Henry H. Rachford, Jr.

Rounding errors in alternating direction methods for parabolic problems

Aplikace matematiky, Vol. 13 (1968), No. 2, 177–180

Persistent URL: `http://dml.cz/dmlcz/103152`

ROUNDING ERRORS IN ALTERNATING DIRECTION METHODS FOR PARABOLIC PROBLEMS

H. H. RACHFORD, JR.

Recently, the rouding error growth in solving a Crank-Nicolson difference analogue of a general second order parabolic problem with smooth coefficients in one space variable was analyzed [1]. It was shown that to maintain a fixed bound on rounding-induced errors the word length of the floating mantissa must be increased in proportion to the logarithm of the number of time-distance mesh points as the time and distance steps, k and h, are taken to zero at constant k/h. The present work shows that the analysis can be extended to the p-dimensional case when the computation is done using a stable, consistent, two-level alternating direction procedure. In this case, the required increase in word length is proportional to $\log (p^2\overline{N}\overline{M}^2)$ where $\overline{N}$ is the maximum number of grid points in any line in R_h, the mesh covering the spatial domain, and $\overline{M}$ is the number of time steps.

Let $L(u) \equiv \sum_{i=1}^{p} [(\partial/\partial x_i)(\bar{\alpha}(x, t)(\partial u/\partial x_i)) + \xi(x, t)(\partial u/\partial x_i)] + \gamma(x, t)\, u$, and consider

$$L(u) = \frac{\partial u}{\partial t} + f(x, t) \tag{1}$$

in a bounded region $R \times (0, T]$, where $R \subset \mathbf{R}^p$, $\bar{\alpha}$, ξ, and γ are scalar valued continuous function of $x \in R^p$ and time, t, $0 < \alpha_0 \leqq \bar{\alpha} \leqq \alpha_m$, $\gamma \leqq 0$, and u is specified such that fourth distance derivatives of u are bounded. We consider the operators

$$-L_{h_i} w(P,t) \equiv \overline{\nabla}_i(\bar{\alpha}(P_i^{+1/2}, t)\, \nabla_i\, w(P,t)) + \left(\frac{1}{2h_i}\right) \xi(P,t)\, [w(P_i^+, t) - w(P_i^-, t)] + \gamma(P,t)\, w(P,t)/p \tag{2}$$

where the grid of points R_h over R is generated by the increment vector $h = (h_1, \ldots, h_p)$, $P \in R_h$ is defined by $P = (x_1, \ldots, x_i, \ldots, x_p)$, $P_i^{\pm} = (x_1, \ldots, x_i \pm h_i, \ldots, x_p)$, $P^{\pm 1/2}(x_1, \ldots, x_i + h_{i/2}, \ldots, x_p)$, $\nabla_i\, w(P) = [w(P_i^+) - w(P)]\, h_i^{-1}$ and $\overline{\nabla}_i\, w(P) = \nabla_i\, w(P_i^-)$. The Crank-Nicolson difference analogue of (1) becomes

$$w(P, t_{n+1}) + \tfrac{1}{2}k \sum_{i=1}^{p} L_{h_i}[w(P, t_{n+1}) + w(P, t_n)] = w(P, t_n) - kf(P, t_n + \tfrac{1}{2}k)\,, \tag{3}$$

which relates the values of the approximation w at points of $(R_h \cup C_h) \times \{t_n\}$, where $t_n = nk$, $n = 0, 1, \ldots, K-1$, $K = T/k$, and where the points C_h are points on $\partial R \times \{t_n\}$. We let N_h be the number of points of R_h and $\|v\| = (h_1, \ldots, h_p \sum_{R_h} v_i^2)^{1/2}$ for all $v \in \mathbf{R}^{N_n}$. The relation (2) is evidently of the form

$$(4) \qquad (I + A)\, w_{n+1} + B w_n = g_n\,, \quad n = 0, 1, \ldots, K-1\,,$$

where A and B depend also on n. Letting $\sum_{i=1}^{p} A_i = A$ and noting that $(I + A_i)\, w = z$ is readily solved, the alternating direction form of (4) is

$$(5a) \qquad (I + A_1)\, \beta_{n+1}^{(1)} + \sum_{j=2}^{p} A_j \beta_n + B_{\beta_n} = g_n$$

$$(5b) \qquad (I + A_i)\, \beta_{n+1}^{(i)} = \beta_{n+1}^{(i-1)} + A_i \beta_n\,, \quad i = 2, \ldots, p\,,$$

and the approximation for u_{n+1}, β_{n+1} is taken to be $\beta_{n+1}^{(p)}$.

The computations using (5) produce not $\{\beta_n\}$ but a sequence $\{\hat{\beta}_n\}$, which differs from $\{\beta_n\}$ due to rounding. We follow the type of analysis of WILKINSON [2] and write $\beta_{n+1}^{(i)} = Q_i d_i$, $\hat{\beta}_{n+1}^{(i)} = \hat{Q}_i \hat{d}_i$, where

$$d_1 = g_n - (B + A - A_1)\, \beta_n\,, \quad d_i = \beta_{n+1}^{(i-1)} + A_i \beta_n\,, \quad i = 2, \ldots, p\,,$$

and

$$\hat{d}_1 = g_n - (B + A - A_1)\, \hat{\beta}_n + e_1\,, \quad \hat{d}_i = \hat{\beta}_{n+1}^{(i-1)} + A_i \hat{\beta}_n + e_i\,, \quad i = 2, \ldots, p\,,$$

where e_i is the error introduced in computing $\hat{d}_i$ from the stated arguments, $Q_i = (I + A_i)^{-1}$, and $\hat{Q}_i$ is a matrix approximating Q_i whose existence and exact form depend upon the procedure used to solve (5).

We assume several quantities relevant to the problem to be solved:

$$(6a) \qquad \|(I + A_i)^{-1}\| < 1/\delta\,, \quad \delta > 0\,,$$

$$(6b) \qquad \max\,(\|g_n\| + 2\sum_i \|A_i \beta_n\| + \|B\beta_n\|, \|\beta_n\|) \leqq \beta\,,$$

$$(6c) \qquad \|R_i\| \leqq M(\tau)\,,$$

where $R_i = \hat{Q}_i Q_i^{-1} - I$, and τ is the number of floating base N digits in the mantissa. The existence of β and δ follow from consistency and stability of (5).

From (5), (6), and an examination of $\prod_{i=j}^{q} \hat{Q}_i - \prod_{i=j}^{q} Q_i$, we conclude that

$$(7) \qquad \|v_{n+1}\| \leqq [(1 + M)^p - 1]\delta^{-p} [g_n + \sum_{j=1}^{p} \|A_j \beta_n\| + \|(B + A)\, \beta_n\|] + \\ + \varrho(\varrho^p - 1)(\varrho - 1)^{-1} \eta_n + [\|G\| + \|\sum_{j=1}^{p} (\prod_{i=j}^{p} \hat{Q}_i - \prod_{i=j}^{p} Q_i)\, A_j - \\ - (\prod_{j=1}^{p} \hat{Q}_j - \prod_{j=1}^{p} Q_j)(B + A)\|]\, \|v_n\|$$

where $v_n = \hat{\beta}_n - \beta_n$, $\varrho = (1 + M)/\delta$, η_n is a bound on $\|e_i\|$, and $G \equiv \sum_{j=1}^{p} \prod_{i=j}^{p} Q_i A_j - \prod_{k=1}^{p} Q_k(B + A)$. It will be seen to be important below that G is the matrix such that $\beta_{n+1} = G\beta_n + Hg_n$ from (5); hence, by stability of (5), $\|G\| \leqq 1 + C_0 k$ for all n.

Using the methods of (2), we find that

$$\eta_n = [(k_1 + S)\beta + (k_2 + \beta\delta^{-p} + \alpha)\|v_n\|]\, v/(1 - \zeta v)\,, \tag{8}$$

where

$$k_1 = \max\{1 + (1 + v) N_0[\| |B| \| + a(p - 1)]\,, \qquad [1 + (1 + v) aN_0]\}\,,$$
$$k_2 = \max\{[\| |B| \| + a(p - 1)][1 + (1 + v) N_0]\,, \quad a[(1 + v) N_0 + 1]\}\,,$$

and a is a bound on $\| |A_i| \|$, $v = sN^{1-\tau_1}$, $s = \frac{1}{2}$ or 1 as rounding or chopping occurs in storage, $\tau_1 = \tau - \log_N 1.053$, N_0 is the maximum number of sums taken for any element of any matrix-by-vector multiplication in d_i, $\mu = \varrho^p - \delta^{-p}$, $S = \mu + \delta^{-p}$, $\alpha = \mu Y$, $Y = A \sum_{j=1}^{p} \|A_j\| + \|B + A\|$, and $\zeta = \varrho(\varrho^p - 1)(\varrho - 1)^{-1}$, It follows from (8) that

$$\|v_n\| \leqq \varphi_2(\varphi_1^n - 1)(\varphi_1 - 1)^{-1}\,, \tag{9}$$

where $\varphi_1 = [\|G\| + \alpha + v\zeta(k_2 + \alpha + \beta\delta^{-p})(1 - \zeta v)^{-1}]$ and $\varphi_2 = [\mu + \zeta v(k_1 + S)(1 - \zeta v)^{-1}]\,\beta$. We assume now that h is fixed and that the computations are carried out sothat M decreases at least linearly with v. Expansion of α shows that $\alpha = c_1' M + O(M^2)$ for M small. Thus, if $M = \check{c}_1 v/c_1'$ and we choose $v = \check{c}_1 k$, then $\alpha \leqq c_1 k$; hence, $\varphi_1 = 1 + c_3 k + O(k^2)$. Since $\mu = c_4 k + O(k^2)$, $\varphi_2 \leqq \beta c_5 k$ for k small, and

$$\|v_n\| \leqq c_5 T e^{c_3 T}\,. \tag{10}$$

This is satisfactory as it is exactly the same result that would obtain were $L(u) = \partial u/\partial t$ an ordinary differential equation in t.

The question of real interest arises when $h/k = c$ while $k \to 0$. A suitable ordering of $P \in R_h$ yields A_i as a diagonal set of m tridiagonal blocks, each irreducible for h_i sufficiently small, where m is the number of physical rows of points of R_h in R associated with the ith direction. Thus, the solution of $(I + A_i)\,w = z$ is the solution of m independent tridiagonal systems of the form

$$\begin{pmatrix} b_1, c_1, 0, \ldots, 0, 0 \\ a_2, b_2, c_2, \ldots, 0, 0 \\ \cdots\cdots\cdots\cdots \\ 0, 0, 0, \ldots, a_j, b_j \end{pmatrix} \begin{pmatrix} w_1 \\ \cdot \\ \cdot \\ w_j \end{pmatrix} \equiv \Gamma_s w = r_i \bar{d} = d\,; \quad s = 1, \ldots, m\,, \tag{11}$$

where $\bar{d}$ is an m-segment of z and r_i is a normalizing factor so that for h_i

sufficiently small

(12) i) $\tilde{\delta} + |a_j| + |c_j| < b_j(1 - 4v - 3v^2 - v^3)$; $j = 1, \ldots, J$,

ii) $-1 \leqq a_i$, $c_j < 0$; $j = 1, \ldots, J - 1$; $i = 2, \ldots, J$, the left hand equality holding for some row of some Γ_s; $s = 1, \ldots, m$,

iii) $a_1 = c_J = 0$,

iv) $\tilde{\delta} > 0$.

It is easy to see that $\|\Gamma_s\|_\infty < \tilde{\delta}^{-1}$, hence, $\delta = 1$ suffices for (6a). Analysis of the floating point operations involved in (11) shows [1] indeed that $\hat{Q}_i$ does exist with M of (6c) given by

$$M = (15 + 2\|\Gamma\|_\infty)\, v\tilde{\delta}^{-1} + O(v\tilde{\delta}^{-1})^2 .$$

Taking $v = c_2 h_j k^2$, $h_j \leqq h_i$, assuming $h_i/k = c$ fixed as $k \to 0$ leads to $v\tilde{\delta}^{-1} = \alpha_m c_2 k^2/2c + O(k^3)$, where $\alpha_m = \max_{P \in R_h} \bar{\alpha}(P_i^{\pm 1/2})$. For k small, φ_1 and φ_2 of (9) now satisfy: $\varphi_1 \leqq 1 + c'_3 k$, and $\varphi_2 \leqq c'_5 k^2$ for any $c_{\prime}^3 > c_0 + 69\alpha_m^2 p^2 c_2 c^{-3}$, and $c'_5 > c_2 c^{-1}\, \alpha_m \beta p(12p - \frac{11}{2})$.

The following theorem follows from the analysis outlined above.

Theorem. *Let* (1) *be solved in a hypercube using* (5) *which is assumed to be stable and consistent with* c_0 *independent of* p. *Computation is performed with* τ*-digit floating-N arithmetic. If* $N^{-\tau} = \hat{c}_2 h_j p^{-2} k^2$, $h_j \leqq h_i$, $i = 1, 2, \ldots p$, $h_j = ck$, *and if* $\hat{\beta}_n$ *and* β_n *are the computed and exact solutions for* (5), *respectively, then as* $k \to 0$

$$\|\hat{\beta}_n - \beta_n\| \leqq k c''_5 T e^{c''_3 T} ,$$

where $c''_3 > c_0 + 73s\, N\bar{\alpha}_M^2 \hat{c}_2 c^{-3}$, $c''_5 > 1{\cdot}053 sN\, \hat{c}_2 c^{-1} \bar{\alpha}_M \beta[12 - 11(2p)^{-1}]$ *and* s *is* $\frac{1}{2}$ *or* 1 *as rounding or truncation occurs, respectively.*

Although the analysis has ignored the variations of A and B with n, we need only note that the bounds may be interpreted over all n, and that stability implies $\|G_n\| \leqq 1 + C_0 k$ independent of n to complete the proof. Further, the analysis does not assume symmetry of A_i, but only the inequalities (12). Thus, for any shape region approximated with difference relations of positive type we shall expect the theorem to hold.

References

[1] *Rachford, H. H. Jr.*, Rounding Errors in Parabolic Problems, Part I. The one space variable case, to appear.

[2] *Wilkinson, J. H.* Rouding Errors in Algebraic Processes, Prentice Hall, 1963.

H. H. Rachford, Jr., Rice University, P. O. Box 1892, Houston, Texas 77001, U.S.A.