SIAM J. SCI. STAT. COMPUT.
Vol. 11, No. 1, pp. 1-32, January 1990

A METHOD FOR THE SPATIAL DISCRETIZATION OF PARABOLIC EQUATIONS IN ONE SPACE VARIABLE*

ROBERT D. SKEEL† AND MARTIN BERZINS‡

Abstract. This paper is concerned with the design of a spatial discretization method for polar and nonpolar parabolic equations in one space variable. A new spatial discretization method suitable for use in a library program is derived. The relationship to other methods is explored. Truncation error analysis and numerical examples are used to illustrate the accuracy of the new algorithm and to compare it with other recent codes.

Key words. Galerkin method, Petrov-Galerkin method, polar coordinates, singular boundary value problems, parabolic equations, method of lines

AMS(MOS) subject classifications. 65M05, 65M10, 65M15, 65M20

C. R. classification. G.1.8

1. Introduction. The aim of this paper is to describe and to give evidence in support of a new spatial discretization for the method of lines solution of parabolic equations in one space variable. The intent is to provide a method that is suitable for use in a general-purpose library program, such as the D03P section of the NAG library.

Ordinary and parabolic partial differential equations in one space variable x often have a singularity due to the use of polar cylindrical or spherical coordinates. Because of their common occurrence, some of the differential equation software—such as the PDEONE code of Sincovec and Madsen [16] and the D03P** code of Dew and Walsh [8]—treat these singularities explicitly in order to reduce accuracy problems that arise for coefficients like $1/x^m$ when x is near zero. Nonetheless, methods that have been proposed (see Eriksson and Thomée [9] for references) or implemented do not obtain the same (local order of) accuracy for the case $m=1$ (and sometimes $m=2$) as they do for $m=0$.

The method we propose is a simple piecewise nonlinear Galerkin/Petrov-Galerkin method that is second-order accurate in space. (It supersedes the method proposed by Skeel [17].) The case $m=1$ involves the use of the logarithm function, which is probably the only accurate way to model the logarithmic behavior that can be present in the solution. A code based on a variant of the proposed method has already been included as part of the SPRINT package of Berzins, Dew, and Furzeland [4] (which is available from M. Berzins). The method we propose here has been implemented and will be distributed in the next or next but one release of the D03P (parabolic equations) section of the NAG library.

A derivation of the method is given in § 2. Rather than simply announcing our choice of trial space, test space, and inner product, we give a more concrete finite-element derivation that motivates these choices. Section 3 gives the concise Galerkin

* Received by the editors January 28, 1987; accepted for publication (in revised form) October 18, 1988.

† Department of Computer Science, University of Illinois at Urbana-Champaign, 240 Digital Computer Laboratory, 1304 West Springfield Avenue, Urbana, Illinois 61801. The research of this author was supported in part by the U.K. Science Research Council and in part by the U.S. Department of Energy under grant DEAC0276ER02383.

‡ Department of Computer Studies, University of Leeds, Leeds LS2 9JT, United Kingdom. The research of this author was partly supported by Shell Research Ltd., Thornton Research Centre, P.O. Box 1, Chester, United Kingdom.

formulation of the method. The primary purpose of the remainder of the paper is to supply the details of the evidence in favor of the new method. Section 4 discusses the variant of the proposed method used in SPRINT as well as other competing methods, § 5 considers the time integrator for the system of ordinary differential equations, and § 6 describes the results of numerical testing. Finally, § 7 summarizes our conclusions. Also, there is an Appendix that contains the error analysis that supports the claim of good accuracy for the proposed methods.

2. Derivation of the spatial discretization method. Consider the system of quasilinear partial differential equations (PDEs):

$$D(x, t, u, u_x)u_t = x^{-m}(x^m g(x, t, u, u_x))_x + f(x, t, u, u_x) \tag{1}$$

for $a \leqq x \leqq b$ where D is a diagonal matrix with nonnegative entries and m is nonnegative. If $m > 0$, we require $a \geqq 0$. The cases $m = 1$ and $m = 2$ represent cylindrical and spherical polar coordinates, respectively. Boundary conditions are

$$p^i(x, t, u) + q^i(x, t)g^i(x, t, u, u_x) = 0 \quad \text{at } x = a, b$$

for $i = 1, 2, \cdots$, NPDE. If $m \geqq 1$ and $a = 0$, the boundedness of the solution near $x = 0$ implies $g^i(x, t, u, u_x) = 0$ at $x = 0$. The problem class defined by (1) has been deliberately chosen so as to have recognizable flux and source terms and to have the possibility of recognizable Cartesian polar and spherical polar coordinates. The general form of the flux function $g(\cdots)$ is designed to permit conservative differencing of both advective and diffusive flux terms. For notational convenience we work with a single PDE and omit the argument t.

We consider a spatial mesh $a = x_0 < x_1 < \cdots < x_J = b$. Because continuity of the solution $u(x)$ and of the (negative) flux (per unit area) $v(x) := g(x, u(x), u_x(x))$ is demanded for all x, we use as unknowns values of u and v at meshpoints. It is assumed that meshpoints are placed at discontinuities in the PDE so that the problem is smooth within each subinterval.

The SPRINT implementation also has two additional features of interest. The first is that the problem class is extended to include the coupled ODE (ordinary differential equation)/PDE problems considered by Schryer [15]. The only restrictions are that the problem class must be linear in the PDE time derivative and that time derivatives must not appear in the flux term. The second feature is that an optional remeshing facility has been provided, based on the padded monitor function of Kautsky and Nichols [11]. The monitor function is chosen by the user and typically depends on the flux or on the solution and its space derivatives. These two features are discussed further by Berzins and Furzeland [5], [6].

We seek a difference scheme that

(i) Uses only one evaluation of D, f and g per subinterval;

(ii) Is elegant;

(iii) Is as accurate as possible for the special case

$$g(x, u, u_x) = G(x, u)u_x;$$

and

(iv) Leads to an explicit system of ODEs (which is desirable for reasons given in § 5). Ideally *local* second-order accuracy is desired, meaning that the contribution to the global error from subinterval $[x_{j-1}, x_j]$ is $O(h_j^3)$ where $h_j := x_j - x_{j-1}$. It is clear what this means for linear problems where the global error is a superposition of propagated local errors; more care would be required to define this idea for nonlinear problems. Global second-order accuracy is a weaker property meaning that the local errors are $O(h^3)$ on the average.

The special problem we consider can be expressed as the two first-order PDEs:

$$u_x = H(x, u)v, \tag{2}$$

$$(x^m v)_x = x^m Q(x, u, u_x, u_t) \tag{3}$$

where

$$H(x, u) := 1/G(x, u)$$

and

$$Q(x, u, u_x, u_t) := D(x, u, u_x)u_t - f(x, u, u_x).$$

The derivation begins by focusing on a typical subinterval, or element, that we denote by $[\alpha, \beta]$. The length $h := \beta - \alpha$ and the midpoint $\gamma := (\alpha + \beta)/2$. The first section considers quadrature and lumping, the second treats interpolation, and the third assembles the element equations into difference equations by eliminating the unknown values of $v(x)$ at meshpoints.

Two separate cases are considered depending on whether or not the $1/x^m$ singularity is present. The "regular case" occurs when $m = 0$ or $a > 0$ and is treated with a Galerkin method. The "singular case" occurs when $m \geqq 1$ and $a = 0$, and a Petrov-Galerkin method is used, in which a special trial space is chosen. In the singular case the presence of a discontinuity has an important effect, and for this reason the error analysis takes into account the location of the first discontinuity from the left, which we denote by c (so that on $[0, c]$ the problem is smooth). In the regular case when $m \geqq 1$ let $c := a$. The factor $1/c$ shows up in some error bounds, but we try to avoid it as much as possible in the belief that there are problems for which $c \ll b$ (such as an annulus with a very small center hole).

The choice of scheme was motivated by an error analysis, which in turn was influenced by numerical experiments. The detailed analysis, given in the Appendix, shows that the proposed scheme is second-order accurate for both the regular and singular cases outlined above. We obtain error bounds depending on problem parameters P consisting of bounds on various derivatives on the *open subintervals* of $[a, b]$. For the singular case we also use $\bar{P}$, which consists of bounds on the same derivatives on *all* of $[0, c]$, including interior meshpoints. Most of the error analysis for $m \geqq 1$ applies to any real $m \geqq 1$ not just $m = 1$ and $m = 2$.

2.1. Quadrature and lumping. Integrating (3) and then (2), we obtain

$$u(\beta) = u(\alpha) + \int_\alpha^\beta \frac{H(\cdots)}{x^m} \left\{ \alpha^m v(\alpha) + \int_\alpha^x y^m Q(\cdots)\, dy \right\} dx, \tag{4}$$

and interchanging α and β, we obtain

$$u(\alpha) = u(\beta) - \int_\alpha^\beta \frac{H(\cdots)}{x^m} \left\{ \beta^m v(\beta) - \int_x^\beta y^m Q(\cdots)\, dy \right\} dx. \tag{5}$$

The basic idea is that of the finite-element method. We form discrete approximations of these two equations, thus obtaining an equation for each of $v(\alpha)$ and $v(\beta)$ in terms of $u(\alpha)$ and $u(\beta)$. Then the boundary conditions and the continuity of $v(x)$ at meshpoints can be used to eliminate these unknown values of $v(x)$, thus obtaining difference equations for the values of $u(x)$ at meshpoints.

Approximations to (4) and (5) are developed by using the values

$$H_0 := H(\xi, u(\xi)), \qquad D_0 := D(\xi, u(\xi), u_x(\xi)),$$

$$f_0 := f(\xi, u(\xi), u_x(\xi))$$

at a point ξ inside the interval $[\alpha, \beta]$. (Error analysis suggests to us that separate choices of ξ for each of H, D, and f would, at best, offer only slight improvement in the accuracy of the scheme. Moreover, this would complicate the user interface, requiring either the provision of more than one PDE routine or the evaluation of different PDE functions at different space points in one call.) The question of approximating $u(\xi)$ and $u_x(\xi)$ is deferred to § 2.2. If we assume that with one degree of freedom H, D, and f are best approximated by constants, then we get

$$u(\beta) = u(\alpha) + \int_\alpha^\beta \frac{H_0}{x^m}\left\{\alpha^m v(\alpha) + \int_\alpha^x y^m (D_0 u_t(\alpha) - f_0)\, dy\right\} dx + \sigma + H_0\tau_\alpha \tag{6}$$

and

$$u(\alpha) = u(\beta) - \int_\alpha^\beta \frac{H_0}{x^m}\left\{\beta^m v(\beta) - \int_x^\beta y^m (D_0 u_t(\beta) - f_0)\, dy\right\} dx - \sigma + H_0\tau_\beta \tag{7}$$

where we have used $u_t(\alpha)$ in (6) and $u_t(\beta)$ in (7) to get an explicit system of ODEs in time and the quadrature errors σ, τ_α, and τ_β are defined by

$$\int_\alpha^\beta H(\cdots) v(x)\, dx =: H_0 \int_\alpha^\beta v(x)\, dx + \sigma, \tag{8}$$

$$\int_\alpha^\beta \frac{1}{x^m}\int_\alpha^x y^m Q(\cdots)\, dy\, dx =: \int_\alpha^\beta \frac{1}{x^m}\int_\alpha^x y^m (D_0 u_t(\alpha) - f_0)\, dy\, dx + \tau_\alpha, \tag{9}$$

and

$$\int_\alpha^\beta \frac{1}{x^m}\int_x^\beta y^m Q(\cdots)\, dy\, dx =: \int_\alpha^\beta \frac{1}{x^m}\int_x^\beta y^m (D_0 u_t(\beta) - f_0)\, dy\, dx + \tau_\beta. \tag{10}$$

Note that (7) is undefined for $\alpha = 0$ and $m \geqq 1$. For this and other reasons we develop an additional equation and its discretization. Combining (4) and (5), we obtain

$$\beta^m v(\beta) = \alpha^m v(\alpha) + \int_\alpha^\beta x^m Q(\cdots)\, dx, \tag{11}$$

and combining (6) and (7), we obtain

$$\beta^m v(\beta) =: \alpha^m v(\alpha) + \int_\alpha^\beta x^m \{D_0(u_t(\alpha)\psi_\alpha(x) + u_t(\beta)\psi_\beta(x)) - f_0\}\, dx + \tau \tag{12}$$

where

$$\psi_\beta(x) := \begin{cases} \displaystyle\int_\alpha^x y^{-m}\, dy \Big/ \int_\alpha^\beta y^{-m}\, dy, & \alpha > 0 \text{ or } m = 0, \\ 1, & \alpha = 0 \text{ and } m \geqq 1, \end{cases}$$

and

$$\psi_\alpha(x) := 1 - \psi_\beta(x).$$

In the special case $\alpha = 0$ and $m \geqq 1$, we use (12) instead of (7). Note that, otherwise, the quadrature error τ satisfies

$$\tau := \frac{\tau_\alpha + \tau_\beta}{\int_\alpha^\beta y^{-m}\, dy}.$$

2.1.1. Choice of quadrature point. The choice of ξ gives us a second degree of freedom in the numerical quadrature of (6) and (7). We choose this point to obtain the most accuracy under the assumption that with two degrees of freedom H, D, and f are best approximated by linear polynomials.

The choice of ξ should take into account the behavior of $v(x)$. Consideration of simple examples, such as $H \equiv 1$ and $Q \equiv 1$, suggests that $v(x)$ behaves like x^{-m} in the regular case and like x in the singular case. Hence, we choose

$$\xi = \gamma_{-\mu}$$

where

$$\mu := \begin{cases} m & \text{regular case,} \\ -1 & \text{singular case,} \end{cases}$$

and where γ_k denotes the Gauss point for the weight function x^k, that is,

$$\int_\alpha^\beta (x - \gamma_k) x^k \, dx := 0.$$

Note that

$$\gamma_k = \gamma + \frac{k}{12} \frac{h^2}{\gamma} + O\left(\frac{h^4}{\gamma^3}\right)$$

and in particular $\gamma_1 = \gamma + h^2/(12\gamma)$ exactly.

This choice of quadrature point is justified in the Appendix where in Theorem 5 it is shown that with $\xi = \gamma_{-m}$ in the regular case the error

$$|\sigma| \leqq h^3 C(P)$$

where C denotes a generic constant (meaning that each occurrence of the symbol "C" is a possibly different symbol), and after this it is shown that other choices of ξ are theoretically inferior if a is small but positive. In Theorem 6 it is shown for the singular case that with $\xi = \gamma_1$ the error

$$|\sigma| \leqq \gamma h^3 C(\bar{P})/(m+1),$$

and after this it is shown that this bound is not possible if $\xi = \gamma$. The additional factor of γ in the error bound indicates an extra order of accuracy near the origin $x = 0$. Recall that because discontinuities in H, D, f and the initial conditions are permitted at meshpoints, the bound is good only on the interval $[0, c]$. On the remainder of the interval, $[c, b]$, Theorem 6 states that for the singular case with $\xi = \gamma_1$ the error

$$|\sigma| \leqq h^3 C(P),$$

which is no worse than the bound for $\xi = \gamma_{-m}$. Finally, it is not possible to choose ξ to be the same for both regular and singular cases because $\gamma_{-m} < \gamma < \gamma_1$.

The effect of τ_α and τ_β on the global error is not straightforward. This is discussed further in the Appendix. It is shown there that the error τ in (12) as an approximation to (11) is important but not so important so as to dictate the precise choice of ξ. However, in the singular case when $m > 1$, error analysis and numerical experiments show that the use of $\xi = \gamma_1$ is better than either the midpoint γ or the point γ_m suggested previously by Bakker [1].

2.2. Interpolation. Each of the functions $G(\cdots)$, $D(\cdots)$, and $f(\cdots)$ will have to be evaluated at ξ; and since our derivation assumes the availability of $u(x)$ only at

$x=\alpha$ and $x=\beta$, it will be necessary to construct an interpolant. Note that $u_x = Hv$ and recall that $v(x)$ behaves like x^{-m} in the regular case and like x in the singular case. This suggests for $\alpha \leqq x \leqq \beta$ the use of

$$U(x) := \begin{cases} u(\alpha)\psi_\alpha(x) + u(\beta)\psi_\beta(x) & \text{regular case,} \\ u(\alpha)\phi_\alpha(x) + u(\beta)\phi_\beta(x) & \text{singular case} \end{cases}$$

where ψ_α and ψ_β are defined in § 2.1 and where

$$\phi_\beta(x) := (x^2-\alpha^2)/(\beta^2-\alpha^2), \qquad \phi_\alpha(x) := 1-\phi_\beta(x).$$

The regular case interpolant has been suggested by Russell and Shampine [14] for collocation. Second-order accuracy for the regular case interpolant is shown in Theorem 1 and for the singular case interpolant in Theorem 3.

The derivative $U_x(\xi)$ is also used. Theorem 2 states that for the ψ interpolant

$$|u_x(\xi) - U_x(\xi)| \leqq \left(\frac{\zeta}{\xi}\right)^m \left(1+\frac{m}{\gamma}\right) h^2 C(P)$$

where

$$\zeta^{m+1} := \int_\alpha^\beta x\,dx \Big/ \int_\alpha^\beta x^{-m}\,dx.$$

The point ζ is between γ_{-m} and γ, and so we have second-order accuracy for the regular case. Linear interpolation would cause this error to be increased by a factor of $1/\xi$ if $m \geqq 1$ and $a > 0$. Theorem 4 states that for the singular case with the ϕ interpolant

$$|u_x(\xi) - U_x(\xi)| \leqq \gamma h^2 C(\bar{P})/(m+1).$$

Again linear interpolation increases this error by a factor $1/\xi$. As before this bound for the singular case holds only on that interval $[0, c]$ on which Q is smooth. On the remainder of the interval Theorem 4 gives the bound

$$|u_x(\xi) - U_x(\xi)| \leqq \frac{m+1}{\gamma} h^2 C(P).$$

Having constructed an accurate interpolant, we replace $1/H_0$, D_0, and f_0 by $\tilde{G}_0 := G(\xi, U(\xi))$, $\tilde{D}_0 := D(\xi, U(\xi), U_x(\xi))$, and $\tilde{f}_0 := f(\xi, U(\xi), U_x(\xi))$ in (6), (7), and (12) and leave the precise form of the truncation errors for the Appendix. Equations (6) and (7) can be put into the simpler form:

$$\zeta^m \left(\frac{\xi}{\zeta}\right)^\mu \tilde{G}_0 U_x(\xi) = \alpha^m v(\alpha) + \frac{\zeta^{m+1}-\alpha^{m+1}}{m+1}(\tilde{D}_0 u_t(\alpha) - \tilde{f}_0) + \text{error} \tag{13}$$

and

$$\zeta^m \left(\frac{\xi}{\zeta}\right)^\mu \tilde{G}_0 U_x(\xi) = \beta^m v(\beta) - \frac{\beta^{m+1}-\zeta^{m+1}}{m+1}(\tilde{D}_0 u_t(\beta) - \tilde{f}_0) + \text{error}. \tag{14}$$

The foregoing derivation does not quite work for the special case $\alpha = 0$ and $m \geqq 1$. Because $\zeta = 0$, we do not want to multiply (6) by ζ^{m+1}. Therefore in (13) we adopt the understanding that we first divide by ζ^{m+1} with $\alpha > 0$ and then take the limit $\alpha \to 0$. (Note that $\alpha^m/\zeta^{m+1} \to 0$.) Also, because (7) is undefined, we use (12) to derive (14).

2.3. Assembly of the equations. If the term $\tilde{G}_0 U_x(\xi)$ in (13) and (14) is generalized to $g(\xi, U(\xi), U_x(\xi))$ and the truncation errors are neglected, we get difference equations

for the semidiscrete solution $\{u_j\}$, $\{v_j\}$. If we identify $[\alpha, \beta]$ with $[x_{j-1}, x_j]$, then with an obvious change of notation we get

$$(15_j) \qquad \zeta_{j-1/2}^{m-\mu}\xi_{j-1/2}^{\mu}g_{j-1/2} = x_{j-1}^m v_{j-1} + \frac{\zeta_{j-1/2}^{m+1} - x_{j-1}^{m+1}}{m+1}(D_{j-1/2}\dot{u}_{j-1} - f_{j-1/2}),$$

and

$$(16_j) \qquad -\zeta_{j-1/2}^{m-\mu}\xi_{j-1/2}^{\mu}g_{j-1/2} = -x_j^m v_j + \frac{x_j^{m+1} - \zeta_{j-1/2}^{m+1}}{m+1}(D_{j-1/2}\dot{u}_j - f_{j-1/2}),$$

for $j = 1, 2, \cdots, J$. At interior meshpoints x_j we add (15_{j+1}) to (16_j) to obtain

$$\zeta_{j+1/2}^{m-\mu}\xi_{j+1/2}^{\mu}g_{j+1/2} - \zeta_{j-1/2}^{m-\mu}\xi_{j-1/2}^{\mu}g_{j-1/2}$$
$$= \frac{\zeta_{j+1/2}^{m+1} - x_j^{m+1}}{m+1}(D_{j+1/2}\dot{u}_j - f_{j+1/2}) + \frac{x_j^{m+1} - \zeta_{j-1/2}^{m+1}}{m+1}(D_{j-1/2}\dot{u}_j - f_{j-1/2}).$$

At the right boundary we can solve (16_J) for v_J and substitute this for $g(b, u(b), u_x(b))$ in the right boundary condition. The same approach is used at the left boundary unless $m \geqq 1$ and $a = 0$ in which case we divide (15_1) by $\zeta_{1/2}^{m+1}$ with $a > 0$ and then take the limit $a \to 0$.

3. Galerkin formulation. Integration of the PDE (1) on $[\alpha, \beta]$ with weight function x^m and test function $\psi(x)$ and then integration by parts yields

$$(17) \qquad \psi(\beta)\beta^m v(\beta) - \psi(\alpha)\alpha^m v(\alpha) - \int_\alpha^\beta \psi_x g x^m\, dx = \int_\alpha^\beta \psi Q x^m\, dx.$$

We use as our shape functions $\psi_\alpha(x)$ and $\psi_\beta(x)$ in the regular case and $\phi_\alpha(x)$ and $\phi_\beta(x)$ in the singular case, and so our interpolant $U(x)$ is as defined in § 2.2. With $\psi = \psi_\alpha$, (17) becomes

$$-\int_\alpha^\beta x^m g \psi_{\alpha x}\, dx = \alpha^m v(\alpha) + \int_\alpha^\beta Q x^m \psi_\alpha\, dx,$$

and after numerical quadrature and lumping we get

$$-\xi^\mu g(\xi, U(\xi), U_x(\xi)) \int_\alpha^\beta x^{m-\mu}\psi_{\alpha x}\, dx$$
$$= \alpha^m v(\alpha) + Q(\xi, U(\xi), U_x(\xi), U_t(\alpha)) \int_\alpha^\beta x^m \psi_\alpha\, dx,$$

which is the same as (15_j). In a similar way as with $\psi = \psi_\beta$ we get (16_j).

In the *regular* case the test and trial (shape) functions are the same and thus the method is of *Galerkin* type. In the *singular* case the test and trial functions are different and thus the method is of *Petrov–Galerkin* type.

4. Other methods. In this section we discuss other low-order methods for solving the problem under consideration. The method described in § 4.1 is nearly the same as the method proposed in §§ 2 and 3, and it has been implemented in the SPRINT package. The remaining sections discuss linear Galerkin methods of Bakker [1], Eriksson and Thomée [9], and Berzins and Dew [3] and finite difference methods used in the algorithm PDEONE [15] and in the D03P** family [8] of NAG library routines.

All of these other methods, including the one implemented in SPRINT, use linear interpolation for $U(x)$. As explained in the Appendix this makes no difference to the order of the truncation error and matters only if a is small but positive.

4.1. The method used in SPRINT. The method implemented in SPRINT software of Berzins, Dew, and Furzeland [4] was developed as a first stage in the eventual development of the method described in § 2. It is a Petrov-Galerkin method that includes only some of the features of the proposed Galerkin method that improve the order of the local truncation error. In the nonpolar regular case the method is identical to that of § 2.

The test functions for an element $[\alpha, \beta]$ are chosen to be

$$\psi_\alpha(x) := \int_x^\beta y^{-m}\,dy \Big/ \int_\alpha^\beta y^{-m}\,dy$$

except if $m > 0$ and $\alpha = 0$, in which case

$$\psi_\alpha(x) = (\beta - x)/(\beta - \alpha)$$

and in both cases

$$\psi_\beta(x) := 1 - \psi_\alpha(x).$$

Instead of evaluating $x^m g$ at $x = \xi$ we evaluate g at $x = \gamma$, and the evaluation of D and f is at γ rather than ξ except for the case $m = 1$ where it is necessary to use ξ to maintain a propagated local error of $O(h^3)$.

4.2. The method used in PDEF1. A piecewise linear Galerkin method has been implemented by Bakker [1]. The crucial difference from the method described in § 2 is this: instead of (8) we get

$$\int_\alpha^\beta H(\cdots)v(x)\,dx = \frac{hH(\gamma, U(\gamma))}{\int_\alpha^\beta x^m\,dx} \int_\alpha^\beta x^m v(x)\,dx + \bar{\sigma}.$$

If we consider the example $H \equiv 1$ and $v(x) = x$, we have

$$\bar{\sigma} = -h(\gamma_m - \gamma) = -\frac{m}{12}\frac{h^3}{\gamma} + O\left(\frac{h^5}{\gamma^3}\right)$$

so that one order of accuracy is lost near the origin. This is serious because the accumulation of such local errors leads to a global error of $O(h^2 \log(1/h))$, an estimate derived by Jespersen [10].

4.3. The method of Eriksson and Thomée. To obtain better accuracy when $m > 1$, Eriksson and Thomée [9] consider the more specialized PDE

$$x^{-1}(xu_x + (m-1)u)_x = Q(x, u, u_x, u_t).$$

A variational equation is obtained by replacing u with its piecewise linear interpolant U, multiplying by a "hat" function $\phi_j(x)$, and integrating with weight function x. The stiffness matrix, in the case $Q(\cdots) = q(x)u - f(x)$, is not symmetric for this method, although it is for the method of § 4.1. However, it is proved [9] that the global error is $O(h^2)$.

Before applying numerical quadrature we obtain the equations

$$\left(\alpha + \frac{mh}{2}\right)\frac{u(\beta) - u(\alpha)}{h} = \alpha v(\alpha) + \int_\alpha^\beta x\phi_\alpha(x)Q(x, U(x), U_x(x), U_t(x))\,dx + \text{error}$$

and

$$-\left(\beta-\frac{mh}{2}\right)\frac{u(\beta)-u(\alpha)}{h}=-\beta v(\beta)+\int_{\alpha}^{\beta} x\phi_{\beta}(x)Q(x,\,U(x),\,U_x(x),\,U_t(x))\,dx+\text{error}.$$

In the case $m=2$ this is identical to the method of § 4.1 if the Q's are replaced by their lumped midpoint values and U by a linear interpolant (and it is identical to the finite difference method of Chawla and Katti [7] if the trapezoidal rule is used for quadrature). If $m\neq 2$, the methods are quite different.

4.4. The method used in SGENCO. If the Berzins and Dew [3] C^0 collocation code SGENCO is used with linear basis functions for the equation

$$(g(x,\,u,\,u_x))_x=Q(x,\,u,\,u_x,\,u_t)-\frac{m}{x}\,g(x,\,u,\,u_x),$$

then we get the equations

$$\frac{\alpha+mh}{2}\,g_{\alpha+}+\frac{\alpha}{2}\,g_{\beta-}=\alpha v(\alpha)+\frac{h}{2}\,\alpha Q_{\alpha+}+\text{error}$$

and

$$-\frac{\beta}{2}\,g_{\alpha+}-\frac{\beta-mh}{2}\,g_{\beta-}=-\beta v(\beta)+\frac{h}{2}\,\beta Q_{\beta-}+\text{error}$$

where

$$Q_{\alpha+}=Q(\alpha+,\,u(\alpha),\,U_x(\alpha+),\,u_t(\alpha)),\quad\text{etc.}$$

If $g(x,\,u,\,u_x)\equiv u_x$, then this is identical to the method of Eriksson and Thomée with trapezoidal quadrature. For $m=1$ it is thus a Galerkin method with weight function x, and the global error is $O(h^2\log(1/h))$.

4.5. The method used in PDEONE. We derive here a scheme like that of Sincovec and Madsen [16] and Varga [18, p. 175]. In (4) and (5) instead of evaluating H and Q at $x=\xi$, we evaluate the entire integrand at $x=\gamma$ giving

$$u(\beta)=u(\alpha)+h\frac{H_0}{\gamma^m}\left\{\alpha^m v(\alpha)+\int_{\alpha}^{\gamma} x^m Q(\cdots)\,dx\right\}+\bar{\sigma}$$

and

$$u(\alpha)=u(\beta)-h\frac{H_0}{\gamma^m}\left\{\beta^m v(\beta)-\int_{\gamma}^{\beta} x^m Q(\cdots)\,dx\right\}-\bar{\sigma}$$

where $H_0=H(\gamma,\,U(\gamma))$. For the integrals of $Q(\cdots)$ we use one-point quadrature rules yielding

$$u(\beta)=u(\alpha)+h\frac{H_0}{\gamma^m}\left\{\alpha^m v(\alpha)+\frac{\gamma^{m+1}-\alpha^{m+1}}{m+1}\,Q_{\alpha+}\right\}+H_0\bar{\tau}_{\alpha}+\bar{\sigma}\tag{18}$$

and

$$u(\alpha)=u(\beta)-h\frac{H_0}{\gamma^m}\left\{\beta^m v(\beta)-\frac{\beta^{m+1}-\gamma^{m+1}}{m+1}\,Q_{\beta-}\right\}+H_0\bar{\tau}_{\beta}-\bar{\sigma}.\tag{19}$$

If we proceed as in § 2 we obtain the difference equation

$$\frac{m+1}{x_{j+1/2}^{m+1}-x_{j-1/2}^{m+1}}\left(x_{j+1/2}^{m}g_{j+1/2}-x_{j-1/2}^{m}g_{j-1/2}\right) \tag{20}$$
$$=\frac{x_{j+1/2}^{m+1}-x_{j}^{m+1}}{x_{j+1/2}^{m+1}-x_{j-1/2}^{m+1}}Q_{j+}+\frac{x_{j}^{m+1}-x_{j-1/2}^{m+1}}{x_{j+1/2}^{m+1}-x_{j-1/2}^{m+1}}Q_{j-}.$$

Remark. The averaging of Q_{j+} and Q_{j-} actually recommended by Sincovec and Madsen [16, p. 242] is different; it is obtained by setting $m=0$ in the right-hand side of (20). Also the spatial derivative term in $Q(x, u, u_x, u_t)$ is approximated by $(u_{j+1}-u_{j-1})/(h_{j+1}+h_j)$, which is not so accurate if u_x has a discontinuity at x_j.

Combining (18) and (19), we get

$$\beta^m v(\beta)=\alpha^m v(\alpha)+\frac{\beta^{m+1}-\gamma^{m+1}}{m+1}Q_{\beta-}+\frac{\gamma^{m+1}-\alpha^{m+1}}{m+1}Q_{\alpha+}+\bar{\tau}$$

where

$$\bar{\tau}:=\int_{\alpha}^{\beta}x^m Q(\cdots)\,dx-\frac{\beta^{m+1}-\gamma^{m+1}}{m+1}Q_{\beta-}-\frac{\gamma^{m+1}-\alpha^{m+1}}{m+1}Q_{\alpha+}.$$

If we consider $H\equiv 1$, $Q(x)=x$, and $\alpha=0$, then

$$\bar{\tau}_S=\int_h^1\frac{dy}{y^m}\cdot\bar{\tau}+\bar{\tau}_\alpha$$
$$=\int_h^1\frac{dy}{y^m}\cdot\frac{m+2-2^{m+1}}{(m+1)(m+2)2^{m+1}}h^{m+2}+\frac{h^3}{4(m+2)},$$

and if $m=1$ the resulting contribution to the global error is $O(h^3\log(1/h))$. Nonetheless the accumulation of such errors is only $O(h^2)$.

The error $\bar{\sigma}$ is simply

$$\bar{\sigma}=\int_{\alpha}^{\beta}u_x(x)-u_x(\gamma)\,dx.$$

If $a>0$, $H\equiv 1$, $Q\equiv 0$, $v(a)=1$, and $\alpha=a$, then

$$\bar{\sigma}=\frac{m(m+1)}{24}\frac{h^3}{a^2}+O(h^4),$$

which is worse than the method of § 4.1 by a factor of $1/a$ and worse than the method of § 2 by a factor of $1/a^2$. If we consider $a=0$, $H\equiv 1$, $Q=0$ for $x<c$, $Q=1$ for $x>0$, and $\alpha=c$, then

$$\bar{\sigma}=-\frac{m}{24}\frac{h^3}{c}+O(h^4),$$

which is worse than the methods of §§ 2 and 4.1 by a factor of $1/c$.

In conclusion we see that this method does achieve global second-order accuracy for general n but is less accurate than the method proposed in § 2.

4.6. The method used in D03P.** The difference scheme of this collection of NAG routines by Dew and Walsh [8] is modeled after that of PDEONE. To simplify somewhat the usage of the routines, the coefficient $G(\gamma, U(\gamma))$ is replaced by $\frac{1}{2}(G(\alpha, U(\alpha))+G(\beta, U(\beta)))$. This is all right except at a discontinuity x_j where there

seems to be no way to define an averaging for $G(x_j, u_j)$ that does not degrade the accuracy in either the left subinterval or right subinterval. The codes D03P** and PDEONE are compared by Berzins and Dew [2].

5. Integration in time. The issue of integration in time is not considered in any detail here. Instead it is noted that the spatial discretization of elliptic-parabolic PDEs using the method of § 2 results in large systems of differential-algebraic equations that are integrated using standard software (see Berzins, Dew, and Furzeland [4] and Petzold [12] for further details).

5.1. Explicit or implicit ODEs. Although it is possible to integrate stiff implicit ODEs with almost the same overhead as explicit ODE systems written in normal form, there are still a number of reasons why it is preferable to reduce systems to normal form whenever possible, such as by the lumping applied in § 2.1.

A substantial difficulty with implicit differential or differential-algebraic equations lies in the calculation of the initial solution values and their time derivatives (see Petzold [12]). This is not a problem with systems written in normal form. A further difficulty is that with implicit equations it is sometimes possible to calculate physically misleading values for the initial time derivatives. This point is easily illustrated by the simple example below and leads to a noticeable deterioration of the performance of the ODE integrator.

5.1.1. Example. Consider the heat equation $u_t = u_{xx}$ with boundary and initial conditions given by

$$u(0, t) = 0,$$

$$u(1, t) = \begin{cases} 0, & t < 1, \\ 1, & t \geqq 1, \end{cases}$$

$$u(x, 0) = 0,$$

and suppose that we semidiscretize in space without lumping using a uniform mesh. This yields the system of ODEs

$$\frac{1}{6}\dot{u}_{j+1} + \frac{2}{3}\dot{u}_j + \frac{1}{6}\dot{u}_{j-1} = \frac{1}{h^2}(u_{j+1} - 2u_j + u_{j-1}).$$

The analytical solution (the limiting case of the backward Euler solution as the time stepsize goes to zero) at time $t = 1$ is

$$u_j(1) = \left(\frac{-1}{2+\sqrt{3}}\right)^{J-j} \frac{1-(2+\sqrt{3})^{-2j}}{1-(2+\sqrt{3})^{-2J}},$$

and in particular

$$u_{J-1}(1) \cong -.268, \quad u_{J-2}(1) \cong .072, \quad u_{J-3}(1) \cong -.019$$

for large J. Thus, the semidiscrete solution at interior meshpoints is discontinuous in time and oscillates in space with every other value having the wrong sign.

More generally, a discontinuity or a rapid variation of a boundary value produces corresponding behavior with alternating signs at nearby meshpoints. A discontinuity is often present at $t = 0$, when the boundary condition is inconsistent with the initial condition and the PDE. For example, if $u(1, t) = \sin t$ in the above example, then the derivatives $\dot{u}_J(0+) = 1$, $\dot{u}_{J-1}(0+) \cong -.268$, etc. We have observed that the effect of all this is to degrade the efficiency of the integration.

A further incentive to derive ODE systems in normal form is provided by a new generation of ODE initial value problem codes for both stiff and nonstiff equations (for example, see Petzold [13]). Such integrators attempt to use functional iteration whenever possible to increase the efficiency of integration. The unsuitability of functional iteration for the systems of equations that arise from implicit ODEs means that these codes cannot be applied to such equations at the moment.

6. Numerical testing on parabolic equations. In the numerical testing that was conducted the following measures were taken to ensure that a fair comparison was made. First, all integrations were performed using the same ODE integrator and the same linear algebra routines. The integrator used was the BDF/Adams code with the LINPACK banded matrix routines as implemented in SPRINT (Berzins, Dew, and Furzeland [4]). All discretization methods compared here were compared in a common framework. Only minor changes had to be made to the PDEONE code of Sincovec and Madsen [16] and to the PDEF1 code of Bakker [1] to fit them into this framework. These codes are abbreviated in this report as follows:

PDEONE: the Sincovec and Madsen [16] code,

SGENCO: the Berzins and Dew [3] C^0-collocation code used with linear basis functions,

SPRINT: the discretization method of § 4.1,

PDEF1: the lumped finite element method of Bakker [1] that uses linear basis functions,

NEW: the discretization method of § 2.

6.1. Nonpolar parabolic equations. Testing over a range of simple nonpolar parabolic equations has shown that the formula used by Bakker [1] and by Skeel [17] gives consistently better results than the finite difference methods of Dew and Walsh [8] and Sincovec and Madsen [16], although only to the same order of accuracy. The following problem gives results typical of those obtained on the nonpolar test problems of Berzins and Dew [3]. The methods of Skeel [17], §§ 4.1 and 2 are identical for nonpolar test problems.

6.1.1. Problem 1.1. This problem has an analytic solution and a material interface at $x=0$. The problem is defined by

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(\frac{1}{C_1}\frac{\partial u}{\partial x}\right)+C_1e^{-2u}+e^{-u}, \qquad x\in[-1,0),$$

$$\frac{\partial u}{\partial t}=\frac{\partial}{\partial x}\left(\frac{1}{C_2}\frac{\partial u}{\partial x}\right)+C_2e^{-2u}+e^{-u}, \qquad x\in(0,1],$$

$$(x,t)\in[-1,1]\times(0,1]$$

subject to the boundary conditions

$$u(-1,t)=\log(-C_1+t+P),$$

$$u(1,t)+(C_2+t+P)\frac{\partial u}{\partial x}(1,t)=\log(C_2+t+P)+1.0.$$

The initial condition is consistent with the analytic solution of

$$u(x,t)=\log(C_ix+t+P)$$

where $i=1$ if $x<0$ and $i=2$ if $x>0$. In this case $P=1.1$, $C_1=0.1$, and $C_2=1.0$. This problem illustrates well the performance of the three codes on nonpolar parabolic

TABLE 1
Error norms for Problem 1.1.

Time	0.01	0.22	0.55	0.77	1.0
$N=11$					
PDEONE	1.9-2	9.0-3	3.9-3	2.7-3	1.9-3
SPRINT	1.9-2	8.3-3	3.7-3	2.3-3	1.5-3
SGENCO	1.4-2	6.4-3	2.8-3	1.9-3	1.3-3
$N=41$					
PDEONE	1.5-3	6.0-4	2.5-4	1.7-4	1.2-4
SPRINT	1.3-3	5.4-4	2.4-4	1.5-4	9.9-5
SGENCO	1.5-3	6.0-4	2.5-4	1.8-4	1.2-4
$N=161$					
PDEONE	9.4-5	3.8-5	1.6-5	1.1-5	7.6-6
SPRINT	7.8-5	3.4-5	1.5-5	9.3-6	6.2-6
SGENCO	9.4-5	3.8-5	1.6-5	1.1-5	7.6-6

TABLE 2
Maximum grid errors.

$N=$	11	21	41	81	161
PDEONE	1.9-2	6.2-3	1.7-3	4.3-3	1.1-4
SPRINT	1.3-2	3.3-3	8.3-4	2.1-4	5.2-5
SGENCO	1.9-2	6.1-3	1.7-3	4.3-3	1.1-4

equations. The Dew and Walsh [8] code cannot be compared as it does not correctly treat the material interface at $x=0$. The Bakker [1] code PDEF1 gives identical results to SPRINT for nonpolar problems. Table 1 shows the L^2 error norms at different time levels as the number of equally-spaced meshpoints is increased. N is the number of evenly-spaced meshpoints used in spatial discretization.

The error norms were formed by using a 201-point trapezoidal rule with evenly-spaced meshpoints and with solution values in between the PDE meshpoints being estimated by linear interpolation. Table 2 shows the maximum grid error for each of the methods sampled over the time values used in Table 1 with the additional points 0.11, 0.33, 0.44, 0.66, 0.88. The factor of two difference between the codes SPRINT and SGENCO can be directly attributed to the fact that both are essentially lumped Galerkin methods but that SPRINT uses a midpoint quadrature rule and SGENCO uses a trapezoidal rule.

6.2. Numerical testing on polar parabolic equations. The following test problems were used to compare the different codes on polar parabolic test problems.

6.2.1. Problem 2.1.

$$u\frac{\partial u}{\partial t}=\frac{1}{x^2}\frac{\partial}{\partial x}\left(x^2u\frac{\partial u}{\partial x}\right)+5u^2+4xu\frac{\partial u}{\partial x}, \qquad (x,t)\in[0,1]\times(0,1].$$

The left-hand boundary condition is the symmetry condition and the right-hand Dirichlet condition and the initial condition are consistent with the analytic solution of

$$u(x,t)=e^{1-x^2-t}.$$

6.2.2. Problem 2.2.

$$\frac{\partial u}{\partial t}=\frac{1}{x^2}\frac{\partial}{\partial x}\left(x^2\frac{1}{6}\frac{\partial u}{\partial x}\right)+\frac{2}{3}x^2e^{-2u}, \qquad (x,t)\in[0,1]\times(0,1].$$

The boundary conditions are the symmetry condition at $x=0$ and the boundary condition at $x=1$ given by

$$u(1,t)+(1.0+p+t)\frac{\partial u}{\partial x}=2.0+\log(1+p+t)$$

where $p=1.0$ and the analytic solution is given by

$$u(x,t)=\log(x^2+p+t).$$

6.2.3. Problem 2.3.

$$\frac{\partial u}{\partial t}=\frac{1}{x^2}\frac{\partial}{\partial x}\left(x^2\frac{\partial u}{\partial x}\right)+F(x,t), \qquad (x,t)\in[0,1]\times(0,1]$$

where

$$F(x,t)=e^{-t}[(6+(1-x^2)(\pi^2t^2-1))\cos(\pi xt) \\ -[(1-x^2)x+4xt-2t/x]\pi\sin(\pi xt)].$$

The boundary conditions are the symmetry condition at $x=0$ and $u=0$ at $x=1$. The exact solution is given by

$$u(x,t)=(1-x^2)e^{-t}\cos(\pi xt).$$

The limiting value of the function F at $x=0$ is obtained by using

$$\lim_{x\to 0}\frac{\sin(\pi xt)}{x}=\pi t.$$

6.2.4. Problem 2.4.

$$\frac{\partial u}{\partial t}=\frac{1}{x^2}\frac{\partial}{\partial x}\left(x^2\frac{\partial u}{\partial x}\right), \qquad (x,t)\in[0,1]\times(0,0.8]$$

where the boundary conditions are the symmetry condition at $x=0$ and

$$u(1,t)=1+6t.$$

The exact solution is given by

$$u(x,t)=x^2+6t.$$

6.2.5. Problem 2.5. This problem consists of the elliptic parabolic system defined by

$$\frac{\partial u}{\partial t}=\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial u}{\partial x}\right)+F(x),$$

$$0=\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial v}{\partial x}\right)+F(x),$$

$$(x,t)\in[0,1]\times(0,1]$$

where $F(x)=x$ for $x<p$ and equal to zero otherwise and $p=0.1$. The boundary conditions are the Neumann condition

$$\frac{\partial u}{\partial x}=\frac{\partial v}{\partial x}=0 \quad \text{at } x=0$$

and the Dirichlet condition

$$u=v=u_{\text{exact}} \quad \text{at } x=1$$

where the exact solution is given by

$$u=v=-\log(x)\frac{p^3}{3}, \qquad x>p$$

$$=-\log(p)\frac{p^3}{3}+\frac{p^3-x^3}{9}, \qquad x\leqq p.$$

6.2.6. Problem 2.6. The heat equation in cylindrical polar coordinates is

$$\frac{\partial u}{\partial t}=\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial x}{\partial x}\right), \qquad (x,t)\in[0,1]\times(0,1]$$

where the exact solution is given by

$$u(x,t)=J_0(px)\,e^{-p^2t}$$

and $p\approx 2.40482557$ is the first zero of the Bessel function $J_0(x)$. The boundary conditions are a Dirichlet condition at $x=1$ and the usual symmetry condition at $x=0$.

6.2.7. Problem 2.7. The following problem is taken from Eriksson and Thomée [9]:

$$\frac{\partial u}{\partial t}=\frac{1}{x^2}\frac{\partial}{\partial x}(x^2u)-3u+\frac{\sinh(2x)}{x\sinh 2}-4e^t+3,$$

$$(x,t)\in[0,1]\times(0,1].$$

The boundary and initial conditions are given by

$$\frac{\partial u}{\partial x}(0,t)=u(1,t)=u(x,0)=0, \qquad t>0$$

and the exact solution is given by

$$u(x,t)=(e^t-1)\frac{\sinh(2x)}{x\sinh 2}-e^t+1.$$

6.2.8. Problem 2.8. The following is a slightly more complicated problem in cylindrical polar coordinates:

$$\frac{\partial u}{\partial t}=\frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial u}{\partial x}\right)+3u+2x\frac{\partial u}{\partial x},$$

$$(x,t)=[0,1]\times(0,1].$$

The boundary conditions are given by

$$\frac{\partial u}{\partial x}(0,t)=0, \quad u(1,t)=e^{-t}, \quad t>0$$

and the exact solution and the initial condition are given by

$$u(x, t) = e^{1-t-x^2}.$$

6.2.9. Problem 2.9. This problem consists of the parabolic equation defined by

$$\frac{\partial u}{\partial t} = \frac{1}{x}\frac{\partial}{\partial x}\left(x\frac{\partial u}{\partial x}\right) + F(x), \qquad (x, t) \in [0, 1] \times (0, 1]$$

where $F(x) = 100$ for $x < 0.1$ and equal to zero otherwise. The boundary conditions are the symmetry condition at $x = 0$

$$\frac{\partial u}{\partial x} = 0 \quad \text{at } x = 0$$

and the Dirichlet condition

$$u = 0 \quad \text{at } x = 1$$

where the exact solution is given by

$$u = \begin{cases} 0.5 \log (0.1) + 25((0.1)^2 - x^2) + J_0(px)\, e^{-p^2 t}, & x \leqq 0.1, \\ 0.5 \log x + J_0(px)\, e^{-p^2 t}, & x \geqq 0.1 \end{cases}$$

where $p \approx 2.40482557$ is the first zero of the Bessel function $J_0(x)$. This problem has a severe discontinuity in the PDE defining function close to the polar origin.

6.3. Summary of numerical testing results. The numerical testing results are summarized by the eight graphs of Fig. 1. The results for Problem 2.4 are not presented graphically and this problem is covered separately below. The procedure employed for each of the test problems was to use five evenly-spaced meshes of 11, 21, 41, 81, and 161 meshpoints. Each of the integrations in time was performed to a local ODE error tolerance that was sufficiently small for the PDE spatial discretization error to dominate the global error in the solution. All the graphs below are of the $\log_{10}$ of the maximum grid error against the $\log_{10}$ of the number of spatial meshpoints. The maximum error at the spatial meshpoints was found by stopping the integration at times $t = 0.01$ and then $t = k/9$ for $k = 1, 2, \cdots, 9$.

For the sake of clarity certain codes were not included on certain graphs. The following points should be noted:

The Bakker code PDEF1 is not applicable to Problem 2.1.

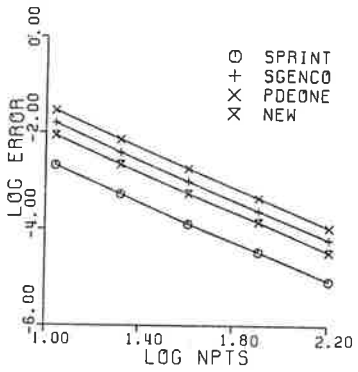
The results produced by PDEONE and SGENCO are indistinguishable on Problems 2.5, 2.6, 2.7, and 2.9.

The graph for Problem 2.6 compares the published results at $t = 1.0$ of Thomée and Eriksson [9] with the results at $t = 1.0$ for the other codes.

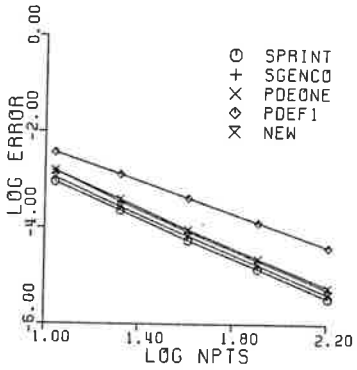
The codes SGENCO and PDEONE have a great deal of difficulty with the discontinuity in Problem 2.9.

The SPRINT code and the method of § 2 produce identical results for Problems 2.8, 2.6, and 2.9.

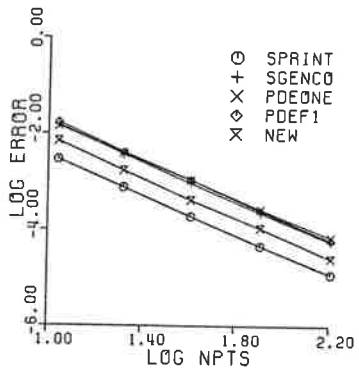
The results for Problem 2.4 are not presented graphically because all the codes apart from PDEF1 produce solutions that are exact at the meshpoints to computer roundoff error. The results for the maximum grid error at the meshpoints (EMAX) for code PDEF1 are as follows.



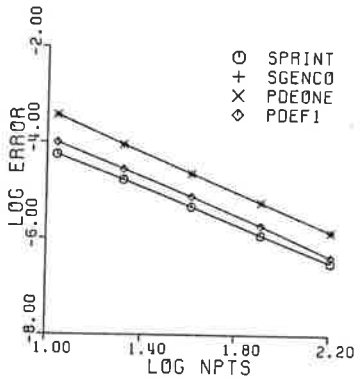
Problem 2.1



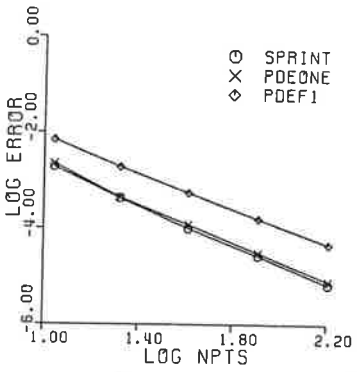
Problem 2.2



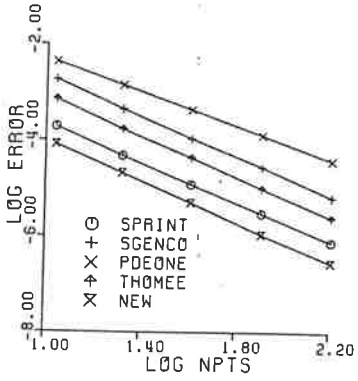
Problem 2.3



Problem 2.5



Problem 2.6



Problem 2.7

FIG. 1. *Summary of numerical testing results.*

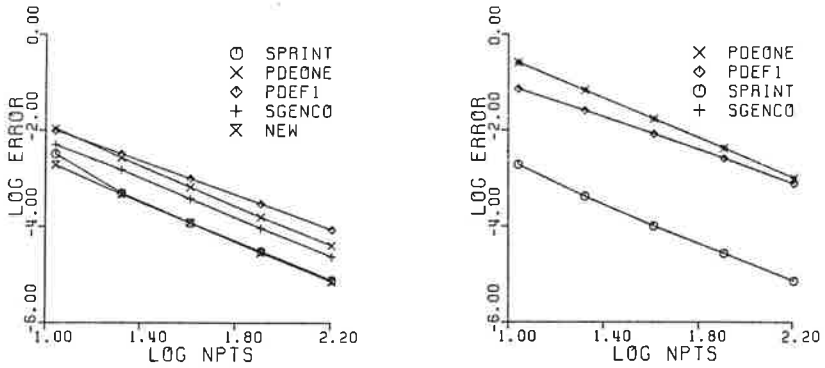


FIG. 1—continued.

Max. Grid errors for PDEF1 on Prob. 2.4.

NPTS	11	21	41	81	161
EMAX	6.1E−1	2.1E−1	6.6E−2	2.0E−3	5.9E−3

The explanation for the poor showing of this code on this problem is that it semidiscretizes the PDE of Problem 2.4 to give the following equation at the internal meshpoints:

$$u_t = \frac{6+h^2/2x^2}{1+h^2/4x^2} \quad \text{at } x = x_j$$

where h is the even mesh spacing; whereas the other codes semidiscretize the PDE to the exact equation given by

$$u_t = 6 \quad \text{at } x = x_j.$$

The results for Problem 2.9 in particular show the advantage of the discretization method that we have developed over PDEONE and SGENCO.

The results for the Dew and Walsh [8] code D03PGF are not presented in the tables and graphs. On all the problems except Problems 2.2 and 2.5 the performance of the code is slightly worse than but very similar to SGENCO. On Problem 2.5 the code does not perform well due to the way it treats the discontinuities in the first derivative of the solution.

The graphical results clearly illustrate the superior performance of the SPRINT discretization and the discretization of § 2. The failure of the latter to perform the best on every problem has been analyzed, and it appears to be due to the fact that the new method has been designed to minimize a bound on the individual contribution of each local error rather than to achieve some cancellation of errors from different elements. Such cancellation can be fully realized, however, only for smooth problems on uniform meshes. Thus, it could be suggested that the method of § 2 sacrifices small possible gains in accuracy in order to achieve greater robustness.

7. Conclusions. Theoretical and experimental evidence indicates that the Galerkin method derived in this paper for the regular case and the Petrov-Galerkin method derived for the singular case produce more accurate results than existing methods and

in particular they seem to be the best methods to use in a general-purpose library subroutine.

Appendix A. Error analysis. This Appendix gives bounds on the local truncation error of the method derived in § 2. Some of the proofs are longer than necessary in order to obtain tighter bounds, although this does not show up in the statements of the theorems. We use P to denote the 17-tuple consisting of the supremum norms on the open subintervals of $[a, b]$ of u_t, u_{tt}, v, v_t, H, H_x, H_t, H_{xx}, H_{xt}, D, D_x, D_t, D_{xx}, f, f_x, f_t, f_{xx}. Here H_x denotes $(H(x, u(x)))_x$ rather than $H_x(x, u)$ evaluated at $u = u(x)$. This is similar for the other derivatives of H, D, and f with respect to x and t. For the singular case we use $\bar{P}$ to denote a similar 17-tuple except that we use the supremum norm on all of $[0, c]$ for all 17 quantities (although for H_{xx} open subintervals would suffice). The error analysis assumes $b = a+1$.

We shall examine the various errors introduced in a typical subinterval $[\alpha, \beta]$. The substitution of $\tilde{H}_0$, $\tilde{D}_0$, $\tilde{f}_0$ for H_0, D_0, f_0 in (8), (9), (10), (12) changes the error terms σ, τ_α, τ_β, τ to, say,

$$\tilde{\sigma} = \sigma + \sigma^I, \qquad \tilde{\tau}_\alpha = \tau_\alpha + \tau_\alpha^I,$$
$$\tilde{\tau}_\beta = \tau_\beta + \tau_\beta^I, \qquad \tilde{\tau} = \tau + \tau^I$$

where the terms σ^I, τ_α^I, τ_β^I, and τ^I are the interpolation errors resulting from replacing $u(x)$ and $u_x(x)$ in H, D, and f by $U(x)$ and $U_x(x)$. (See, for example, (23) and (24).) Quadrature/lumping errors are considered in § A.2 and interpolation errors in § A.1.

Of course, it is the effect of the local truncation errors on the global error that is important, especially since there is considerable latitude in how the local errors might be defined. Hence, we define the propagated local error to be the global error that would result from the use of a single finite element for $[\alpha, \beta]$ and the use of infinitesimal elements outside of this subinterval. This can be rather complicated, and so we model the problem outside of $[\alpha, \beta]$ by the differential equation $x^{-m}(x^m u_x / H(x))_x = Q(x)$ with linearized boundary conditions. Thus, the effect of the singularity is included. A lengthy argument indicates that the boundary conditions which give the largest errors are the following:

(i) *Singular case.* $v(a) =$ given, $u(b) =$ given, for which the propagated local error $\tilde{\tau}_s$ can be as large as

$$\tilde{\tau}_s = \int_\beta^b \frac{H(y)}{y^m}\, dy \cdot \tilde{\tau} + \tilde{H}_0 \tilde{\tau}_\alpha + \tilde{\sigma}; \tag{21}$$

and

(ii) *Regular case.* $u(a) =$ given, $v(b) =$ given (which supposes that $a > 0$ or $m = 0$) for which the propagated local error $\tilde{\tau}_r$ can be as large as

$$\tilde{\tau}_r = \int_a^\alpha \frac{H(y)}{y^m}\, dy \cdot \tilde{\tau} + \tilde{H}_0 \tilde{\tau}_\beta - \tilde{\sigma}. \tag{22}$$

Because the primary purpose of our error analysis is to support our choice of interpolant and quadrature point, we stop just short of giving an explicit bound for either (21) or (22) and instead produce just the necessary constituents for such a bound.

A.1. Interpolation error. The local truncation errors due to interpolation are given by

$$\sigma^I := (H_0 - \tilde{H}_0) \int_\alpha^\beta v(x)\, dx \tag{23}$$

and

$$\tau^I := (D_0 - \tilde{D}_0) \int_\alpha^\beta x^m (u_t(\alpha)\psi_\alpha(x) + u_t(\beta)\psi_\beta(x))\, dx - (f_0 - \tilde{f}_0) \int_\alpha^\beta x^m\, dx \tag{24}$$

together with suitable equations for τ_α^I and τ_β^I. All these errors involve partial derivative of H, D, and f times the differences $u(x) - U(x)$ and $u_x(x) - U_x(x)$.

In the proofs that follow let $H(x) := H(x, u(x))$, $D(x) := D(x, u(x), u_x(x))$, and $f(x) := f(x, u(x), u_x(x))$. Also define $Q(x) := D(x)u_t(x) - f(x)$.

THEOREM 1. *The ψ interpolant satisfies*

$$\|u - U\| \leqq h^2 C(P)$$

where $\|\cdot\|$ denotes the maximum norm on $[\alpha, \beta]$.

Proof. Consider first the case $\alpha > 0$ or $m = 0$. Integrating by parts, we obtain

$$u(x) = u(\alpha) + \int_\alpha^x \frac{dy}{y^m} H(x)x^m v(x) - \int_\alpha^x \int_\alpha^y \frac{dz}{z^m} (H(y)y^m v(y))_y\, dy.$$

A second equation is obtained by replacing α by β. Multiplying the first equation by $\psi_\alpha(x)$ and the second by $\psi_\beta(x)$, we obtain

$$u(x) = U(x) - \psi_\alpha(x) \int_\alpha^x \int_\alpha^y \frac{dz}{z^m} (H(y)y^m v(y))_y\, dy - \psi_\beta(x) \int_x^\beta \int_y^\beta \frac{dz}{z^m} (H(y)y^m v(y))_y\, dy.$$

Because

$$(H(x)x^m v(x))_x = x^m (H_x(x)v(x) + H(x)Q(x))$$

we have

$$|u(x) - U(x)| \leqq \omega(x)(\|H_x v + HQ\|)$$

where

$$\omega(x) := \psi_\alpha(x) \int_\alpha^x \int_\alpha^y \frac{dz}{z^m} y^m\, dy + \psi_\beta(x) \int_x^\beta \int_y^\beta \frac{dz}{z^m} y^m\, dy.$$

Integration once by parts gives

$$\omega(x) = \frac{1}{2(m+1)} \left\{ (\beta^2 - x^2) \int_\alpha^x \frac{dy}{y^m} - (x^2 - \alpha^2) \int_x^\beta \frac{dy}{y^m} \right\} \Big/ \int_\alpha^\beta \frac{dy}{y^m}$$

and a second time gives

$$\omega(x) = \frac{1}{4} \left\{ (\beta^2 - x^2) \int_\alpha^x \frac{y^2 - \alpha^2}{y^2} \frac{dy}{y^m} + (x^2 - \alpha^2) \int_x^\beta \frac{\beta^2 - y^2}{y^2} \frac{dy}{y^m} \right\} \Big/ \int_\alpha^\beta \frac{dy}{y^m}.$$

Clearly,

$$\omega(x) \leqq \frac{(\beta^2 - x^2)(x^2 - \alpha^2)}{4x^2},$$

which is maximized for $x^2 = \alpha\beta$ to yield the stated bound. For the case $\alpha = 0$ and $m \geqq 1$, we have

$$u(x) = u(\beta) - \int_x^\beta \int_0^y (H(z)z^m v(z))_z\, dz \frac{dy}{y^m}.$$

When we note that $U(x) = u(\beta)$ and

$$|(H(z)z^m v(z))_z| \leqq z^m C(P)$$

the theorem easily follows. □

In an identical fashion we can show that

$$\|u_t - U_t\| \leqq h^2 C(P). \tag{25}$$

Less accuracy can be expected for the approximation $U_x(\xi)$.

LEMMA. *If* $k \neq 0$, *then*

$$\frac{\gamma_k - \gamma}{k} \leqq \frac{h^2}{4\gamma}.$$

Furthermore,

$$\zeta \geqq \gamma_{-m}.$$

Proof of first inequality. We have $\gamma_k = \gamma I'/I$ where

$$I' := \gamma^{-k-2} \int_\alpha^\beta x^{k+1}\, dx$$

and

$$I := \gamma^{-k-1} \int_\alpha^\beta x^k\, dx.$$

With a change of variables $x =: \gamma(1+s)$ and $\theta := h/(2\gamma)$, we get

$$\begin{aligned} I &= \int_{-\theta}^{\theta} (1+s)^k\, ds \\ &= \int_0^\theta \{(1+s)^k + (1-s)^k\}\, ds. \end{aligned} \tag{26}$$

For I' we get (26) with m reduced by 1. Subtraction gives

$$\begin{aligned} I' - I &= \int_0^\theta s\{(1+s)^k - (1-s)^k\}\, ds \\ &= k \int_0^\theta \frac{\theta^2 - s^2}{2} \{(1+s)^{k-1} + (1-s)^{k-1}\}\, ds. \end{aligned} \tag{27}$$

Therefore, using (26) and (27), we have

$$\begin{aligned} \frac{k}{4}\frac{h^2}{\gamma} + \gamma - \gamma_k &= k\theta^2\gamma + \gamma - \gamma I'/I \\ &= \frac{\gamma}{I}\{k\theta^2 I + I - I'\} \\ &= \frac{k\gamma}{2I}\left\{2\theta^2 \int_0^\theta \{(1+s)^k + (1-s)^k\}\, ds \right. \\ &\quad \left. - \int_0^\theta (\theta^2 - s^2)\{(1+s)^{k-1} + (1-s)^{k-1}\}\, ds \right\} \\ &= \frac{k\gamma}{2I} \int_0^\theta \{(\theta^2 + 2s\theta^2 + s^2)(1+s)^{k-1} + (\theta^2 - 2s\theta^2 + s^2)(1-s)^{k-1}\}\, ds. \end{aligned}$$

The integral is nonnegative because $-2s\theta^2 \geqq -2s\theta$. □

Proof of second inequality. The inequality $\gamma_{-m} \leqq \zeta$ is a consequence of

$$\int_\alpha^\beta f(x)g(x)\,dx \leqq \left(\int_\alpha^\beta |f(x)|^{m+1}\,dx\right)^{1/(m+1)} \left(\int_\alpha^\beta |g(x)|^{(m+1)/m}\,dx\right)^{m/(m+1)}$$

with $f(x) = x^{1/(m+1)}$ and $g(x) = x^{-m^2/(m+1)}$. □

THEOREM 2. *The ψ interpolant satisfies*

$$|u_x(\xi) - U_x(\xi)| \leqq \left(\frac{\zeta}{\xi}\right)^m \left(1 + \frac{m}{\gamma}\right) h^2 C(P).$$

Proof. Defining $w(x) = x^m u_x(x)$, we note that

$$w_x = x^m (H_x v + HQ),$$

$$w_{xx} = x^m (H_{xx} v + 2H_x Q + HQ_x) + m x^{m-1} HQ,$$

and

$$Q_x = D_x U_t + D(Hv)_t - f_x. \tag{28}$$

If $m \geqq 1$ and $\alpha \to 0$, the bound goes to $+\infty$, and therefore assume either $m = 0$ or $\alpha > 0$. Hence we consider

$$U_x(\xi) - u_x(\xi) = \int_\alpha^\beta (w(x) - w(\xi)) \frac{dx}{x^m} \Big/ \left(\xi^m \int_\alpha^\beta \frac{dx}{x^m}\right).$$

The numerator

$$\int_\alpha^\beta (w(x) - w(\xi)) \frac{dx}{x^m} = w_x(\xi) \int_\alpha^\beta (x - \xi) \frac{dx}{x^m} + \int_\alpha^\beta \int_\xi^x \int_\xi^y w_{zz}(z)\,dz\,dy \frac{dx}{x^m}.$$

The first term vanishes because $\xi = \gamma_{-m}$. Replacing w_{zz} by bounds and using the definition of ζ, we obtain

$$|U_x(\xi) - u_x(\xi)| \leqq \frac{\zeta^{m+1}}{h\gamma\xi^m} \int_\alpha^\beta \int_\xi^x \int_\xi^y (z^m + m z^{m-1})\,dz\,dy \frac{dx}{x^m} C(P)$$

$$= \left(\frac{\gamma_1 \zeta^{m+1} - \xi^{m+2}}{(m+2)\xi^m} + \frac{\zeta^{m+1} - \xi^{m+1}}{\xi^m}\right) \frac{C(P)}{m+1}.$$

We have

$$\gamma_1 \zeta^{m+1} - \xi^{m+2} = (\gamma_1 - \zeta)\zeta^{m+1} + \zeta^{m+2} - \xi^{m+2}$$

$$\leqq (\gamma_1 - \zeta)\zeta^{m+1} + (\zeta - \xi)(m+2)\zeta^{m+1}$$

$$\leqq (\gamma_1 - \gamma)\zeta^{m+1} + (\gamma - \xi)(m+2)\zeta^{m+1}$$

where we use the fact that $\xi < \zeta < \gamma$. Similarly,

$$\zeta^{m+1} - \xi^{m+1} \leqq (\gamma - \xi)(m+1)\zeta^m.$$

The theorem follows from the lemma and the fact that $\zeta/\gamma < 1$. □

In the case where a is small but positive, linear interpolation is not as good. For example if $H \equiv 1$, $Q \equiv 0$, $v(a) = 1$, and $\alpha = a$, then $u_x(x) = (a/x)^m$ and the error for *linear* interpolation of $u(\xi)$ is

$$-\frac{m}{8}\frac{h^2}{a} + O(h^3)$$

and the error for $u_x(\xi)$ is

$$-\frac{m(m-1)}{24}\frac{h^2}{a^2}+O(h^3)$$

using the fact that $\xi=\gamma-(m/12)(h^2/\gamma)+O(h^4)$. Thus, the error is worse by a factor of $1/a$ than it is for ψ interpolation.

THEOREM 3. *For the singular case the ϕ interpolant satisfies*

$$\|u-U\|\leqq\frac{h^2\gamma}{m+1}C(\bar{P}) \quad \text{if } \beta\leqq c$$

and

$$\|u-U\|\leqq h^2C(P) \quad \text{in any case.}$$

Proof of first inequality. Setting $w(x)=u_x(x)/x$ and integrating by parts, we obtain

$$u(x)=u(\alpha)+\frac{x^2-\alpha^2}{2}w(x)-\int_\alpha^x\frac{y^2-\alpha^2}{2}w_y(y)\,dy.$$

A second equation is obtained by replacing α by β. Multiplication of the first equation by $\phi_\alpha(x)$ and the second by $\phi_\beta(x)$ yields

$$u(x)=U(x)-\phi_\alpha(x)\int_\alpha^x\frac{y^2-\alpha^2}{2}w_y(y)\,dy-\phi_\beta(x)\int_x^\beta\frac{\beta^2-y^2}{2}w_y(y)\,dy.$$

Note that

$$w_x=H_x\frac{v}{x}+H\left(\frac{v}{x}\right)_x, \qquad \frac{v(x)}{x}=\frac{1}{x^{m+1}}\int_0^x y^mQ(y)\,dy,$$

and

$$\left(\frac{v(x)}{x}\right)_x=\frac{Q(x)}{x}-\frac{m+1}{x^{m+2}}\int_0^x y^mQ(y)\,dy \tag{29}$$
$$=\frac{1}{x^{m+2}}\int_0^x y^{m+1}\bar{Q}_y(y)\,dy$$

where the overbar indicates a spatial derivative defined across meshpoints. Hence $|v(x)/x|\leqq\|Q\|^*/(m+1)$ and $|(v(x)/x)_x|\leqq\|\bar{Q}_x\|^*/(m+2)$ where $\|\cdot\|^*$ denotes the maximum norm on $[0,\beta]$. Combining these facts with (28), we obtain

$$|u(x)-U(x)|\leqq w(x)C(\bar{P})/(m+1)$$

where

$$w(x)=\phi_\alpha(x)\int_\alpha^x\frac{y^2-\alpha^2}{2}dy+\phi_\beta(x)\int_x^\beta\frac{\beta^2-y^2}{2}dy$$
$$=\frac{(\beta-x)(x-\alpha)(\alpha\beta+2\gamma x)}{6\gamma}<\frac{h^2}{8}\gamma. \qquad \square$$

Proof of second inequality. From (29) we have $|w_x(x)|\leqq(\|H_xv\|+2\|H\|\,\|Q\|^*)/x$ and so

$$|u(x)-U(x)|\leqq\tilde{w}(x)C(P)$$

where

$$\tilde{w}(x) = \phi_\alpha(x) \int_\alpha^x \frac{y^2-\alpha^2}{2}\frac{dy}{y} + \phi_\beta(x) \int_x^\beta \frac{\beta^2-y^2}{2}\frac{dy}{y}.$$

Using $1/y \leqq \beta/y^2$, we get

$$\tilde{w}(x) \leqq \beta(\beta-x)(x-\alpha)/(\beta+\alpha) \leqq \frac{h^2}{4}. \qquad \square$$

In an identical fashion it can be shown that (25) holds for the ϕ interpolant in the singular case.

THEOREM 4. *For the singular case the ϕ interpolant satisfies*

$$|u_x(\xi) - U_x(\xi)| \leqq \frac{h^2\gamma}{m+1} C(\bar{P}) \quad \text{if } \beta \leqq c$$

and

$$|u_x(\xi) - U_x(\xi)| \leqq \frac{m+1}{\gamma} h^2 C(P) \quad \text{in any case.}$$

Proof of first inequality. With $w(x) := u_x(x)/x$ we have

$$\begin{aligned} U_x(\xi) - u_x(\xi) &= \frac{2\xi}{\beta^2-\alpha^2} \int_\alpha^\beta (w(x) - w(\xi)) x\, dx \\ &= \frac{\xi}{h\gamma} \int_\alpha^\beta \int_\xi^x w_y(y)\, dy\, x\, dx \\ &= \frac{\xi}{h\gamma} \int_\alpha^\beta \int_\xi^x \left(w_x(\xi) + \int_\xi^y w_{zz}(z)\, dz \right) dy\, x\, dx. \end{aligned}$$

Using the results in the proof of Theorem 3, we have that $|w_x(x)| \leqq C(\bar{P})/(m+1)$ and

$$\begin{aligned} w_{xx} &= H_{xx}\frac{v}{x} + 2H_x\left(\frac{v}{x}\right)_x + H\left(\frac{v}{x}\right)_{xx}, \\ \left(\frac{v(x)}{x}\right)_{xx} &= \frac{1}{x^{m+3}} \int_0^x y^{m+2} \bar{Q}_{yy}(y)\, dy, \\ \bar{Q}_{xx} &= \bar{D}_{xx} u_t + 2\bar{D}_x (Hv)_t + D(\overline{Hv})_{xt} - \bar{f}_{xx}, \\ (\overline{Hv})_{xt} &= \left(\bar{H}_x v + HQ - mH\frac{v}{x} \right)_t, \\ \left(\frac{v}{x}\right)_t &= \frac{1}{x^{m+1}} \int_0^x y^m Q_t(y)\, dy, \\ Q_t &= D_t u_t + D u_{tt} - f_t, \end{aligned}$$

whence $|w_{xx}(x)| \leqq C(\bar{P})/(m+1)$. Therefore,

$$\begin{aligned} |u_x(\xi) - U_x(\xi)| &\leqq \frac{\xi}{h\gamma} \left(\left| \int_\alpha^\beta \int_\xi^x dy\, x\, dx \right| + \int_\alpha^\beta \int_\xi^x \int_\xi^y dz\, dy\, x\, dx \right) \frac{C(\bar{P})}{m+1} \\ &= \xi \left(|\xi - \gamma_1| + \frac{1}{h\gamma} \int_\alpha^\beta \frac{(x-\xi)^2}{2} x\, dx \right) \frac{C(\bar{P})}{m+1} \\ &\leqq \frac{h^2\gamma}{24} \frac{C(\bar{P})}{m+1} \end{aligned}$$

where the last inequality follows because $\xi = \gamma_1$. $\square$

Proof of second inequality. Modifying the previous proof, we get

$$U_x(\xi)-u_x(\xi)=\frac{\xi}{h\gamma}\int_\alpha^\beta\int_\xi^x\left(\xi^2 w_x(\xi)+\int_\xi^y (z^2 w_z)_z\,dz\right)\frac{dy}{y^2}\,x\,dx.$$

We have that

$$x^2 w_x = x\left(H_x v + HQ - (m+1)H\frac{v}{x}\right)$$

and

$$(x^2 w_x)_x = m\left((m+1)H\frac{v}{x} - HQ - 2H_x v\right) + x(H_{xx}v + 2H_x Q + HQ_x),$$

whence

$$|x^2 w_x| \leqq xC(P) \quad \text{and} \quad |(x^2 w_x)_x| \leqq (m+x)C(P).$$

Therefore,

$$|u_x(\xi)-U_x(\xi)| \leqq \frac{\xi}{h\gamma}\left(\left|\int_\alpha^\beta\int_\xi^x \frac{dy}{y^2}\,x\,dx\right|\xi + \int_\alpha^\beta\int_\xi^x\int_\xi^y (m+z)\,dz\frac{dy}{y^2}\,x\,dx\right)C(P).$$

When we use $\xi=\gamma_1$,

$$\text{first term} = h|\gamma-\xi| \leqq \frac{h^3}{12\gamma}.$$

The

$$\begin{aligned}\text{second term} &\leqq \int_\alpha^\beta\int_\xi^x\int_\xi^x (m+z)\,dz\frac{dy}{y^2}\,x\,dx\\ &= \frac{1}{\xi}\int_\alpha^\beta\left(m+\frac{x+\xi}{2}\right)(x-\xi)^2\,dx\\ &\leqq (m+\xi)\frac{h^3}{12\gamma}\end{aligned}$$

when we use $\xi=\gamma_1=\gamma+h^2/(12\gamma)$. Putting this together, we get

$$|u_x(\xi)-U_x(\xi)| \leqq \frac{h^3}{12\gamma}(m+1+\xi). \qquad \square$$

For the singular case linear interpolation is not as good. For example, if $H\equiv 1$, $Q\equiv 1$, and $\alpha=0$, then $u_x(x)=x/(m+1)$ and the error

$$\frac{u(h)-u(0)}{h}-u_x(\xi) = -\frac{mh}{2(m+1)(m+2)}.$$

A.2. Quadrature and lumping error. The local truncation errors due to quadrature and lumping are given by

$$\sigma = \int_\alpha^\beta (H(x) - H(\xi))v(x)\,dx,$$

$$\tau_\alpha = \int_\alpha^\beta \int_x^\beta \frac{dy}{y^m} x^m \{D(x)u_t(x) - f(x) - D(\xi)u_t(\alpha) + f(\xi)\}\,dx, \tag{30}$$

$$\tau_\beta = \int_\alpha^\beta \int_\alpha^x \frac{dy}{y^m} x^m \{D(x)u_t(x) - f(x) - D(\xi)u_t(\beta) + f(\xi)\}\,dx,$$

$$\tau = \int_\alpha^\beta x^m \{D(x)u_t(x) - f(x) - D(\xi)U_t(x) + f(\xi)\}\,dx.$$

Because there can be cancellation between τ and either τ_α or τ_β (see Theorem 8, Proof of first inequality), we obtain bounds on

$$\tau_R := \int_a^\alpha \frac{H(y)}{y^m}\,dy \cdot \tau + H_0\tau_\beta \tag{31}$$

in the regular case and on

$$\tau_S := \int_\beta^b \frac{H(y)}{y^m}\,dy \cdot \tau + H_0\tau_\alpha \tag{32}$$

in the singular case. These come from the worst case propagated local errors given by (21) and (22).

THEOREM 5. *With $\xi = \gamma_{-m}$ the truncation error* (8) *satisfies*

$$|\sigma| \leqq h^3 C(P).$$

Proof. From (30) we have

$$\sigma = \int_\alpha^\beta \int_\xi^x H_y(y)\,dy\,v(x)\,dx \tag{33}$$

$$= \int_\alpha^\beta \int_\xi^x \left(H_x(\xi) + \int_\xi^y H_{zz}(z)\,dz \right) dy\,v(x)\,dx.$$

Using $v(x) = x^{-m}(\alpha^m v(\alpha) + \int_\alpha^x y^m Q(y)\,dy)$ and integrating by parts, we obtain

$$\sigma = \int_\alpha^\beta (x-\xi)\frac{dx}{x^m} H_x(\xi)\alpha^m v(\alpha) + \int_\alpha^\beta \int_x^\beta (y-\xi)\frac{dy}{y^m} x^m H_x(\xi)Q(x)\,dx \tag{34}$$

$$+ \int_\alpha^\beta \int_\xi^x \int_\xi^y H_{zz}(z)\,dz\,dy\,v(x)\,dx.$$

Because $\xi = \gamma_{-m}$, the first term vanishes and the inner integral of the second term is positive. Therefore,

$$|\sigma| \leqq \int_\alpha^\beta \left\{ \int_x^\beta (y-\xi)\frac{dy}{y^m} x^m + \int_\xi^x \int_\xi^y dz\,dy \right\} dx \cdot C(P)$$

$$= \int_\alpha^\beta \frac{x-\xi}{x^m} \frac{x^{m+1} - \alpha^{m+1}}{m+1}\,dx \cdot C(P) + \frac{1}{2}\int_\alpha^\beta (x-\xi)^2\,dx \cdot C(P).$$

Because $\xi = \gamma_{-m}$,

$$\text{first term} = \frac{1}{m+1}\int_\alpha^\beta (x-\xi)x\,dx\,C(P)$$

$$= \frac{H\gamma}{m+1}(\gamma_1 - \xi)C(P)$$

$$\leqq \frac{h^3}{4}C(P)$$

when we use the lemma. The second term is bounded by $(h^3/6)C(P)$. □

In the case where a is small but positive, a choice of ξ other than γ_{-m} is not likely to be as good. For example, if $H(x) = 1+x$, $Q \equiv 0$, $v(a) = 1$, and $\alpha = a$, then the error

$$\sigma = (\gamma_{-m} - \xi)a^m \int_a^\beta \frac{dx}{x^m};$$

and if, for example, $\xi = \gamma$, then

$$\sigma = -\frac{m}{12}\frac{h^3}{a} + O(h^4).$$

This is worse by a factor of $1/a$ than the bound given by Theorem 5.

THEOREM 6. *For the singular case with* $\xi = \gamma_1$

$$|\sigma| \leqq \frac{h^3\gamma}{m+1}C(\bar{P}) \quad \textit{if } \beta \leqq c$$

and

$$|\sigma| \leqq h^3 C(P) \quad \textit{in any case.}$$

Proof of first inequality. Using the fact that $\xi = \gamma_1$, we have from (33) that

$$\sigma = \int_\alpha^\beta (x-\xi)x \int_\xi^x \left(\frac{v}{y}\right)_y dy\,dx \cdot H_x(\xi) + \int_\alpha^\beta x \int_\xi^x \int_\xi^y H_{zz}(z)\,dz\,dy \frac{v(x)}{x}\,dx.$$

Using bounds from the proof of Theorem 3, we obtain

$$|\sigma| \leqq \frac{3}{2}\int_\alpha^\beta x(x-\xi)^2\,dx \frac{C(\bar{P})}{m+1} \leqq \frac{h^3}{8}\gamma\frac{C(\bar{P})}{m+1}$$

where the last inequality follows because $\xi = \gamma_1$. □

Proof of second inequality. Equation (34) gives

$$|\sigma| \leqq \left(\left|\int_\alpha^\beta \frac{x-\xi}{x^m}dx\right| \frac{\alpha^{m+1}}{m+1} + \int_\alpha^\beta \left|\int_x^\beta \frac{y-\xi}{y^m}dy\right| x^m\,dx\right)\|H_x\|\,\|Q\|$$

$$+ \int_\alpha^\beta \frac{(x-\xi)^2}{2}dx\,\|H_{xx}\|\,\|v\|.$$

Let $\eta < \beta$ be such that

$$\int_\eta^\beta \frac{x-\xi}{x^m}dx = 0.$$

Then

$$|\sigma| \leqq \left(-\int_\alpha^\eta \frac{x-\xi}{x^m}\, dx \frac{\alpha^{m+1}}{m+1} - \int_\alpha^\eta \int_x^\eta \frac{y-\xi}{y^m}\, dy\, x^m\, dx + \int_\eta^\beta \int_x^\beta \frac{y-\xi}{y^m}\, dy\, x^m\, dx \right) C(P)$$
$$+ \frac{h^3}{6} C(P).$$

Interchanging the order of integration in the two double integrals and using twice again the definition of η, we obtain

$$|\sigma| \leqq 2 \int_\eta^\beta \frac{x-\xi}{x^m} \frac{x^{m+1}-\eta^{m+1}}{m+1}\, dx \cdot C(P) + \frac{h^3}{6} C(P)$$
$$\leqq 2 \int_\eta^\beta (x-\xi)(x-\eta)\, dx \cdot C(P) + \frac{h^3}{6} C(P).$$

The

$$\text{integral} \leqq \int_\xi^\beta (x-\xi)(x-\eta)\, dx$$
$$= \frac{1}{3}(\beta-\xi)^3 + \frac{1}{2}(\beta-\xi)^2(\xi-\eta) \leqq \frac{5}{6}(\beta-\xi)^3 \leqq \frac{5}{48} h^3. \qquad \square$$

A choice of ξ other than γ_1 is not likely to be as good. If $H(x) = 1+x$ and $Q \equiv 1$, then the error

$$\sigma = \frac{\gamma_1 - \xi}{m+1} h\gamma;$$

and if, for example, $\xi = \gamma$, then

$$\sigma = \frac{h^3}{12(m+1)}.$$

This is worse by a factor of $1/\gamma$ than the bound given by the first inequality of Theorem 6.

THEOREM 7. *For the regular case with $\xi = \gamma_{-m}$ the propagated truncation error τ_R defined by equation* (31) *satisfies*

$$|\tau_R| \leqq \left(1 + m\beta^{m-1} \int_a^\beta \frac{dy}{y^m} \right) h^3 C(P).$$

Proof. We can write $\tau = \tau_1 + \tau_2 + \tau_3$ where

$$\tau_1 = \int_\alpha^\beta x^m (D(x) - D(\xi))(u_t(x) - u_t(\xi))\, dx,$$

$$\tau_2 = \int_\alpha^\beta x^m \{(D(x) - D(\xi)) u_t(\xi) - (f(x) - f(\xi))\}\, dx,$$

$$\tau_3 = D(\xi) \int_\alpha^\beta x^m (u_t(x) - U_t(x))\, dx.$$

Clearly,

$$|\tau_1| \leqq \int_\alpha^\beta x^m (x-\xi)^2\, dx\, \|D_x\|\, \|u_{xt}\|.$$

Expanding D and f in τ_2 in a Taylor series about $x=\xi$, we obtain $\tau_2=\tau_2'+\tau_2''$ where

$$\tau_2'=\int_\alpha^\beta x^m(x-\xi)\,dx\,(D_x(\xi)u_t(\xi)-f_x(\xi))$$

and

$$\tau_2''=\int_\alpha^\beta x^m\frac{(x-\xi)^2}{2}(D_{xx}(\xi')u_t(\xi)-f_{xx}(\xi'))\,dx.$$

Applying the lemma, we have

$$|\tau_2'|\leqq\int_\alpha^\beta x^m\,dx(\gamma_m-\gamma_{-m})C(P)$$

$$\leqq\frac{m}{2}h^3\beta^{m-1}C(P).$$

Also we have

$$|\tau_2''|\leqq\frac{1}{2}\int_\alpha^\beta x^m(x-\xi)^2\,dx\,(\|D_{xx}\||u_t(\xi)|+\|f_{xx}\|)$$

and

$$|\tau_3|\leqq|D(\xi)|\int_\alpha^\beta x^m\,dx\,\|u_t-U_t\|,$$

and so

$$|\tau|\leqq\left(1+\frac{m}{\beta}\right)h^3\beta^mC(P).$$

The other part of τ_R can be expressed

$$\tau_\beta=\int_\alpha^\beta\int_\alpha^x\frac{dy}{y^m}x^m\{Q(x)-Q(\xi)+D(\xi)(u_t(\xi)-u_t(\beta))\}\,dx,$$

and so

$$|\tau_\beta|\leqq h^2\beta^m\int_\alpha^\beta\frac{dx}{x^m}C(P).$$

Note that for $x\leqq\beta$

$$\beta^m=x^m+\beta^m-x^m\leqq x^m+m(\beta-x)\beta^{m-1}\leqq x^m+mh\beta^{m-1},$$

and so

$$\beta^m\int_\alpha^\beta\frac{dx}{x^m}\leqq h+mh\beta^{m-1}\int_\alpha^\beta\frac{dx}{x^m}.$$

Therefore,

$$|\tau_R|\leqq h^3\left(\beta^m\int_a^\alpha\frac{dx}{x^m}+m\beta^{m-1}\int_a^\beta\frac{dx}{x^m}+1\right)C(P).$$

However,

$$\beta^m\int_a^\alpha\frac{dx}{x^m}\leqq\alpha-a+m(\beta-a)\beta^{m-1}\int_a^\alpha\frac{dx}{x^m}\leqq1+m\beta^{m-1}\int_a^\beta\frac{dx}{x^m},$$

from which the result follows. □

THEOREM 8. *For the singular case with $\xi = \gamma_1$ the propagated truncation error defined by equation* (32) *satisfies*

$$|\tau_S| \leqq \left(\beta \log \frac{c}{\beta} + \frac{\beta}{c}\right) h^3 C(\bar{P}) \quad \text{if } \beta \leqq c \text{ and } m \geqq 2$$

and

$$|\tau_S| \leqq h^3 C(P) \quad \text{in any case.}$$

Proof of first inequality. We must modify the proof of Theorem 7. We have

$$\tau = \tau_2' + (\tau_1 + \tau_2'' + \tau_3)$$

where

$$|\tau_1 + \tau_2'' + \tau_3| \leqq h^2 \int_\alpha^\beta x^m \, dx \, C(P). \tag{35}$$

Also

$$\tau_\alpha = \int_\alpha^\beta \int_x^\beta \frac{dy}{y^m} x^m (x-\xi) \, dx \, (D_x(\xi) u_t(\xi) - f_x(\xi)) + \tau_\alpha'$$

where

$$\tau_\alpha' = \int_\alpha^\beta \int_x^\beta \frac{dy}{y^m} x^m \left\{ (D(x) - D(\xi))(u_t(x) - u_t(\xi)) + \frac{(x-\xi)^2}{2} (D_{xx}(\xi') u_t(\xi) - f_{xx}(\xi')) + D(\xi)(u_t(x) - u_t(\alpha)) \right\} dx.$$

Use of the fact that $|u_{tx}(x)| \leqq xC(P)/(m+1)$ gives us

$$\begin{aligned} |\tau_\alpha'| &\leqq \int_\alpha^\beta \int_x^\beta \frac{dy}{y^m} x^m \left\{ h^2 + \frac{1}{m+1} \int_\alpha^x y \, dy \right\} dx \cdot C(P) \\ &\leqq \left\{ \int_\alpha^\beta \frac{x^{m+1} - \alpha^{m+1}}{(m+1)x^m} \, dx + \frac{1}{m+1} \int_\alpha^\beta y \, dy \right\} h^2 C(P) \\ &\leqq h^3 \gamma C(P)/(m+1). \end{aligned}$$

We have

$$\int_\beta^1 \frac{H(y)}{y^m} \, dy = \int_\beta^c \frac{H_0}{y^m} \, dy + \int_\beta^c \frac{H(y) - H_0}{y^m} \, dy + \int_c^1 \frac{H(y)}{y^m} \, dy$$

where the

$$|\text{last two terms}| \leqq \int_\beta^c \frac{dy}{y^{m-1}} \|\bar{H}_x\|^* + \int_c^1 \frac{dy}{y^m} \|H\|^*.$$

Combining all this, we obtain

$$\begin{aligned} |\tau_S| \leqq & \left| \int_\alpha^\beta \int_x^c \frac{dy}{y^m} x^m (x-\xi) \, dx \, H_0 (D_x(\xi) u_t(\xi) - f_x(\xi)) \right| \\ & + \left(\int_\beta^c \frac{dy}{y^{m-1}} \|\bar{H}_x\|^* + \int_c^1 \frac{dy}{y^m} \|H\|^* \right) |\tau_2'| \\ & + \int_\beta^1 \frac{H(y)}{y^m} \, dy \, h^2 \int_\alpha^\beta x^m \, dx \, C(P) + H_0 h^3 \gamma C(P)/(m+1). \end{aligned}$$