Numerical Analysis of Boundary-Layer Problems in Ordinary Differential Equations

By W. D. Murphy

1. Summary. We categorize some of the finite-difference methods that can be used to treat the initial-value problem for the boundary-layer differential equation

$$\mu y' = f(y, x) ; \qquad y(0) = y^0 . \tag{1}$$

These methods take the form

$$\sum_{i=0}^{k} \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^{k} \beta_i f(Y_{n+i}, x_{n+i}) + R_n , \tag{2}$$

where α_ν and β_ν ($\nu = 0, 1, \cdots, k$) denote real constants which do not depend upon n, R_n is the round-off error, $\mu = h^\gamma$, $0 < \gamma < 1$, and h is the mesh size. We define a new kind of stability called μ-stability and prove that under certain conditions μ-stability implies convergence of the difference method. We investigate μ-stability and the optimal methods which it allows, i.e., methods of maximum accuracy.

The idea of relating μ to h allows us to study the nature of the difference equation for very small μ. We can, however, look at this in another way. Given a differential equation in the form of Eq. (1) we ask how can we choose h so that the associated difference equation will give an accurate approximation. If μ is sufficiently small, choose h by the formula $h = \mu^{1/\gamma}$ where $0 < \gamma < 1$.

2. Boundary-Layer Phenomena. Eq. (1) characterizes the boundary-layer problem for the first order in one unknown. The small interval near the initial point ($x = 0$) where the slope of the curve $y(x, \mu)$ is changing most rapidly is called the boundary layer. An estimate of this interval is $[0, -A\mu \ln \mu]$ where A is a positive constant that is independent of μ. This theory has been well investigated in recent years and a rather complete study can be found in Vasil′eva [7]. We briefly describe the treatment found there.

We first introduce some definitions. Let $y = \phi(x)$ be one of the solutions of the degenerate equation $f(y, x) = 0$.

Definition. The root $y = \phi(x)$ is *isolated* on the set [0, 1] if there exists an $\epsilon > 0$ such that $f(y, x) = 0$ has no solution other than $\phi(x)$ for $|y - \phi(x)| < \epsilon$.

Definition. The isolated root $y = \phi(x)$ will be called *positively stable* in [0, 1] if $\partial f(\phi(x), x)/\partial y \leqq -\bar{L} < 0$ for all $x \in [0, 1]$.

Definition. The *domain of influence* of an isolated positively stable root $y = \phi(x)$ is the set of points (y^*, x^*) such that the solution to the adjoined equation

$$dy/d\tau = f(y, x^*) \tag{3}$$

(x^* is regarded as a parameter) satisfying the initial conditions $y\,|_{\tau=0} = y^*$ tends to the value $\phi(x^*)$, as $\tau \rightarrow \infty$.

Received November 9, 1966. Revised January 30, 1967.

The main theorem about boundary-layer equations is the following:

THEOREM 1. *If some root* $y = \phi(x)$ *of the equation* $f(y, x) = 0$ *is an isolated positively stable root in* $[0, 1]$, *and if the initial point* $(y^0, 0)$ *belongs to the domain of influence of this root, then the solution* $y(x, \mu)$ *of Eq.* (1) *tends to the* $\phi(x)$ *of the degenerate equation, as* $\mu \to 0$, *for* $0 < x \leqq 1$.

Proof. See Vasil′eva [7]. The paper by Vasil′eva [7] goes on to explain how to find an asymptotic expansion of the solution of Eq. (1) in terms of the small parameter μ. Here in addition to the conditions of Theorem 1 we assume that $f(y, x)$ has continuous partial derivatives of order up to $n + 2$. With this condition, Vasil′eva finds an asymptotic expansion for $y(x, \mu)$ which contains n terms. Inside the boundary layer $[0, -A\mu \ln \mu]$, where A is a constant independent of μ, each term of the asymptotic expansion contains three functions found by solving three separate differential or transcendental equations. Outside the boundary layer $(-A\mu \ln \mu, 1]$ the terms are much simpler and can be determined from the variational equations. This procedure for finding the asymptotic expansion is a very tedious one and can only be explicitly calculated for the simplest problems.

It is the aim of this paper to tie together the known numerical analysis theory with the boundary layer theory in such a way that this problem can be solved with computers even as $\mu \to 0$. If we attempt to apply the standard proof of convergence to the difference Eq. (2), we run into serious difficulties because the following limit occurs:

$$(4) \qquad \lim_{h\to 0; x_n = nh} (1 + Mh^{1-\gamma})^n = \infty ,$$

where M is a positive constant. However, if we are a bit more careful, we can make use of the fact that

$$(5) \qquad \lim_{h\to 0; x_n = nh} (1 - Mh^{1-\gamma})^n = 0 .$$

This limit will be directly related to the condition $-\bar{L} \leqq (\partial f/\partial y) - \underline{L} < 0$. A price is paid for the privilege of using Eq. (5); namely, we must restrict ourselves to a smaller class of difference equations than is generally done in ordinary differential equations (ODE). In fact, this class will contain optimal methods of order at most $k + 1$ instead of $k + 2$ as is the case in Dahlquist [1]. See Murphy [6] for the proof of this last result.

3. μ-Stability.

We associate with the difference Eq. (2) two polynomials

$$\rho(\zeta) = \alpha_k\zeta^k + \alpha_{k-1}\zeta^{k-1} + \cdots + \alpha_0 \qquad (\alpha_k \neq 0) ,$$
$$\sigma(\zeta) = \beta_k\zeta^k + \beta_{k-1}\zeta^{k-1} + \cdots + \beta_0 ,$$

and we assume for convenience that $\rho(\zeta)$ and $\sigma(\zeta)$ have no common factors. Furthermore, our consistency condition is that $\rho(1) = 0$ and $\rho'(1) = \sigma(1)$. The stability condition proposed by Henrici [2] and Hull and Luxemburg [3] is that the roots of $\rho(\zeta) = 0$ lie in or on the unit circle in the complex ζ-plane, and are simple if they lie on the circle.

This stability condition is not satisfactory for us, as can easily be seen by looking at the difference Eq. (2) without the terms R_n associated with the differential equation

$$\mu y' = -L_0 y\,, \tag{6}$$

where L_0 is a positive constant. Here the solution takes the form

$$Y_n = C_0\zeta_0^n + C_1\zeta_1^n + \cdots + C_{k-1}\zeta_{k-1}^n\,, \tag{7}$$

where the C_i's are constants depending on the initial conditions and

$$\zeta_j = \zeta_{j0} + \left(-m!\,\frac{\sigma(\zeta_{j0})}{\rho^{(m)}(\zeta_{j0})}\,L_0h^{1-\gamma}\right)^{1/m} + O(h^{2(1-\gamma)/m})\,, \tag{8}$$

where $\rho(\zeta_{j0}) = 0$ and m is the multiplicity of the root ζ_{j0}.

It is clear that roots of $\rho(\zeta)$ which lie inside the unit circle will not cause any problems with regard to boundedness of the solution of the difference equation. However, the simple roots on the unit circle may lead to divergent methods. $\zeta = 1$ is always an acceptable root by the consistency condition. Furthermore, the condition

$$\sigma(-1)/\rho'(-1) < 0 \tag{9}$$

insures boundedness of $|\zeta_j|^n$ for the root $\zeta_{j0} = -1$.

If $\zeta_{j0} = e^{i\theta_j}$ and $p_j + iq_j = -\sigma(e^{i\theta_j})/\rho'(e^{i\theta_j})$ then

$$|\zeta_j^n| = |\zeta_j|^n = [1 + 2L_0(p_j\cos\theta_j + q_j\sin\theta_j)h^{1-\gamma} + O(h^{2-2\gamma})]^{n/2}\,.$$

Therefore, we require

$$p_j\cos\theta_j + q_j\sin\theta_j < 0\,. \tag{10}$$

Inequalities (9) and (10) in addition to stability will categorize a new kind of stability which we choose to call μ-stability.

If we have m roots on the unit circle, the condition (10) reduces to $m/2$ conditions because we are dealing with complex conjugates. See Murphy [6] for the details.

The condition of μ-stability can be thought of as merely conditions on the coefficients, β_ν. An example will clarify this point.

Example 1. Let $\rho(\zeta) = \zeta^2 - 1$; the roots are $\zeta = \pm 1$. By consistency

$$\rho'(1) = 2 = \beta_2 + \beta_1 + \beta_0\,. \tag{11}$$

By condition (9)

$$\sigma(-1)/\rho'(-1) = (\beta_2 - \beta_1 + \beta_0)/(-2) < 0\,. \tag{12}$$

Combining Eqs. (11) and (12) gives

$$\beta_1 < 1\,. \tag{13}$$

Thus the inequality (13) is equivalent to the condition of μ-stability for this example. Note that Simpson's rule ($\beta_1 = 4/3$) is not μ-stable.

In the analysis to follow it will be desirable to also consider a stronger kind of stability called relative stability.

Definition. A difference scheme characterized by the polynomials $\rho(\zeta)$ and $\sigma(\zeta)$ will be called *relatively stable* if the roots of $\rho(\zeta) + h^{1-\gamma}\sigma(\zeta) = 0$ have the property that

$$(14) \qquad |\zeta_i| \leqq |\zeta_0| = 1 - h^{1-\gamma} + O(h^{2(1-\gamma)}) , \qquad i = 1, 2, \cdots, k-1 .$$

For a relatively stable scheme we must require

$$(15) \qquad \sigma(-1)/\rho'(-1) < -1$$

and

$$(16) \qquad p_j \cos\theta_j + q_j \sin\theta_j < -1/2 .$$

4. Convergence. A few lemmas will be required for the main theorem of this paper.

LEMMA 1. *Let the consistent difference equation*

$$\sum_{i=0}^{k} \alpha_i Y_{n+i} = -L_0 h^{1-\gamma} \sum_{i=0}^{k} \beta_i Y_{n+i}$$

be μ-stable. L_0 is a positive constant with the property that $-\overline{\overline{L}} \leqq -L_0 \leqq -\overline{L} < 0$. Let Φ_i be the solution of this difference equation with initial conditions

$$\Phi_0 = \Phi_1 = \cdots = \Phi_{k-2} = 0, \qquad \Phi_{k-1} = (\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1} .$$

Then, for all $n > 1$,

$$(17) \qquad \sum_{i=0}^{n-1} |\Phi_i| \leqq \frac{C}{h^{1-\gamma}}$$

for h sufficiently small and where C, a constant, may be chosen independent of h and L_0.

Proof. The solution to the difference equation is given by Eqs. (7) and (8). By Cramer's rule, we can write

$C_j = D_j/W$, where D_j and W reduce to Vandermonde determinants. Consequently,

$$|D_j| = |(\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1}| \prod_{t<s;s\neq j;t\neq j} |\zeta_s - \zeta_t|$$

and $W = \prod_{j<i} (\zeta_i - \zeta_j)$. Therefore,

$$|C_j| = \frac{|(\alpha_k + h^{1-\gamma} L_0 \beta_k)^{-1}|}{\prod_{i=0;i\neq j}^{k-1} |\zeta_i - \zeta_j|} .$$

A positive power of h is the leading term of a difference expression $\zeta_i - \zeta_j$ only when $\zeta_{i0} = \zeta_{j0}$. Assume that $\zeta_{00} = 1$ and ζ_{j0} has multiplicity m_j. Then

$$|C_j| \leqq \frac{C}{h^{(1-\gamma)(m_j-1)/m_j}} .$$

If $m_j = 1$, then by μ-stability there exist a constant $L > 0$ such that $|\zeta_j|^2 \leqq 1 - L h^{1-\gamma}$. In this case

$$|C_j| \sum_{i=0}^{n-1} (1 - L h^{1-\gamma})^{i/2} \leqq C/h^{(1-\gamma)}$$

for h sufficiently small. If $m_j > 1$, then by μ-stability and for h sufficiently small $|\zeta_j| \leqq r < 1$. Here

$$|C_j| \sum_{i=0}^{n-1} |\zeta_j|^i \leqq \frac{C}{h^{(1-\gamma)(m_j-1)/m_j}} \frac{(1-r^n)}{1-r} \leqq \frac{C}{h^{1-\gamma}}.$$

Combining these results gives Eq. (17). Q.E.D.

Using the same conditions as in Lemma 1, we have the immediate consequence

LEMMA 2.

$$|\Phi_n| \leqq \Phi\left(\frac{r^n}{h^{(1-\gamma)(m-1)/m}} + (1 - Lh^{1-\gamma})^n\right) \tag{18}$$

for h sufficiently small, where m equals the maximum multiplicity of the roots ζ_{j0} *and* $r = (1/2)(1 + \max_{|\zeta_{j0}|<1} |\zeta_{j0}|) < 1$.

L and Φ *are positive constants independent of h and* L_0, *i.e., L and* Φ *are uniform bounds for all* L_0 *such that* $-\bar{L} \leqq -L_0 \leqq -\underline{L} < 0$.

We will naturally assume that all of the conditions of the hypothesis of Theorem 1 are satisfied in proving the next result.

THEOREM 2. *Let the consistent finite-difference equation*

$$\sum_{i=0}^{k} \alpha_i Y_{n+i} = h^{1-\gamma} \sum_{i=0}^{k} \beta_i F_{n+i} + R_n\,, \tag{19}$$

where $F_{n+i} = f(Y_{n+i}, x_{n+i})$ *and* R_n *is the round-off error, satisfy the following conditions:*

(a) $R_n = O(h^{2(1-\gamma)})$;

(b) *The finite-difference equation is relatively stable*;

(c) $\partial f/\partial x$, $\partial f/\partial y$, *and* $\partial^2 f/\partial y^2$ *are continuous and bounded* $(-\bar{\bar{L}} \leqq \partial f/\partial y \leqq -\underline{L} < 0)$ *rof* $0 \leqq x \leqq 1$ *and* $-\infty < y < +\infty$;

(d) $|e_i| = |Y_i - y_i| \leqq Th^{1-\gamma}$ *for* $i = 0, 1, \cdots, k-1$ *where T is a positive constant independent of h. Then for h sufficiently small there exists a constant C such that* $|e_n| \leqq Ch^{1-\gamma}$ *for* $n = 0, 1, \cdots, N$ *where* $0 \leqq x_n \leqq x_N = 1$.

Proof. The exact solution of $\mu y' = f(y, x)$ satisfies the difference equation

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h^{1-\gamma} \sum_{i=0}^{k} \beta_i f_{n+i} + T_n \tag{20}$$

where T_n, the truncation error, is $O(h^{2(1-\gamma)})$ by the consistency condition and the hypothesis (c).

Subtracting Eq. (20) from (19) and letting $e_i = Y_i - y_i$, we obtain

$$\begin{aligned}\sum_{i=0}^{k} \alpha_i e_{n+i} &= h^{1-\gamma} \sum_{i=0}^{k} \beta_i (F_{n+i} - f_{n+i}) + R_n - T_n \\ &= h^{1-\gamma} \sum_{i=0}^{k} \beta_i \frac{\partial \bar{f}_{n+i}}{\partial y} e_{n+i} + R_n - T_n\end{aligned} \tag{21}$$

where

$$\frac{\partial \bar{f}_{n+i}}{\partial y} = \frac{\partial f}{\partial y}(Y_{n+i} + \delta_{n+i}(Y_{n+i} - y_{n+i}), x_{n+i})\,.$$

Denoting the right side of Eq. (21) by Q_n, we find that

$$|Q_n| \leqq h^{1-\gamma}\beta\bar{L} \sum_{i=0}^{k} |e_{n+i}| + |R_n - T_n| \tag{22}$$

where $\beta = \max |\beta_i|$.

The solution to Eq. (21) is given in Hull and Luxemburg [3] as

$$(23) \qquad e_n = \begin{cases} \sum_{i=0}^{n-k} g_{n-i-1} Q_i + \theta_n\,, & n \geqq k\,, \\ \theta_n\,, \quad n < k\,, \end{cases}$$

where g_n is defined as the solution to $\sum_{i=0}^{k} \alpha_i g_{n+i} = 0$ with initial conditions $g_0 = g_1 = \cdots = g_{k-2} = 0$, $g_{k-1} = \alpha_k^{-1}$; and where

$$\theta_n = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} \alpha_{k-j} g_{n+k-i-j-1} \right) e_i \quad \text{for} \quad n \geqq k\,.$$

Consequently, by relative stability we can set $\max |g_n| = g < \infty$. Then

$$(24) \qquad |\theta_n| \leqq k\alpha g \sum_{i=0}^{k-1} |e_i| \leqq k^2 \alpha g T h^{1-\gamma} \leqq K_1 h^{1-\gamma}$$

by condition (d). Here $\alpha = \max |\alpha_i|$ and $K_1 = k^2 \alpha g T$.

From Eqs. (22), (23), and (24) we obtain

$$(25) \quad |e_n| \leqq h^{1-\gamma} \beta g \bar{\bar{L}} |e_n| + h^{1-\gamma} \beta g \bar{\bar{L}} (k+1) \sum_{i=0}^{n-1} |e_i| + g \sum_{i=k}^{n} O(h^{2(1-\gamma)}) + |\theta_n|\,.$$

Now if $h \leqq h_0$ where $h_0^{1-\gamma} \beta g \bar{\bar{L}} < 1$ and $0 \leqq n \leqq N_0 = \ln h^{1-\gamma} / \ln r$, where $r = (1/2)(1 + \max_{|\zeta_{j0}|})$, as $|\zeta_{j0}| < 1$, then

$$(26) \qquad |e_n| \leqq h^{1-\gamma} A \sum_{i=0}^{n-1} |e_i| + K_3 h^{1-\gamma}$$

where

$$K_3 \geqq \frac{g K_2 N_0 h^{1-\gamma} + K_1}{1 - h_0^{1-\gamma} \beta g \bar{\bar{L}}} \quad \text{and} \quad A \geqq \frac{\beta g \bar{\bar{L}} (k+1)}{1 - h_0^{1-\gamma} \beta g \bar{\bar{L}}}\,.$$

K_2 is a bound for R_n and T_n. By a simple induction it follows that

$$(27) \qquad \begin{aligned} |e_n| &\leqq K_3 h^{1-\gamma} (1 + A h^{1-\gamma})^n \leqq K_3 h^{1-\gamma} e^{nAh^{1-\gamma}}\,, \\ |e_n| &\leqq K_3 h^{1-\gamma} \exp\,(A h^{1-\gamma} \ln h^{1-\gamma} / \ln r) \end{aligned}$$

for $0 \leqq n \leqq N_0$.

Although the last three inequalities leading to Eq. (27) assure us that $e_n = O(h^{1-\gamma})$ for the interval $[0, x_{N_0}]$, we cannot use this approach for the whole interval $[0, 1]$ because for $nh = 1$

$$\exp\,(nAh^{1-\gamma}) = \exp\,(A/h^{\gamma}) \to \infty \text{ as } h \to 0 \qquad (nh = 1)\,.$$

However, use has not yet been made of the fact that $\partial f / \partial y$ is continuous and $-\bar{\bar{L}} \leqq \partial f / \partial y \leqq -\bar{L} < 0$. To incorporate these suppositions into the proof requires a rather subtle argument. Basically, we translate the smallest value of $\partial f / \partial y$ to the lefthand side of Eq. (21) and then make use of Lemmas 1 and 2. This technique leads to the introduction of the maximum norm ($E_n = \max\,(|e_0|, |e_1|, \cdots, |e_n|)$) and consequently the continuity condition is imposed so that the coefficient multiplying E_n remains less than one in absolute value for some initial interval $[0, x_{N_1}]$

where N_1 can be chosen greater than N_0 under certain assumptions. Finally, the estimate (27) is used together with the one for E_n to obtain an upper bound for E_n on the interval $[0, x_{N_1}]$. The argument is repeated $l - 1$ times where $N_l = N$ and $x_N = 1$.

In order to consider $n > N_0$ we proceed as follows: Find an interval $[0, x_{N1}]$ where $-\bar{\bar{L}}_1 \leqq \partial f(y(x), x)/\partial y \leqq -\bar{L}_1 < 0$ and

$$h^{1-\gamma} \sum_{i=0}^{k-1}{}' (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^{k} |\beta_j| \frac{|C_i|}{1 - |\zeta_i|} < 1 \tag{28}$$

where the $'$ means the sum is taken over those i's corresponding to roots $|\zeta_{i0}| = 1$ and the C_i's are defined with respect to the difference equation

$$\sum_{i=0}^{k} (\alpha_i + h^{1-\gamma}\beta_i \bar{\bar{L}}_1)\Phi_{n+i} = 0 \tag{29}$$

with initial conditions $\Phi_0 = \Phi_1 = \Phi_2 = \cdots = \Phi_{k-2} = 0$ and $\Phi_{k-1} = (\alpha_k + h^{1-\gamma}\beta_k\bar{\bar{L}}_1)^{-1}$. The Green's function takes the form

$$\Phi_n = C_0\zeta_0^{\,n} + C_1\zeta_1^{\,n} + \cdots + C_{k-1}\zeta_{k-1}^{n}$$

and by relative stability

$$|\zeta_i| \leqq |\zeta_0| \,, \qquad i = 1, 2, \cdots, k - 1 \,,$$

where $\zeta_0 = 1 - \bar{\bar{L}}_1 h^{1-\gamma} + O(h^{2(1-\gamma)})$. Eq. (28) may now be rewritten as

$$\frac{\bar{\bar{L}}_1 - \bar{L}_1}{\bar{\bar{L}}_1 + O(h^{1-\gamma})} \sum_{i=0}^{k-1}{}' |C_i| \sum_{j=0}^{k} |\beta_j| < 1 \,.$$

We will assume that h is sufficiently small and

$$\sum_{i=0}^{k-1}{}' |C_i| \sum_{j=0}^{k} |\beta_j| \tag{30}$$

is close enough to 1 so that $N_0 < N_1$. Corollary 1 will show how this double sum can be minimized.

If we now add $h^{1-\gamma} \sum_{i=0}^{k} \beta_i \bar{\bar{L}}_i \, e_{n+i}$ to both sides of Eq. (21), we obtain

$$\sum_{i=0}^{k} (\alpha_i + h^{1-\gamma}\beta_i\bar{\bar{L}}_1)e_{n+i} = h^{1-\gamma} \sum_{i=0}^{k} \beta_i\left(\bar{\bar{L}}_1 + \frac{\partial \bar{f}_{n+i}}{\partial y}\right)e_{n+i} + R_n - T_n \,. \tag{31}$$

Let the right side of Eq. (31) be denoted by q_n and note that

$$\left|\bar{\bar{L}}_1 + \frac{\partial \bar{f}}{\partial y} n + i\right| \leqq \bar{\bar{L}}_1 - \bar{L}_1 + O(e_{n+i})$$

for $0 \leqq n \leqq N_1 - k$ and $i = 0, 1, \cdots, k$, where we have used condition (c).

Now it follows that

$$\begin{aligned} |q_i| &\leqq h^{1-\gamma}(\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^{k} |\beta_j| \, |e_{i+j}| + O(h^{2(1-\gamma)}) \\ &\quad + O\left(h^{1-\gamma} \sum_{j=0}^{k} |e_{i+j}|^2\right), \qquad 0 \leqq i \leqq N_1 - k \,, \end{aligned} \tag{32}$$

using condition (a) and our knowledge about T_n.

The solution to the difference equation, Eq. (31), is

$$e_n = \begin{cases} \sum_{i=0}^{n-k} \Phi_{n-i-1} q_i + \Psi_n , & n \geqq k , \\ \Psi_n , \quad n < k , \end{cases} \tag{33}$$

where Φ_n was defined by Eq. (29) and Ψ_n is given by

$$\Psi_n = \sum_{i=0}^{k-1} \left(\sum_{j=0}^{k-i-1} (\alpha_{k-j} + h^{1-\gamma} \beta_{k-j} \bar{\bar{L}}_1) \Phi_{n+k-i-j-1} \right) e_i \qquad \text{for } n \geqq k .$$

However, by Lemma 2 we can write for $n \geqq N_0$

$$\begin{aligned} |\Psi_n| &\leqq k^2 (\alpha + h^{1-\gamma} \beta \bar{\bar{L}}) \Phi \left[\frac{r^{N_0}}{h^{1-\gamma}} + (1 - L h^{1-\gamma})^{N_0} \right] T h^{1-\gamma} \\ &\leqq K_4 T h^{1-\gamma} \qquad \text{for } n \geqq N_0 . \end{aligned} \tag{34}$$

Defining $E_n = \max (|e_0|, |e_1|, \cdots, |e_n|)$ and using Lemma 1, we have

$$\sum_{i=0}^{n-k} |\Phi_{n-i-1}| O(h^{2(1-\gamma)}) \leqq K_5 h^{1-\gamma} \tag{35}$$

and

$$\sum_{i=0}^{n-k} |\Phi_{n-i-1}| O(h^{1-\gamma} E_n^{\,2}) \leqq K_6 E_n^{\,2} . \tag{36}$$

A bound can now be obtained for e_n using Eqs. (32) through (36):

$$\begin{aligned} |e_n| &\leqq h^{1-\gamma} \left((\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^{k} |\beta_j| E_n \sum_{i=0}^{n-k} |\Phi_{n-i-1}| \right) + K_4 T h^{1-\gamma} + K_5 h^{1-\gamma} + K_6 E_n^{\,2} \\ &\leqq \left[\sum_{i=0}^{k-1}{}' h^{1-\gamma} (\bar{\bar{L}}_1 - \bar{L}_1) \sum_{j=0}^{k} |\beta_j| \frac{|C_i|}{1 - |\zeta_i|} + O(h^{(1-\gamma)/m}) \right] E_n \\ &\quad + K_4 T h^{1-\gamma} + K_5 h^{1-\gamma} + K_6 E_n^{\,2} , \qquad N_0 \leqq n \leqq N_1 , \end{aligned}$$

where the $O(h^{(1-\gamma)/m})$ term results from the roots $|\zeta_{j0}| < 1$ and Lemma 2.

By our choice of N_1 (see Eq. (28)) and relative stability the term in brackets multiplying E_n will be less than $a < 1$.

Hence,

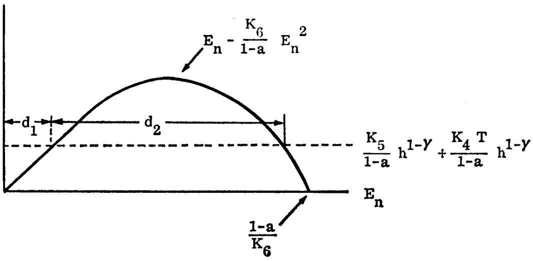
$$|e_n| \leqq a E_n + K_5 h^{1-\gamma} + K_4 T h^{1-\gamma} + K_6 E_n^{\,2} \qquad \text{for } N_0 \leqq n \leqq N_1 . \tag{37}$$

If $E_n = |e_j|$ for $j \leqq N_0$, then we may use our estimate, Eq. (27). If $j > N_0$, we can replace $|e_n|$ in Eq. (37) by E_n and obtain

$$E_n \leqq \frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4 T h^{1-\gamma}}{1-a} + \frac{K_6 E_n^{\,2}}{1-a} , \qquad 0 \leqq n \leqq N_1 , \tag{38}$$

since this bound is larger than the one for $|e_n|$ for $0 \leqq n \leqq N_0$, i.e., see Eq. (27).

Geometrically, we have the following picture


$E_n - \frac{K_6}{1-a} E_n^2$
d_1
d_2
$\frac{K_5}{1-a} h^{1-\gamma} + \frac{K_4 T}{1-a} h^{1-\gamma}$
E_n
$\frac{1-a}{K_6}$

We wish to show that $E_n \leqq d_1$ for $0 \leqq n \leqq N_1$. It will then follow that

$$d_1 = \frac{-1 + \left(1 - 4\frac{(K_4T + K_5)}{(1-a)^2} K_6 h^{1-\gamma}\right)^{1/2}}{\frac{-2K_6}{1-a}} = \frac{(K_4T + K_5)h^{1-\gamma}}{1-a} + O(h^{2(1-\gamma)}) .$$

The proof that $E_n \leqq d_1$ for $0 \leqq n \leqq N_1$ is by induction. Rewriting the difference equation, Eq. (21), with n replaced by $n - k$ and assuming $h \leqq h_0$ where

$$|\alpha_k| > \bar{L}|\beta_k| h_0^{1-\gamma} ,$$

we have

$$|e_n| \leqq \sum_{i=0}^{k-1} \frac{|-\alpha_i + h^{1-\gamma}\beta_i(\partial \bar{f}n - k + i/\partial y)|\,|e_{n-k+i}|}{|\alpha_k| - \bar{L}|\beta_k| h_0^{1-\gamma}} + \frac{|R_{n-k} - T_{n-k}|}{|\alpha_k| - \bar{L}|\beta_k| h_0^{1-\gamma}}$$
$$\leqq K_7 E_{n-1} + K_8 h^{2(1-\gamma)} .$$

$E_{k-1} \leqq Th^{1-\gamma} < d_1$ by hypothesis. By induction assume $E_{n-1} \leqq d_1$ and h is so small that

$$|e_n| \leqq K_7 d_1 + K_8 h^{2(1-\gamma)} < d_1 + d_2$$

and therefore $E_n < d_1 + d_2$. But now by Eq. (38) $E_n < d_1 + d_2$ implies $E_n \leqq d_1$. Thus by induction

$$(39) \qquad E_n \leqq \frac{K_5 h^{1-\gamma}}{1-a} + \frac{K_4 T h^{1-\gamma}}{1-a} + O(h^{2(1-\gamma)})$$

for $0 \leqq n \leqq N_1$.

We may now repeat the argument and extend the interval to $[0, x_{N_2}]$ where $N_1 < N_2$. We of course must use Eq. (39) for the initial conditions in the second interval.

Finally, at the end of l intervals a simple induction argument will show that

$$(40) \qquad E_n \leqq \frac{\left[1 - \left(\frac{K_4}{1-a}\right)^l\right]}{1 - \frac{K_4}{1-a}} \frac{K_5 h^{1-\gamma}}{1-a} + \left(\frac{K_4}{1-a}\right)^l T h^{1-\gamma} + O(h^{2(1-\gamma)})$$

for $0 \leqq n \leqq N_l = N$. Q.E.D.

Naturally, the bound for E_n given by Eq. (40) may be large if l is large, i.e., if

$$\sum_{i=0}^{k-1}{}' |C_i| \sum_{j=0}^{k} |\beta_j|$$

is much greater than 1. We therefore wish to minimize this sum. Since the roots $\zeta_{j0} \neq 1$ on the unit circle do not yield μ-stable difference schemes of higher precision than those roots inside the unit circle, we will exclude such roots for now.

A few important corollaries are:

COROLLARY 1. *If the only essential root (root on the unit circle) is $\zeta_{j0} = 1$, and if $\alpha_k = 1$ and $\beta_j \geqq 0$ for $j = 0, 1, \cdots, k$, then the value of l in Theorem 2 (inequality (40)) is one.*

Proof. Since $\beta_j \geqq 0$ by consistency

$$\rho'(1) = \sum_{j=0}^{k} |\beta_j| = (1 - r_1)(1 - r_2) \cdots (1 - r_{k-1}) ,$$

where $\rho(r_i) = 0$, $i = 1, 2, \cdots, k - 1$.

In Lemma 1 it was shown that

$$|C_0| \leqq \frac{1/(\alpha_k + O(h^{1-\gamma}))}{(\zeta_0 - \zeta_1)(\zeta_0 - \zeta_2) \cdots (\zeta_0 - \zeta_{k-1})} \approx \frac{1 + O(h^{1-\gamma})}{\rho'(1)} .$$

Therefore, $|C_0| \sum_{j=0}^{k} |\beta_j| \approx 1$ and l can be chosen equal to 1. Q.E.D.

COROLLARY 2. *If the only essential root is $\zeta_{j0} = 1$ and*

$$\frac{(\bar{L} - \underline{L})|C_0|}{\bar{\bar{L}} + O(h^{1-\gamma})} \sum_{j=0}^{k} |\beta_j| < 1$$

where $|\zeta_0| = 1 - \bar{\bar{L}}h^{1-\gamma} + O(h^{2(1-\gamma)})$, then the value of l in Theorem 2 is one.

Proof. The proof is obvious.

An example will illustrate how the value of l may be estimated in practice. Consider the differential equation $\mu y' = -y(y - 1)(20x + 10)$; $y(0) = 2$.

In the boundary layer for all $0 < h \leqq h_0$, $-30 \leqq f_y \leqq -10$.

Example 2. Suppose Adam's method

$$Y_{n+3} - Y_{n+2} = \frac{h^{1+\gamma}}{24} (9F_{n+3} + 19F_{n+2} - 5F_{n+1} + F_n)$$

is used to calculate the solution to the above differential equation. Here

$$|C_0| \sum_{i=0}^{3} |\beta_i| \approx \frac{34}{24} .$$

Since f_y is monotonic in the boundary layer

$$\bar{L}_1 = \bar{\bar{L}}_2 = L_1, \bar{L}_2 = \bar{\bar{L}}_3 = L_2, \quad \text{etc.} , \quad \frac{30 - L_1}{30} \frac{34}{24} = 0.8 ,$$

where we have let $a = 0.8 < 1$,

$$L_1 = 30(0.435) = 13.1 , \qquad L_2 = 30(0.435)^2 = 5.66 .$$

Outside of the boundary layer we make use of the following fact from the asymptotic theory:

$$f_y(y(x, \mu), x) = f_y(1, x) + O(h^\gamma) = -20x - 10 + O(h^\gamma)$$

for

$$-(2/\bar{L})h^\gamma 1nh^\gamma \leqq x \leqq 1 .$$

Here $f_y(1, x)$ is independent of h and is monotonic $(-30 \leqq f_y(1, x) \leqq -10)$. Thus we must increase l by 2. Hence $l = 4$. In practice it was observed that there was no error build up outside of the boundary-layer region for μ-stable schemes. Therefore, the estimate $l = 4$ is to be considered an absolute maximum for the value of l in this example.

Remark 1. The same proof of Theorem 2 could be used to obtain bounds for μ-stable methods instead of the less general relatively stable methods, but these μ-stable methods would require a much larger value of l.

Remark 2. Instead of considering $-\infty < y < \infty$ we could have considered a strip: $0 \leqq x \leqq 1$ and $|y - y(x)| < t$ where t is as large as is necessary in the proof.

Remark 3. If in Theorem 2 R_n and T_n are $O(h^{(p+1)(1-\gamma)})$ and $e_i = O(h^{p(1-\gamma)})$ for $i = 0, 1, \cdots, k - 1$, then the same proof will lead to the result that $e_n = O(h^{p(1-\gamma)})$ for $n = 0, 1, \cdots, N$.

5. Optimal Methods. By the "best method" or optimal method we will mean the μ-stable method which allows both l and T_n to be a minimum simultaneously.

By Corollary 1, l will have the value 1 if $\beta_i \geqq 0$, $i = 0, 1, \cdots, k$ and the only essential root is $\zeta = 1$. By using the methods outlined in Henrici [2] on optimal methods we find that the "best methods" for the roots $\zeta = 1$ and $\zeta = r$ where $|r| < 1$ take the form

$$\begin{aligned} Y_{n+2} &- (1 + r)Y_{n+1} + rY_n \\ &= \frac{h^{1-\gamma}}{12}[(5 + r)F_{n+2} + (8 - 8r)F_{n+1} + (-5r - 1)F_n] + T_n \end{aligned}$$

where

$$T_n = \frac{(1 + r)}{24} y^{(1v)} h^4 .$$

Now if $-1 < r \leqq -1/5$ all β_i's will be greater than or equal to zero. Therefore, we merely pick r close to -1 in order to make T_n small.

For the case $k = 3$ let the roots be $\zeta = 1$, r_1, and r_2 with $|r_1| < 1$ and $|r_2| < 1$. Of course if r_1 is complex then r_2 must be its complex conjugate. The optimal methods are characterized by:

$$\begin{aligned} \beta_3 &= (1/24)[9 + r_1 + r_2 + r_1r_2] , \\ \beta_2 &= (1/24)[19 - 13(r_1 + r_2) - 5r_1r_2] , \\ \beta_1 &= (1/24)[-5 - 13(r_1 + r_2) + 19r_1r_2] , \\ \beta_0 &= (1/24)[1 + r_1 + r_2 + 9r_1r_2] . \end{aligned}$$

The condition that guarantees $\beta_1 \geqq 0$ is the most restrictive; we must require $\operatorname{Re} r_1 \leqq 0$, $\operatorname{Re} r_2 \leqq 0$ and $-13(r_1 + r_2) \geqq 5$ or $19r_1r_2 \geqq 5$.

These inequalities will be satisfied if r_1 and r_2 lie in the shaded region of Fig. 1 and are complex conjugates if either is complex. Note also that

$$T_n = (1/720)(-19 - 11r_1 - 11r_2 - 19r_1r_2)$$

and has a minimum at $r_1 = r_2 = -11/19$, which lies in the shaded region of Fig. 1.

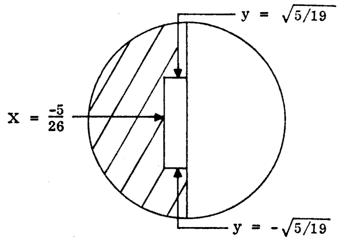


FIGURE 1

For the case $k = 4$ we choose the roots, 1, $r_1 = -s$, $r_2 = se^{i\theta}$ and $r_3 = se^{-i\theta}$, $0 < s < 1$ and $\pi/2 \leqq \theta \leqq \pi$, in order to simplify the arithmetic and to insure that

$$T_n = \frac{1}{1440}[-27 - 27r_1r_2r_3 - 11(r_1 + r_2 + r_3) - 11(r_1r_2 + r_1r_3 + r_2r_3)]y^{(v1)}(x)h^6$$

will be small. Note that the minimum occurs at $s = 1$.

The corresponding β's for the optimal methods are given by:

$$\beta_4 = \frac{1}{720}[251 + 19(r_1 + r_2 + r_3) + 11(r_1r_2 + r_1r_3 + r_2r_3) + 19r_1r_2r_3],$$

$$\beta_3 = \frac{1}{720}[646 - 346(r_1 + r_2 + r_3) - 74(r_1r_2 + r_1r_3 + r_2r_3) - 106r_1r_2r_3],$$

$$\beta_2 = \frac{1}{720}[-264 - 456(r_1 + r_2 + r_3) + 456(r_1r_2 + r_1r_3 + r_2r_3) + 264r_1r_2r_3],$$

$$\beta_1 = \frac{1}{720}[106 + 74(r_1 + r_2 + r_3) + 346(r_1r_2 + r_1r_3 + r_2r_3) - 646r_1r_2r_3],$$

$$\beta_0 = \frac{1}{720}[-19 - 11(r_1 + r_2 + r_3) - 19(r_1r_2 + r_1r_3 + r_2r_3) - 251r_1r_2r_3].$$

The analysis for the β_i's is straightforward, and the conclusion is that we must choose $0.695 \leqq s < 1$ and $\pi/2 \leqq \theta \leqq \pi$ to insure $\beta_i \geqq 0$, $i = 0, 1, 2, 3, 4$.

As k increases the analysis becomes much more difficult and even calculating general expressions for the β_i's and T_n in terms of $r_1, r_2, \cdots, r_{k-1}$ is very tedious. We therefore resort to a slightly different approach.

From our analysis of $k = 2$, 3, and 4 we suspect that for the r_i's in the negative half plane and near the unit circle there is some hope that for $k > 4$ all β_i's will be greater than zero. We make use of the following formulas derived by Hull and Newbery [5] for optimal methods:

$$T_n = \frac{Ry^{(k+2)}h^{k+2}}{(k+1)!} + O(h^{k+3}), \qquad R = \sum_{i=1}^{k} \bar{\alpha}_{i-1} \int_{i-1}^{i} x(x-1)\cdots(x-k)dx,$$

where

$$\bar{\alpha}_{i-1} = \alpha_i + \alpha_{i+1} + \cdots + \alpha_k$$

and

$$\beta_j = \sum_{i=1}^{k} \bar{\alpha}_{i-1} \int_{i-1}^{i} \frac{x(x-1)\cdots(x-k)}{(x-j)}\,dx$$

for $j = 0, 1, \cdots, k$.

The above integrals can be calculated exactly by using the Newton-Cotes formulas. We have programmed the CDC 6600 computer to calculate the values of β_j for the limiting values of r_i where we suspect favorable results for the β_j's; that is, for k even let one root be at $\zeta = 1$, another at $\zeta = -1$, and all the remaining ones at $\zeta = \pm i$.

For k odd let one root be at $\zeta = 1$ and all others at $\zeta = \pm i$. We refer to this choice of the roots at α-min. This is in contrast to α-max, where one root is chosen at $\zeta = 1$ and the remaining ones at $\zeta = -1$.

Although both α-min and α-max define unstable schemes, in practice we would choose one root at $\zeta = 1$ and the other roots inside the unit circle but near the roots of α-min or α-max when they lead to $\beta_i \geqq 0$, $i = 0, 1, \cdots, k$.

TABLE 1. *Computer Calculations*

Degree	α-max	α-min
$k = 3$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 4$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 5$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 6$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 7$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 8$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 9$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 10$	$\beta \geqq 0$	$\beta < 0$
$k = 11$	$\beta \geqq 0$	$\beta \geqq 0$
$k = 12$	$\beta \geqq 0$	$\beta < 0$
$k = 13$	$\beta \geqq 0$	$\beta < 0$
$k = 14$	$\beta \geqq 0$	$\beta < 0$
$k = 15$	$\beta \geqq 0$	$\beta < 0$
$k = 16$	$\beta \geqq 0$	$\beta < 0$
$k = 17$	$\beta \geqq 0$	$\beta < 0$
$k = 18$	$\beta \geqq 0$	$\beta < 0$

The results of the computer calculations are given in Table 1 for $k = 3, 4, \cdots, 18$. By $\beta \geqq 0$ we mean that $\beta_i \geqq 0$ for $i = 0, 1, \cdots, k$. $\beta < 0$ means that at least one β_i was less than zero.

6. Concluding Remarks. A series of nonlinear boundary-layer problems was solved on the IBM 7094 and the CDC 6600 by over 100 finite-difference schemes with various choices for the value of γ $(0 < \gamma < 1)$.

In every case when h became sufficiently small, schemes which were predicted to converge by the theory did so, while schemes which were predicted to diverge overflowed in the computer. The best accuracy (nine significant figures on the CDC

6600) was achieved by the optimal methods described in Section 5.

These data together with an exhaustive study of this subject including the extension to higher dimensions and the system

$$\mu y' = f(y, z, x) , \qquad z' = g(y, z, x) ,$$

can be found in Murphy [6].

7. Acknowledgment. I would like to express my deep sense of gratitude to Professor Eugene Isaacson of New York University for spending many hours discussing and reading the original report and for making a number of valuable suggestions and comments.

Autonetics
Anaheim, California

1. G. Dahlquist, "Convergence and stability in the numerical integration of ordinary differential equations," *Math. Scand.*, v. 4, 1956, pp. 33–53. MR **18,** 338.

2. Peter Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962. MR **24** #B1772.

3. T. E. Hull & W. A. J. Luxemburg, "Numerical methods and existence theorems for ordinary differential equations," *Numer. Math.*, v. 2, 1960, pp. 30–41. MR **22** #4847.

→ T. E. Hull & A. C. R. Newbery, "Error bounds for a family of three-point integration procedures," *J. Soc. Indust. Appl. Math.*, v. 7, 1959, pp. 402–412. MR **24** #B2118.

→ T. E. Hull & A. C. R. Newbery, "Integration procedures which minimize propagated errors," *J. Soc. Indust. Appl. Math.*, v. 9, 1961, pp. 31–47. MR **22** #11519.

6. W. D. Murphy, "Numerical analysis of boundary layer problems," AEC Research and Development Report NYO-1480-63, New York University.

7. A. B. Vasil'eva, "Asymptotic behavior of solutions to certain problems involving nonlinear differential equations containing a small parameter multiplying the highest derivatives," *Russian Math. Surveys*, v. 18, 1963, no. 3, pp. 13–84.