.\end{aligned}$$

For I_1, an use of inequality (i) yields

$$\left|\int_0^t I_1(\tau)\, d\tau\right| \leq K(\lambda; \epsilon)\left[\|\zeta(t)\|^2 + \int_0^t (\|\eta_t\|^2 + \|\eta\|^2)\, d\tau\right] + \epsilon \int_0^t \|\zeta_t\|^2\, d\tau.$$

To estimate I_2, apply inverse property and the inequalities (i)–(iv) to obtain

$$\left|\int_0^t I_2(\tau)\, d\tau\right| \leq K(K_0, K_1; \epsilon)h^{-1}\int_0^t (\|\zeta\|_1^2 + |\eta(1)|^2)\, d\tau + \epsilon\int_0^t \|\zeta_t\|^2\, d\tau.$$

Finally for the term I_3, we apply inverse property and the inequalities (i)–(iii) to have

$$\begin{aligned}\left|\int_0^t I_3\, d\tau\right| \leq &\ K(K_0, T, \epsilon; \|\theta_{xx}(l, \cdot)\|_{L^\infty})h^{-1}\int_0^t (|\eta(0)|^2 + |\eta(1)|^2 + \|\zeta\|_1^2)\, d\tau \\ &+ 2\epsilon\int_0^t \|\zeta_t\|^2\, d\tau.\end{aligned}$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

Altogether we obtain with $\epsilon = 1/8$,

$$\int_0^t \|\zeta\|^2 \, d\tau + \|\zeta\|_1^2 \leq K(K_0, K_1, \lambda, T; \|\theta_{xx}(l, \cdot)\|_{L^\infty}) \times \left[\|\zeta\|^2 + \int_0^t (\|\eta_t\|^2 + \|\eta\|^2) \, d\tau + h^{-1} \int_0^t (\|\zeta\|_1^2 + |\eta(0)|^2 + |\eta(1)|^2) \, d\tau \right].$$

Application of Lemmas 3.1, 3.2 and Theorem 4.1 now completes the proof.

REMARKS. From Theorem 4.3 the optimal error estimate for the error e in $L^\infty(L^\infty)$–norm is obtained for C^0– polynomial spline spaces of degree r with $r \geq 2$. This is due to the fact that the superconvergent result for ζ in $L^\infty(L^2)$–norm is used along with the inverse property. In order to obtain an optimal rate of convergence even for piecewise-linear finite element spaces. i.e. for $r \geq 1$, we modify the proof of Theorem 4.4 in the following way: first we obtain a superconvergence for ζ in H^1–norm and then use inequality (ii) as well as the triangle inequality to complete the proof. Now to estimate I_2, we note

$$g(t, U(1)) - g(t, u(1)) = -\tilde{g}_u(\eta(1) - \zeta(1)),$$

where $\tilde{g}_u = \dfrac{\partial g}{\partial u}(u(1) - \xi e(1))$, $0 \leq \xi \leq 1$, and rewrite I_2 as

$$I_2 = \frac{1}{2}\frac{d}{dt}(\zeta^2(1))\tilde{g}_u - \zeta_t(1)\tilde{g}_u\eta(1).$$

On integration by parts with respect to t, it follows that

$$\int_0^t I_2 \, d\tau = \left[\frac{1}{2}\zeta^2(1) - \zeta(1)\eta(1)\right]\tilde{g}_u + \int_0^t \zeta(1)\eta_t(1)\tilde{g}_u \, d\tau - \int_0^t \left[\frac{1}{2}\zeta(1) - \eta(1)\right]\zeta(1)\,(\tilde{g}_{ut} + \tilde{g}_{uu}(u_t - \xi e_t)) \, d\tau$$

and hence using inequalities (i)–(iii) it yields

$$\left|\int_0^t I_1(\tau) \, d\tau\right| \leq K(K_1; \epsilon)\left[\|\zeta(t)\|^2 + |\eta(1)|^2\right] + \epsilon\|\zeta(t)\|_1^2 + K(K_1, \|\zeta_t\|_{L^2(L^\infty)}; \epsilon) \int_0^t (\|\zeta\|_1^2 + |\eta_t(1)|^2 + |\eta(1)|^2) \, d\tau.$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

Here one may assume temporarily that $\|\zeta_t\|_{L^2(L^\infty)} \leq 1$, for sufficiently small h. This is in fact not a restriction, since using the inverse property,

$$\|\zeta_t\|_{L^2(L^\infty)} \leq K_0 h^{-1/2} \|\zeta_t\|_{L^2(L^2)}$$

and the latter estimate is $O(h^{r+1})$. Similarly, the estimate for the term I_3 can be obtained with only a change in the dependence of the constant. This time the constant may depend on $\|\theta_{xxt}(l, \cdot)\|_{L^\infty}$, $l = b, b-1$. The estimate of $|\eta_t(1)|$ is an easy consequence of Lemma 3.3. Now the rest of the analysis follows from Theorem 4.4. Hence we obtain

$$\|\zeta\|_{L^\infty(H^1)} \leq K_4 h^{r+1}, \qquad r \geq 1.$$

Finally, the optimal error estimate for e in $L^\infty(L^\infty)$ is obtained using the inequality (ii) and the triangle inequality.

5. Error estimates for fully discrete scheme

We shall begin with the backward Euler–Galerkin scheme (10). Note that for a given initial function U_0, (10) is a nonlinear system of algebraic equations and it has a unique solution for small k. Let $e_n = u(t_n) - U_n$ be the error at the time level t_n. Then we have the following error estimates.

THEOREM 5.1. *Under the regularity conditions* (i_1) *in the Assumptions B, the estimates*

$$\|e_n\|^2 + k \sum_{j=0}^{n} \|e_j\|_1^2 \leq K_4 (h^{2(r+1)} + k^2),$$

hold for $r \geq 1$ *and for small* k.

PROOF. As before, we write $e_n = \eta_n - \zeta_n$ with $\eta_n = u_n - \tilde{u}_n$ and $\zeta_n = U_n - \tilde{u}_n$, where $\tilde{u}_n$ is the auxiliary projection defined by the equation (12) at the time level $t = t_n$. Since the estimates for η at $t = t_n$ are known from Lemmas 3.1 and 3.2, it therefore remains to estimate ζ_n. Now from (8), (12) at $t = t_n$ and (10), it follows that

$$(\bar{\partial}_t \zeta_n, V) + ((\zeta_n)_x, V_x) =$$
$$(\bar{\partial}_t \eta_n, V) - \lambda(\eta_n, V) + (\sigma_n, V)$$

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

$$
\begin{aligned}
&-2V(0)\sum_{j=0}^{n-1}\omega_{nj}\left[\theta_{xx}(b,t_n-t_j)(\zeta_j(0)-\eta_j(0))\theta_{xx}(b-1,t_n-t_j)(\zeta_j(1)-\eta_j(1))\right]\\
&-2V(0)\left[\int_0^{t_n}\theta_{xx}(b,t_n-\tau)u(0,\tau)\,d\tau\sum_{j=0}^{n-1}\omega_{nj}\theta_{xx}(b,t_n-t_j)u_j(0)\right]\\
&+2V(0)\left[\int_0^{t_n}\theta_{xx}(b-1,t_n-\tau)u(0,\tau)\,d\tau\sum_{j=0}^{n-1}\omega_{nj}\theta_{xx}(b-1,t_n-t_j)u_j(1)\right]\\
&+V(1)\left[g(t_n,U_n(1))-g(t_n,u_n(1))\right]\\
&=(\bar\partial_t\eta_n,V)-\lambda(\eta_n,V)+(\sigma_n,V)+I_1^n+I_2^n+I_3^n+I_4^n,
\end{aligned}
\tag{14}
$$

where $\sigma_n = u_t(t_n) - \bar\partial_t u_n$. Choose $V = \zeta_n$ in (14) and note that

$$(\bar\partial_t\zeta_n,\zeta_n)\geq(\|\zeta_n\|^2-\|\zeta_{n-1}\|^2)/(2k).$$

For the term I_1^n, we have

$$|I_1^n|\leq K(K_1,T;\|\theta(l,\cdot)\|_{L^\infty})|\zeta_n(0)|\sum_{j=0}^{n-1}\left[|\zeta_j(0)|+|\zeta_j(1)|+|\eta_j(0)|+|\eta_j(1)|\right],$$

for $l = b, b-1$.

Further, each one of the two terms I_2^n and I_3^n consists of two parts, the one arising from the trapezoidal rule with steplength k_1, and the other from the rectangle rule. Therefore these two terms are bounded by

$$K(K_1,T;\|\theta_{ttxx}(l,\cdot)\|_{L^\infty})[k_1^2+k_1k]|\zeta(0)|,$$

for $l = b$ or $b-1$. Altogether, we obtain, on replacing $|\zeta(0)|\leq\|\zeta\|_1$ and using inequality (iii) in Section 4 for $|\zeta(1)|$ with Young's inequality, the result

$$
\begin{aligned}
&(\|\zeta_n\|^2-\|\zeta_{n-1}\|^2)+2\|\zeta_n\|_1^2\\
&\qquad\leq Kk\left[\|\bar\partial_t\eta_n\|^2+\|\eta_n\|^2+|\eta_n(1)|^2+\|\zeta_n\|^2+\|\sigma_n\|^2\right]\\
&\qquad\quad+Kk^2\sum_{j=0}^{n-1}\left[|\eta_j(0)|^2+|\eta_j(1)|^2+\|\zeta_j\|^2+\|\zeta_j\|_1^2\right]+8\epsilon k\|\zeta_n\|_1^2,
\end{aligned}
$$

where the constant K depends on $K_1, \lambda, T, \|\theta_{ttxx}(l,\cdot)\|_{L^\infty}; \epsilon$ for $l = b, b-1$. Summing up on n and using $k\sum_{j=1}^{n}\sum_{i=0}^{j-1}\phi_i\leq T\sum_{j=0}^{n-1}\phi_j$, it follows for appro-

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

priately chosen k and ϵ that

$$\|\zeta_n\|^2 + k\sum_{j=0}^{n} \|\zeta_j\|_1^2 \le Kk \sum_{j=0}^{n-1} [\|\zeta_j\|^2 + \sum_{i=0}^{j} \|\xi_i\|_1^2]$$
$$+Kk \sum_{j=0}^{n} \left[\|\bar{\partial}_t \eta_j\|^2 + \|\eta_j\|^2 + |\eta_j(1)|^2 + |\eta_j(0)|^2 + \|\sigma_j\|^2\right].$$

An application of Gronwall's Lemma along with the estimates for η yields

$$\|\zeta_n\|^2 + k\sum_{j=0}^{n} \|\zeta_j\|_1^2 \le K[h^{2(r+1)} + h^{4r} + k^2],$$

and this completes the proof.

For H^1 error estimate, we choose $V = k\bar{\partial}_t \zeta_n$ in equation (14). Using Lemmas 3.1, 3.2 and Theorem 5.1, it is easy to obtain the following results. As the method of analysis is quite similar to the previous Theorem, we state only the estimates without proof.

THEOREM 5.2. *Let the assumptions B be satisfied with regularity condition* (i_1). *Then there is a constant (say)* K_6 *such that*

$$k\sum_{j=0}^{n} \|\bar{\partial}_t e_j\|^2 + \|e_n\|_1^2 \le K_6[h^{2r} + k^2], \qquad n \ge 1,$$

holds for $r \ge 1$ *and for small* k. *Here* K_6 *is a generic constant.*

For second-order accuracy in time, we shall consider the Crank–Nicolson Galerkin scheme (11). The analysis is similar in spirit to Theorems 5.1 and 5.2 with added regularity conditions (i_2) and higher truncation errors. Therefore, we only state the results without proof.

THEOREM 5.3. *Let the solution* u *satisfy the regularity conditions* (i_2) *in the assumptions B.Then there is a constant* K_6 *such that for* $r \ge 1$ *and for small* k

$$\|e_n\|^2 + k\sum_{j=0}^{n} \|e_{j+1/2}\|_1^2 \le K_6[h^{2(r+1)} + k^4]$$

and

$$k\sum_{j=1}^{n} \|\bar{\partial}_t e_j\|^2 + \|e_n\|_1^2 \le K_6[h^{2r} + k^4]$$

hold.

https://doi.org/10.1017/S0334270000007281 Published online by Cambridge University Press

REMARKS. In case we do not use the dual quadrature rules, we need only less smoothness for the solution u in the direction of t both for backward Euler scheme and Crank–Nicolson scheme.

The present analysis can be modified without any difficulty, for the problem with nonhomogeneous right-hand side i.e.with a source term say $f = f(x, t)$ and with $b = b(t)$ (cf. Cannon and van der Hoek [4]). But we understand the computational complexities while dealing with a variable domain of integration and the double integral involving the function f.

Acknowledgments

The author greatfully acknowledges the support of the Centre for Mathematical Analysis, ANU (Canberra), Australia in the preparation of this material. Further, he would like to thank Professor Vidar Thomée and Dr Bob Anderssen for their helpful suggestions. Finally the suggestions of the referees to improve the earlier version of the manuscript are highly appreciated.

References

[1] J. R. Cannon, "The solution of the heat equation subject to the specification of energy", *Quart. Appl. Math* **21** (1963) 155–160.

[2] J. R. Cannon, "The one-dimensional heat equation", in *Encyclopadia of Mathematics and its Applications 23*, (Addison-Wesley, CA, 1984).

[3] J. R. Cannon, S. Perezesteva and J. van der Hoek, "Galerkin procedure for the diffusion equation subject to the specification of mass", *SIAM J. Numer. Anal.* **24** (1987) 499–515.

[4] J. R. Cannon and J. van der Hoek, "Diffusion subject to the specification of mass", *J. Math. Anal. Appl.* **115** (1986) 517–529.

[5] I. H. Sloan and V. Thomée, "Time discretization of an integro-dfferential equation of parabolic type", *SIAM J. Numer. Anal.* **23** (1986) 1052–1061.

[6] V. Thomée, "Galerkin finite element methods for parabolic problems", in *Lecture Notes in Mathematics 1054*, (Springer-Verlag, Berlin, 1984).

[7] V. Thomée and L. B. Wahlbin, "On Galerkin methods in semilinear parabolic problems", *SIAM J. Numer. Anal* **12** (1975) 378–389.