I. NUMERICAL SOLUTIONS OF STEADY VISCOUS FLOW PAST SPHERES AND GAS BUBBLES

II. NUMERICAL SOLUTION OF SINGULAR ENDPOINT BOUNDARY VALUE PROBLEMS

Thesis by

Donald Campbell Brabston, Jr.

In Partial Fulfillment of the Requirements

For the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

1974

(Submitted December 5, 1973)

ACKNOWLEDGMENTS

I wish to thank Professor Herbert B. Keller for suggesting these problems and for providing many valuable ideas. I wish to thank my fellow graduate students for many interesting discussions on these and other problems.

My sincere appreciation is extended to Elizabeth Fox for her excellent typing of this very hard-to-type work.

During the course of the research I was supported by the Woodrow Wilson National Fellowship Foundation, the National Science Foundation, Caltech, and the Veterans Administration.

Most of all I thank my wife, Martha, for her patience and understanding when the going was rough and her pressure when I procrastinated.

ABSTRACT

In Part I, numerical solutions of the Navier-Stokes equations are given for steady, viscous, incompressible, axisymmetric flow past a rigid sphere and a spherical gas bubble. The problem is formulated in terms of a stream function and the vorticity which are expanded in finite Legendre series. The coefficients in these series satisfy a finite system of ordinary differential equations. A finite-difference scheme is used to solve the system with Newton's method used to solve the nonlinear difference equations. The results agree very well with low and high Reynolds number theories.

In Part II, systems of ordinary differential equations with singular points of the first kind are considered. The singular point may be at either end, at both ends, or in the interior of a finite interval. A two-point linear system of boundary conditions is imposed at the endpoints. A theory is developed stating the conditions under which a unique solution will exist. A numerical method is developed for solving these problems. In this method, a series solution about the singular point is matched to a finite difference solution away from the singular point. Error estimates are developed, and numerical examples are given.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont'd)

PART I - LIST OF FIGURES

PART 1 - LIST OF TABLES

PART I

NUMERICAL SOLUTION OF STEADY VISCOUS FLOW PAST SPHERES AND GAS BUBBLES

1. INTRODUCTION

In this thesis, we study the numerical solution of axisymmetric steady, viscous, incompressible, laminar flow past a rigid sphere and past a spherical gas bubble for low to moderate Reynolds numbers.

One reason for studying this problem is to better understand numerical solution techniques for investigating fluid flow problems. Also we wish to check the theoretical results of D. W. Moore [15] for high Reynolds number flow past a gas bubble. Exact solutions of the Navier-Stokes equations for moderate or larger Reynolds numbers cannot be expected. Since experimental work is limited to low Reynolds number flow because of instability and turbulence, this problem is usually approached by numerical means. Because large amounts of computer time are usually needed for these problems, considerable emphasis is placed on developing an efficient computation scheme.

The method we use may be logically broken into two parts. First is the representation of the solution as infinite Legendre series in the polar angle θ and the derivation of the coupled system of ordinary differential equations for the coefficients in the series. In this part, we follow the work of Dennis and Walker [3] with the exception of the treatment of the boundary conditions.

The second part of the method is the numerical solution of the coupled nonlinear system of differential equations. For this, we use the centered Euler finite difference scheme of H. B. Keller [7] using

Newton's method for systems to solve the resulting nonlinear difference equations. Newton's method gives rise to large block tridiagonal systems which we solve with direct factorization. Several factorings were investigated and one proved superior to the rest. Finally, Richardson extrapolation was used to increase the order of accuracy. This increased accuracy was verified by examining the drag coefficient as calculated at various distances from the sphere.

Keller and Nieuwstadt [9] used much the same methods in investigating the two-dimensional flow past a cylinder with good results.

Very good agreement is obtained with the results of Dennis and Walker [3] for the rigid sphere. Good agreement is also obtained with both low Reynolds number theory and high Reynolds number theory for the gas bubble.

2. EQUATIONS OF MOTION

There are two basic numerical methods used for solving fluid flow problems past blunt bodies. The first involves integrating the time-dependent flow equations in time until little change is found in the flow field, and the flow is presumed steady. Work of this type for the rigid sphere has been done by Dennis and Walker [4].

In the second method, the equations of motion for steady flow are solved directly. This method is the one used here.

The Navier-Stokes equations for three-dimensional steady state flow are:

$$\nabla \cdot \underset{\sim}{q}^* = 0$$

$$\underset{\sim}{q}^* \cdot \nabla \underset{\sim}{q}^* = -\frac{1}{\rho} \nabla p^* + \nu \nabla^2 \underset{\sim}{q}^* \qquad (2.1)$$

where $\underset{\sim}{q}^*$ is the fluid velocity. These equations can be nondimensionalized with the velocity at infinity, $\underset{\sim}{q}^*_\infty = (U, 0)$, and a characteristic body length (in our case, the radius of the sphere, a). These dimensionless equations are:

$$\nabla \cdot \underset{\sim}{q} = 0$$

$$\underset{\sim}{q} \cdot \nabla \underset{\sim}{q} = -\text{grad}\, p + \frac{2}{R} \nabla^2 \underset{\sim}{q}$$

$$\underset{\sim}{q} = \underset{\sim}{q}^*/U \qquad (2.2)$$

$$p = p^*/\rho U^2$$

$$R = 2aU/\nu \, .$$

For spherical coordinates, Landau and Lifshitz [11] give these equations as:

$$\frac{\partial u}{\partial a} + \frac{2u}{a} + \frac{1}{a\sin\theta}\frac{\partial}{\partial\theta}(v\sin\theta) = 0$$

$$\frac{v}{a}\frac{\partial v}{\partial\theta} + u\frac{\partial v}{\partial a} + \frac{uv}{a} = -\frac{1}{a}\frac{\partial p}{\partial\theta} + \frac{2}{R}\left[\nabla^2 v - \frac{v}{a^2\sin^2\theta} + \frac{2}{a^2}\frac{\partial u}{\partial\theta}\right] \qquad (2.3)$$

$$\frac{v}{a}\frac{\partial u}{\partial\theta} + u\frac{\partial u}{\partial a} - \frac{v^2}{a} = -\frac{\partial p}{\partial a} + \frac{2}{R}\left[\nabla^2 u - \frac{2u}{a^2} - \frac{2\cot\theta}{a^2}v - \frac{2}{a^2}\frac{\partial v}{\partial\theta}\right]$$

where here u and v are the dimensionless radial and transverse components of the velocity, respectively, and account has been taken of the axisymmetric nature of the flow to eliminate dependence on the other angle, ϕ. Here, a is the dimensionless radial variable and θ is the polar angle with $\theta = 0$ pointing downstream into the wake. Also, because of symmetry, only the half plane needs to be treated (i.e., $0 \leq \theta \leq \pi$). (See fig. 1.)

In order to look at the far field more effectively and still retain significance near the sphere, the transformation $a = e^{\xi}$ is introduced and the equations become:

(a) $$\frac{\partial u}{\partial\xi} + 2u + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(v\sin\theta) = 0$$

(b) $$v\frac{\partial v}{\partial\theta} + u\frac{\partial v}{\partial\xi} + uv = -\frac{\partial p}{\partial\theta} + \frac{2e^{-\xi}}{R}\left[\frac{\partial^2 v}{\partial\xi^2} + \frac{\partial v}{\partial\xi} + \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\frac{\partial v}{\partial\theta}\right) - \frac{v}{\sin^2\theta} + 2\frac{\partial u}{\partial\theta}\right]$$

(c) $$v\frac{\partial u}{\partial\theta} + u\frac{\partial u}{\partial\xi} + v^2 = -\frac{\partial p}{\partial\xi} + \frac{2e^{-\xi}}{R}\left[\frac{\partial^2 u}{\partial\xi^2} + \frac{\partial u}{\partial\xi} - 2u - 2\cot\theta\, v - 2\frac{\partial v}{\partial\theta}\right]. \qquad (2.4)$$

We introduce a stream function ψ and a vorticity ζ which are defined by:

$$u = \frac{e^{-2\xi}}{\sin\theta}\frac{\partial\psi}{\partial\theta}, \quad v = -\frac{e^{-2\xi}}{\sin\theta}\frac{\partial\psi}{\partial\xi}$$
$$e^{\xi}\zeta = \frac{\partial v}{\partial\xi} + v - \frac{\partial u}{\partial\theta} . \tag{2.5}$$

Then the continuity equation, (2.4a), is automatically satisfied by the definition of ψ. Substituting (2.5) into (2.4b, c) and eliminating the pressure, we have:

$$\text{(a)} \quad \frac{\partial^2\psi}{\partial\xi^2} - \frac{\partial\psi}{\partial\xi} + \sin\theta\frac{\partial}{\partial\theta}\left(\frac{1}{\sin\theta}\frac{\partial\psi}{\partial\theta}\right) + \sin\theta e^{3\xi}\zeta = 0$$
$$\text{(b)} \quad \frac{\partial^2\zeta}{\partial\xi^2} + \frac{\partial\zeta}{\partial\xi} + \cot\theta\frac{\partial\zeta}{\partial\theta} + \frac{\partial^2\zeta}{\partial\theta^2} - \frac{\zeta}{\sin^2\theta} = \frac{1}{2}\mathrm{Re}^{\xi}\left(u\frac{\partial\zeta}{\partial\xi} + v\frac{\partial\zeta}{\partial\theta} - u\zeta - v\zeta\cot\theta\right) \tag{2.6}$$
$$\text{on } 0 \leqslant \xi < \infty, \quad 0 \leqslant \theta \leqslant \pi .$$

For both the rigid sphere and gas bubble, we have $u \equiv 0$ on the surface of the sphere since no fluid can enter or leave the sphere. This implies ψ = constant on the sphere; we take this constant to be zero. On the axis of symmetry, $\theta = 0$ and $\theta = \pi$, we have $\psi = \zeta = 0$ due to symmetry.

For an outer boundary condition as $\xi \to \infty$, we use the Oseen asymptotic expansion for the stream function and the vorticity instead of merely the free stream conditions. We impose this boundary condition at a finite radius from the sphere, call it ξ_∞, and require the stream function and vorticity at ξ_∞ to be equal to the value of the asymptotic expansion there. The Oseen expansion as given by Batchelor [2] is

$$\psi_\infty(\xi,\theta) = \frac{1}{2}e^{2\xi}\sin^2\theta + \frac{1}{4}e^{-\xi}\sin^2\theta - \frac{3}{R}(1+\cos\theta)(1-e^{-\frac{1}{4}Re^{\xi}(1-\cos\theta)})$$

$$\zeta_\infty(\xi,\theta) = -\frac{3}{2}e^{-2\xi}\sin\theta\, e^{-\frac{1}{4}Re^{\xi}(1-\cos\theta)}(1+\tfrac{1}{4}Re^{\xi}) . \qquad (2.7)$$

The only difference between the rigid sphere and the gas bubble is in the second boundary condition prescribed at the surface of the sphere. For the rigid sphere, we have the "no slip" condition which says that the fluid velocity on the sphere is zero; that is, $v = 0$ and hence $\frac{\partial\psi}{\partial\xi}\Big|_{\xi=0} = 0$. For the gas bubble, the no slip condition is replaced by the requirement that there be no stress on the bubble surface. This condition is (where μ is the fluid viscosity)

$$p_{r\theta} = \mu[e^{-\xi}\frac{\partial u}{\partial\theta} + \frac{\partial}{\partial\xi}(e^{-\xi}v)] = 0 \text{ on } \xi = 0 . \qquad (2.8)$$

To summarize, the boundary conditions for the rigid sphere are:

(a) $\psi(\xi,\theta) = 0$ on $\xi = 0$

(b) $\frac{\partial\psi(\xi,\theta)}{\partial\xi} = 0$ on $\xi = 0$

(c) $\psi(\xi,\theta) = \zeta(\xi,\theta) = 0$ on $\theta = 0,\ \theta = \pi$ (2.9)

(d) $\psi(\xi_\infty,\theta) = \psi_\infty(\xi_\infty,\theta)$

(e) $\zeta(\xi_\infty,\theta) = \zeta_\infty(\xi_\infty,\theta)$.

For the gas bubble, (2.9b) is replaced by

(b') $\mu[e^{-\xi}\frac{\partial u}{\partial\theta} + \frac{\partial}{\partial\xi}(e^{-\xi}v)] = 0$ on $\xi = 0$. (2.9)

Equations (2.6) along with boundary conditions (2.9) constitute an elliptic boundary value problem. We now wish to solve this problem numerically for various values of R.

3. METHODS OF SOLUTION

Different methods of solution of (2.6, 2.9) have been attempted in the past for flow past spheres and (in the two-dimensional case) cylinders. Several, such as Jenson [6] have used finite differences and relaxation methods on the stream function and vorticity equations. Others, for example Rimon and Cheng [16] and Dennis and Walker [4] have obtained steady solutions by integrating the time-dependent equations to a steady solution.

The method of series truncation applied to the stream function has been used for the cylinder by Underwood [19], by Keller and Nieuwstadt [9], and for the rigid sphere, by Dennis and Walker[4]. Many investigators compute, iteratively, a finite difference solution to the vorticity equation, and then use this solution to compute a solution to the stream function equation, etc. There are difficulties with this due to the overspecification and underspecification of boundary conditions for the equations. Our method here solves both equations simultaneously, eliminating the boundary condition specification problem.

Our method here breaks logically into two parts. The first part is much like the method of Dennis and Walker in representing the stream function and vorticity by Legendre series. Dennis and Walker only treat the rigid sphere case and not the gas bubble; also, we handle the boundary conditions in a much different manner. The second part of the method is to solve the resulting infinite coupled nonlinear system of ordinary differential equations. Here we differ completely from Dennis and Walker since we treat the entire

(truncated) system as a unit using H. B. Keller's centered Euler finite difference scheme.

3.1 Series Representation

We wish now to represent the stream function and vorticity as series each of whose terms is a product of an unknown function of the radial variable ξ and a known function of the angle θ. The known functions of the angle θ must vanish at $\theta = 0$ and $\theta = \pi$ to satisfy (2.9c). We choose Legendre functions for the angular functions and assume that the stream function and vorticity may be represented as:

$$\text{(a)} \qquad \psi(\xi,\theta) = e^{\frac{1}{2}\xi} \sum_{n=1}^{\infty} f_n(\xi) \int_z^1 P_n(t)dt$$

$$\text{(b)} \qquad \zeta(\xi,\theta) = \sum_{n=1}^{\infty} g_n(\xi)\, P_n^{(1)}(z) \tag{3.1.1}$$

where $z = \cos\theta$ and $P_n(z)$ and $P_n^{(1)}(z)$ are, respectively, the Legendre function and first associated Legendre function of order n.

We insert (3.1.1) in the stream function equation (2.6a) and use standard orthogonality relationships (given in Abramowitz and Stegun [1]) along with

$$\int_z^1 P_n(t)dt = \frac{1}{2n+1}\left[P_{n-1}(z) - P_{n+1}(z)\right]$$

$$P_n^{(1)}(z) = -(1-z^2)^{\frac{1}{2}} \frac{dP_n(z)}{dz} \tag{3.1.2}$$

to arrive at the equations:

$$f_n''(\xi) - (n+\tfrac{1}{2})^2 f_n(\xi) = e^{5/2\xi} n(n+1) g_n(\xi), \quad n = 1, 2, \ldots \tag{3.1.3}$$

In order to determine the equations for the g_n, we first use (3.1.1) in (2.5) to get

$$u = e^{-3/2\xi} \sum_{n=1}^{\infty} f_n(\xi)\, P_n(z)$$

$$v = -\frac{e^{-3/2\xi}}{\sin\theta} \sum_{n=1}^{\infty} \left[f_n'(\xi) + \tfrac{1}{2} f_n(\xi)\right] \int_z^1 P_n(t)dt \quad . \tag{3.1.4}$$

Calculation of the linear portion of the g_n equation from the left hand side of (2.6b) is straightforward. Calculation of the nonlinearity from the right hand side, however, requires use of the 3-J symbols. These are tabulated, along with formulas for their calculation, by Rotenberg, et al. [17]. Talman [18] gives many relations between the 3-J symbols and Legendre functions, among which we needed

$$\int_{-1}^{1} P_\ell^{(1)}(z) P_n(z) P_m^{(1)}(z) dz = -2\left[m(m+1)\ell(\ell+1)\right]^{\frac{1}{2}} \times \begin{pmatrix} m & n & \ell \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} m & n & \ell \\ 0 & 0 & 0 \end{pmatrix} \tag{3.1.5}$$

where $\begin{pmatrix} j_1 & j_2 & j_3 \\ m_1 & m_2 & m_3 \end{pmatrix}$ are the 3-J symbols. The resulting equation for g_n, $n = 1, 2, \ldots$, is

$$g_n''(\xi) + g_n'(\xi) - n(n+1) g_n(\xi) =$$

$$\tfrac{1}{2} R e^{-\frac{1}{2}\xi} \sum_{m=1}^{\infty} \sum_{\ell=1}^{\infty} \left\{ a^n_{m,\ell} f_m(\xi)\left[g_\ell'(\xi) - g_\ell(\xi)\right] + \beta^n_{m,\ell}\, g_\ell(\xi)\left[f_m'(\xi) + \tfrac{1}{2} f_m(\xi)\right] \right\} \tag{3.1.6}$$

where:
$$a^n_{m,\ell} = -(2n+1)\left[\frac{\ell(\ell+1)}{n(n+1)}\right]^{\frac{1}{2}} \begin{pmatrix} n & m & \ell \\ -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} n & m & \ell \\ 0 & 0 & 0 \end{pmatrix}$$

$$\beta^n_{m,\ell} = -(2n+1)\left[\frac{\ell(\ell^2-1)(\ell+2)}{nm(n+1)(m+1)}\right]^{\frac{1}{2}} \begin{pmatrix} n & m & \ell \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} n & m & \ell \\ 0 & 0 & 0 \end{pmatrix}$$

For boundary conditions, we simply substitute the series (3.1.1) into the boundary conditions (2.9). Condition (2.9c) is satisfied automatically by the series because of (3.1.2). The zero streamline condition (2.9a) is satisfied if and only if $f_n(0) = 0$, $n = 1, 2, \ldots$. For the rigid sphere, the no slip condition (2.9b) implies $f_n'(0) = 0$, $n = 1, 2, \ldots$. It can be shown that the gas bubble zero stress condition (2.9b') is satisfied if and only if

$$f_n'(0) - \tfrac{1}{2}n(n+1)g_n(0) = 0, \quad n = 1, 2, \ldots \ .$$

Using the definitions of ψ_∞ and ζ_∞, (2.7), and a bit of calculation, it can be shown that the $f_n(\xi_\infty)$ and $g_n(\xi_\infty)$ must satisfy

$$f_1(\xi) = f_1^\infty(\xi) \equiv \frac{3}{2} e^{-\frac{1}{2}\xi}\left\{\frac{2}{3}\left(\tfrac{1}{2}e^{-\xi}+e^{2\xi}\right)-\frac{3}{R}\left(2-\frac{2}{p}+\frac{1}{p^2}-\frac{1}{p^2}e^{-2p}\right)\right\}$$

$$f_2(\xi) = f_2^\infty(\xi) \equiv \frac{45}{2R} e^{-\frac{1}{2}\xi}\left\{-\frac{2}{3}-\frac{3}{p^2}+\frac{2}{p}+\frac{2}{p^3}-\frac{1}{p^2}e^{-2p}-\frac{2}{p^3}e^{-2p}\right\}$$

$$f_{n+1}(\xi) = f_{n+1}^\infty(\xi) \equiv \frac{2n+3}{2n-1} f_{n-1}(\xi)+(2n+3)(n+\tfrac{1}{2})e^{-\frac{1}{2}\xi}$$

$$\times\left\{-\frac{4}{15}\delta_{n,2}\left(\tfrac{1}{4}e^{-\xi}+\tfrac{1}{2}e^{2\xi}\right)\right.$$

$$\left.+\frac{3}{R}\frac{e^{-p}}{2^n}\sum_{j=0}^{m}\frac{(-1)^{n-j}(2n-2j)!}{j!\,p^{n-2j+1}(n-j)!}\left[S_{n-2j}(p)-\frac{n-2j+1}{p}S_{n-2j+1}(p)\right]\right\}$$

$$g_n(\xi) = g_n^\infty(\xi) \equiv \frac{3}{2}\frac{n+\frac{1}{2}}{n(n+1)} e^{-2\xi}(1+p)e^{-p}\frac{1}{2^n} \times$$

$$\sum_{j=0}^{m}\frac{(-1)^{n-j+1}(2n-2j)!}{j!\,(n-j)!\,p^{n-2j}}\left[S_{n-2j-1}(p)-\frac{(n-2j+1)(n-2j)}{p^2}S_{n-2j+1}(p)\right]$$

$$m = \begin{cases}\frac{1}{2}(n-1), & n \text{ odd}\\ \frac{1}{2}n-1, & n \text{ even}\end{cases} \qquad (3.1.7)$$

$$p = \tfrac{1}{4} \mathrm{Re}^{\xi}$$

$$S_k(p) = e^{p} \sum_{\ell=0}^{k} \frac{(-p)^{\ell}}{\ell!} - e^{-p} \sum_{\ell=0}^{k} \frac{p^{\ell}}{\ell!}, \quad k = 0, 1, \ldots$$

$$\delta_{i,k} = \begin{cases} 0, & i \neq k \\ 1, & i = k \end{cases} \qquad \text{(3.1.7) cont'd}$$

$n = 1, 2, \ldots$, at $\xi = \xi_{\infty}$.

To summarize, the boundary conditions for the f_n and g_n are, for the rigid sphere case:

(a) $f_n(0) = 0$, $n = 1, 2, \ldots$

(b) $f_n'(0) = 0$, $n = 1, 2, \ldots$

(c) $f_n(\xi_{\infty}) = f_n^{\infty}(\xi_{\infty})$, $n = 1, 2, \ldots$ (3.1.8)

(d) $g_n(\xi_{\infty}) = g_n^{\infty}(\xi_{\infty})$, $n = 1, 2, \ldots$.

For the gas bubble case, (3.1.8b) is replaced by

(b') $f_n'(0) - \tfrac{1}{2} n(n+1) g_n(0) = 0$, $n = 1, 2, \ldots$. (3.1.8)

The equations (3.1.3), (3.1.6) along with boundary conditions (3.1.8) form a coupled nonlinear infinite system of ordinary differential equations to be solved for the $f_n(\xi)$ and $g_n(\xi)$. Once this system is solved, the flow field is known since the stream function and vorticity can be reconstructed from the (assumed convergent) series expansions (3.1.1).

From the solution of this system, we can also calculate the velocity from (3.1.4) and various physical parameters such as the drag coefficient and pressure coefficient. The drag coefficient is defined by

$$C_D = \frac{D}{\pi \rho U^2 a^2} \tag{3.1.9}$$

where ρ is the (constant) fluid density and D is the drag on the sphere. Milne-Thompson [14] gives the force on the sphere as

$$\underset{\sim}{D} = -\int_S p^* \underset{\sim}{n} dS - \mu \int_S \underset{\sim}{n} \times \underset{\sim}{\zeta}^* dS - \rho \int_S \underset{\sim}{q}^* (\underset{\sim}{n} \cdot \underset{\sim}{q}^*) dS \tag{3.1.10}$$

where p^*, $\underset{\sim}{\zeta}^*$, $\underset{\sim}{q}^*$ are the dimensional pressure, vorticity, and velocity; the integration is performed over any sphere enclosing the rigid sphere or gas bubble, and $\underset{\sim}{n}$ is the outward unit normal vector. Evaluating this at the sphere ξ = constant, we calculate the drag coefficient (in terms of dimensionless quantities) as

$$C_D = -e^{2\xi}\left[\int_0^\pi p(\xi, \theta)\sin 2\theta d\theta + \frac{4}{R}\int_0^\pi \zeta(\xi,\theta)\sin^2\theta d\theta \right.$$
$$\left. -2\int_0^\pi uv \sin^2\theta d\theta + \int_0^\pi u^2 \sin 2\theta d\theta\right] . \tag{3.1.11}$$

In order to compute C_D in terms of the $f_n(\xi)$ and $g_n(\xi)$, we use the expansions (3.1.1), (3.1.4), Legendre function identities, the Navier-Stokes equations, and the following addition formula [18]:

$$P_\ell^{(m)}(\cos\theta) P_{\ell'}^{(m')}(\cos\theta) = (-1)^M \left[\frac{(\ell+m)!\,(\ell'+m')!}{(\ell-m)!\,(\ell'-m')!}\right]^{\frac{1}{2}} \times$$
$$\sum_L (2L+1)\begin{pmatrix} \ell & \ell' & L \\ m & m' & -M \end{pmatrix}\begin{pmatrix} \ell & \ell' & L \\ 0 & 0 & 0 \end{pmatrix}\left[\frac{(L-M)!}{(L+M)!}\right]^{\frac{1}{2}} P_L^{(M)}(\cos\theta)$$

(3.1.12)

where $M = m+m'$ and the sum is over all L for which the 3-J symbols are nonzero. This lengthy calculation gives

$$C_D(\xi) = -e^{2\xi}\cdot\Bigg[\frac{8}{3R}\left(g_1(\xi)+g_1'(\xi)\right)+4e^{-3\xi}\sum_{n=2}^{\infty} f_{n-1}(\xi)f_n(\xi)\frac{n}{(2n-1)(2n+1)}$$

$$+4e^{-3\xi}\sum_{n=1}^{\infty}\left[f_n'(\xi)+\tfrac{1}{2}f_n(\xi)\right]\left[f_{n+1}'(\xi)+\tfrac{1}{2}f_{n+1}(\xi)\right]\frac{1}{(n+1)(n(n+2))^{\frac{1}{2}}}\begin{pmatrix}1 & n+1 & n\\ 0 & -1 & 1\end{pmatrix}\begin{pmatrix}1 & n+1 & n\\ 0 & 0 & 0\end{pmatrix}$$

$$+2\sqrt{2}\,e^{-\frac{1}{2}\xi}\sum_{n=1}^{\infty} g_n(\xi)\sqrt{n(n+1)}\Bigg[f_{n+1}(\xi)\begin{pmatrix}n+1 & n & 1\\ 0 & 1 & -1\end{pmatrix}\begin{pmatrix}n+1 & n & 1\\ 0 & 0 & 0\end{pmatrix}$$

$$+f_{n-1}(\xi)\begin{pmatrix}n-1 & n & 1\\ 0 & 1 & -1\end{pmatrix}\begin{pmatrix}n-1 & n & 1\\ 0 & 0 & 0\end{pmatrix}\Bigg]$$

$$-\frac{16}{3R}g_1(\xi) \qquad (3.1.13)$$

$$+2e^{-3\xi}\sum_{n=1}^{\infty} f_n(\xi)\frac{1}{n+\frac{1}{2}}\left[f_{n+1}'(\xi)\frac{1}{2n+3}-f_{n-1}'(\xi)\frac{1}{2n-1}\right]\Bigg].$$

On the rigid sphere surface making use of the boundary conditions (3.1.8), this reduces to

$$C_D(0) = \frac{8}{3R}\left[g_1(0) - g_1'(0)\right]. \qquad (3.1.14)$$

For the gas bubble, we have

$$C_D(0) = \frac{8}{3R}\left[g_1(0) - g_1'(0)\right]$$

$$-\sum_{n=1}^{\infty} g_n(0)g_{n+1}(0)(n+1)\sqrt{n(n+2)}\begin{pmatrix}1 & n+1 & n\\ 0 & -1 & 1\end{pmatrix}\begin{pmatrix}1 & n+1 & n\\ 0 & 0 & 0\end{pmatrix}. \qquad (3.1.14')$$

We define a pressure coefficient, $k(\theta)$, as follows:

$$k(\theta) = \frac{p^*(0,\theta)-p^*_\infty}{\frac{1}{2}\rho U^2}$$

$$= 2\int_\pi^\theta \left.\frac{\partial p'(\xi,\theta)}{\partial\theta}\right|_{\xi=0} d\theta + k(\pi) \qquad (3.1.15)$$

where $k(\pi) = 2\int_\infty^0 \left.\frac{\partial p'(\xi,\theta)}{\partial\xi}\right|_{\theta=\pi} d\xi$.

Then, we have for both the gas bubble and the rigid sphere,

$$k(\pi) = 1 - \frac{4}{R}\sum_{n=1}^{\infty} n(n+1)\int_0^{\xi_\infty} g_n(\xi)d\xi + O(Re^{-\frac{1}{2}Re^{\xi_\infty}-\xi_\infty}) \ . \qquad (3.1.16)$$

For the rigid sphere

$$k(\theta) = k(\pi) + \frac{4}{R}\sum_{n=1}^{\infty}\left[g_n'(0) + g_n(0)\right]\left[P_n(\cos\theta)-(-1)^n\right] \ , \qquad (3.1.17)$$

and for the gas sphere,

$$k(\theta) = k(\pi) + \frac{4}{R}\sum_{n=1}^{\infty}\left[g_n'(0) + g_n(0)\right]\left[P_n(\cos\theta)-(-1)^n\right] - \frac{1}{8}\zeta^2(0,\theta) \ . \qquad (3.1.17')$$

The problem is now reduced to solving the system (3.1.3, 3.1.6) and boundary conditions (3.1.8). The first step is to truncate the infinite series to reduce the system to a finite system. But solving even this reduced system is far from simple. Dennis and Walker use a finite difference method to solve the g_n equation and a shooting method to solve the f_n equation and alternate between these two to couple the system. They are also forced to use a relaxation parameter which decreases as the Reynolds number increases, thereby entailing more and more iterations as $R \to \infty$.

Our method, as described in the next section, seems much more efficient, as well as appearing capable of going to much higher Reynolds numbers.

3.2 Numerical Solution

The infinite system given by (3.1.3), (3.1.6), (3.1.8) is made finite by truncating the stream function and vorticity series (3.1.1) at N. This is equivalent to taking

$$f_n(\xi) = g_n(\xi) \equiv 0, \quad n > N . \tag{3.2.1}$$

The problem is then reduced to the finite system:

$$\text{(a)} \qquad f_n'' - (n+\tfrac{1}{2})^2 f_n - e^{5/2\xi} n(n+1) g_n = 0$$

$$\text{(b)} \qquad g_n'' + g_n' - n(n+1) g_n = \tfrac{1}{2} \mathrm{Re}\, e^{-\frac{1}{2}\xi} \sum_{m=1}^{N} \sum_{i=1}^{N} \Big\{ a^n_{m,i} f_m [g_i' - g_i] + \beta^n_{m,i} g_i [f_m' + \tfrac{1}{2} f_m] \Big\}, \xi \in [0, \xi_\infty]$$

$$\text{(c)} \qquad f_n(0) = 0 \tag{3.2.2}$$

$$\text{(d)} \qquad f_n'(0) = 0 \qquad \text{(rigid sphere)}$$

$$\text{(d')} \qquad f_n'(0) - \tfrac{1}{2} n(n+1) g_n(0) = 0 \qquad \text{(gas bubble)}$$

$$\text{(e)} \qquad f_n(\xi_\infty) = f_n^\infty(\xi_\infty)$$

$$\text{(f)} \qquad g_n(\xi_\infty) = g_n^\infty(\xi_\infty) \qquad n = 1, 2, \ldots, N .$$

We now rewrite the problem (3.2.2) as a first order system by defining

$$\mathcal{Z}(\xi) = \begin{bmatrix} u_1 \\ \vdots \\ u_N \\ v_1 \\ \vdots \\ v_N \\ f_1 \\ \vdots \\ f_N \\ g_1 \\ \vdots \\ g_N \end{bmatrix} = \begin{bmatrix} f'_1 \\ \vdots \\ f'_N \\ g'_1 \\ \vdots \\ g'_N \\ f_1 \\ \vdots \\ f_N \\ g_1 \\ \vdots \\ g_N \end{bmatrix} , \quad \xi \in [0, \xi_\infty] \qquad (3.2.3)$$

Then the differential equations (3.2.2a, b) become

$$\mathcal{Z}'(\xi) = \begin{bmatrix} \vdots \\ (n+\tfrac{1}{2})^2 f_n + e^{5/2\xi} g_n \\ \vdots \\ -v_n + n(n+1) g_n + \tfrac{1}{2} R e^{-\frac{1}{2}\xi} \sum_{m,i=1} \gamma^n_{m,i}(u_m, v_i, f_m, g_i) \\ \vdots \\ u_1 \\ \vdots \\ u_N \\ v_1 \\ \vdots \\ v_N \end{bmatrix} \begin{matrix} \} N \text{ rows} \\ \\ \} N \text{ rows} \end{matrix} \qquad (3.2.4)$$

$$\gamma^n_{m,i} = \alpha^n_{m,i} f_m [v_i - g_i] + \beta^n_{m,i} g_i [u_m + \tfrac{1}{2} f_m] \quad , \quad \xi \in [0, \xi_\infty] \ .$$

Breaking this system into its linear and nonlinear parts, we can rewrite it as

$$\underset{\sim}{\mathcal{Z}}'(\xi) = A(\xi)\,\underset{\sim}{\mathcal{Z}}(\xi) + \underset{\sim}{\mathcal{N}}(\xi, \underset{\sim}{\mathcal{Z}}),$$

$$A(\xi) \equiv \begin{bmatrix} 0 & 0 & (3/2)^2 \cdot & e^{5/2\xi_2} \cdot & \\ & & \cdot\,(N+\tfrac{1}{2})^2 & \cdot\, e^{5/2\xi_{N(N+1)}} & \\ 0 & -I & 0 & 2 \cdot & \\ & & & \cdot\, N(N+1) & \\ I & 0 & 0 & 0 & \\ 0 & I & 0 & 0 & \end{bmatrix} \tag{3.2.5}$$

$$\underset{\sim}{\mathcal{N}}(\xi, \underset{\sim}{\mathcal{Z}}) \equiv \tfrac{1}{2}\,\mathrm{Re}^{-\frac{1}{2}\xi} \sum_{m=1}^{N} \sum_{i=1}^{N} \begin{bmatrix} 0 \\ \gamma^n_{m,i} \\ 0 \\ 0 \end{bmatrix} \begin{matrix} \text{N rows} \\ \text{N rows} \\ \text{N rows} \\ \text{N rows} \end{matrix} .$$

Here the I's are N x N identity matrices.

The boundary conditions are linear and of the separated end-point type. Hence, for the rigid sphere case, they may be represented by:

$$\text{(a)} \qquad B_o\,\underset{\sim}{\mathcal{Z}}(0) = \underset{\sim}{0}, \quad B_1\,\underset{\sim}{\mathcal{Z}}(\xi_\infty) = \underset{\sim}{\Gamma}$$

$$\text{(b)} \qquad B_o = \begin{bmatrix} 0 & 0 & I & 0 \\ I & 0 & 0 & 0 \end{bmatrix} \begin{matrix} \text{N rows} \\ \text{N rows} \end{matrix} \tag{3.2.6}$$

$$(c) \qquad B_1 = \begin{bmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$

$$(d) \qquad \underset{\sim}{\Gamma} = \begin{bmatrix} f_1^{\infty}(\xi_{\infty}) \\ \vdots \\ f_N^{\infty}(\xi_{\infty}) \\ g_1^{\infty}(\xi_{\infty}) \\ \vdots \\ g_N^{\infty}(\xi_{\infty}) \end{bmatrix} \qquad (3.2.6) \text{ cont'd}$$

For the gas bubble, B_o is replaced by

$$(b') \qquad B_o = \begin{bmatrix} 0 & 0 & I & 0 & & \\ & & & -1 & & \\ & & & & -3 & \\ & & & & & \ddots \\ I & 0 & 0 & & & -\frac{1}{2}N(N+1) \end{bmatrix} . \qquad (3.2.6)$$

Equation (3.2.5) and boundary conditions (3.2.6) form a 4N component first order nonlinear system of ordinary differential equations with two-point linear separated end conditions.

To solve this system numerically, we use the centered Euler finite difference scheme proposed and analyzed by H. B. Keller [7]. Keller shows that this scheme has $O(h^2)$ truncation error where h is the mesh spacing. He also shows that under certain conditions the error in the computed solution has a certain asymptotic form and that Richardson extrapolation may be used to increase the order of accuracy. Keller outlines the use of Newton's method for solving the

nonlinear equations and gives sufficient conditions for it to converge quadratically. Finally, he discusses LU decompositions for the solution of the Newton iterates. We will follow basically this same plan of attack in solving the system (3.2.5), (3.2.6).

We first define a uniform mesh on the interval $[0, \xi_\infty]$ as follows

$$\xi_o = 0, \quad \xi_j = jh, \quad j = 1, 2, \ldots, J, \; h = \xi_\infty / J.$$

Then the centered Euler scheme for (3.2.5), (3.2.6) is the difference equations

$$\frac{\underset{\sim}{Z}_j - \underset{\sim}{Z}_{j-1}}{h} = A(\xi_{j-\frac{1}{2}}) \; \frac{\underset{\sim}{Z}_j + \underset{\sim}{Z}_{j-1}}{2} + \underset{\sim}{N}(\xi_{j-\frac{1}{2}}, \; \frac{\underset{\sim}{Z}_j + \underset{\sim}{Z}_{j-1}}{2})$$

$$\xi_{j-\frac{1}{2}} = \xi_j - h/2 \qquad j = 1, 2, \ldots, J \; . \tag{3.2.7}$$

$$B_o \underset{\sim}{Z}_o = \underset{\sim}{0} \; , \quad B_1 \underset{\sim}{Z}_J = \underset{\sim}{\Gamma} \; ,$$

where $\underset{\sim}{Z}_j = \begin{bmatrix} u_{1,j} \\ \vdots \\ u_{N,j} \\ v_{1,j} \\ \vdots \\ g_{N,j} \end{bmatrix}$ is to be the computed approximation to $\underset{\sim}{Z}(\xi_j)$.

We now define a "maxi-vector" of 4N(J+1) components by

$$\underset{\sim}{F} \equiv \begin{bmatrix} \underset{\sim}{z}_0 \\ \vdots \\ \underset{\sim}{z}_J \end{bmatrix}$$ and write (3.2.7) as a maxi-vector system:

$$\underset{\sim}{\Phi}(\underset{\sim}{F}) \equiv \begin{bmatrix} B_0 \underset{\sim}{z}_0 \\ \vdots \\ \underset{\sim}{z}_j - \underset{\sim}{z}_{j-1} - \frac{h}{2} A(\xi_{j-\frac{1}{2}})(\underset{\sim}{z}_j + \underset{\sim}{z}_{j-1}) - h\underset{\sim}{N}(\xi_{j-\frac{1}{2}}, \frac{\underset{\sim}{z}_j + \underset{\sim}{z}_{j-1}}{2}) \\ \vdots \\ B_1 \underset{\sim}{z}_J - \underset{\sim}{\Gamma} \end{bmatrix} = \underset{\sim}{0} . \qquad (3.2.8)$$

(3.2.8) is a nonlinear algebraic system for $\underset{\sim}{F}$. We wish to solve this problem using Newton's method. To formulate Newton's method, assume $\underset{\sim}{F}^{(\nu)}$ is an approximation to $\underset{\sim}{F}^*$, the desired root of (3.2.8), and define

$$\underset{\sim}{F}^{(\nu+1)} = \underset{\sim}{F}^{(\nu)} + \delta\underset{\sim}{F}^{(\nu)} . \qquad (3.2.9)$$

Then we want

$$\underset{\sim}{\Phi}(\underset{\sim}{F}^{(\nu+1)}) = \underset{\sim}{0} ;$$

hence, $\underset{\sim}{\Phi}(\underset{\sim}{F}^{(\nu)} + \delta\underset{\sim}{F}^{(\nu)}) = \underset{\sim}{\Phi}(\underset{\sim}{F}^{(\nu)}) + \frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}}(\underset{\sim}{F}^{(\nu)})\, \delta\underset{\sim}{F}^{(\nu)} + \ldots = 0$,

where $\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}}$ is the Jacobian matrix of $\underset{\sim}{\Phi}$ with respect to $\underset{\sim}{F}$. Linearizing, we can determine $\delta\underset{\sim}{F}^{(\nu)}$ from the equation

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}}(\underset{\sim}{F}^{(\nu)})\, \delta\underset{\sim}{F}^{(\nu)} = - \underset{\sim}{\Phi}(\underset{\sim}{F}^{(\nu)}) . \qquad (3.2.10)$$

Thus, with an appropriate selection of an initial guess $\underset{\sim}{F}^{(0)}$, we can determine successively closer iterates $\underset{\sim}{F}^{(\nu)}$ to $\underset{\sim}{F}^{*}$. We stop the process when we are as close as desired as determined by a preset error criterion.

For our problem (3.2.8), then, all we need do is calculate $\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}}(\underset{\sim}{F})$ and specify the method of solving the linear system (3.2.10). We can calculate the Jacobian of $\underset{\sim}{\Phi}$ by blocks; that is,

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}}(\underset{\sim}{F}) = \frac{\partial \underset{\sim}{\Phi}_i}{\partial \underset{\sim}{F}_j} \qquad (i^{\underline{\text{th}}} \text{ block row, } j^{\underline{\text{th}}} \text{ block column}).$$

Using (3.2.5), this calculation yields

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}} = \begin{bmatrix} B_0 & & & & \\ -L_1 & R_1 & & 0 & \\ & -L_2 & R_2 & \ddots & \\ & & \ddots & \ddots & \\ 0 & & & -L_J & R_J \\ & & & & B_1 \end{bmatrix} \tag{3.2.11}$$

$$L_j = I + \frac{h}{2} A(\xi_{j-\frac{1}{2}}) + \frac{h}{2} \frac{\partial \underset{\sim}{N}}{\partial \underset{\sim}{Z}}\left(\xi_{j-\frac{1}{2}}, \frac{\underset{\sim}{Z}+\underset{\sim}{Z}_{j-1}}{2}\right), \text{ (4Nx4N matrices)}$$

$$R_j = I - \frac{h}{2} A(\xi_{j-\frac{1}{2}}) - \frac{h}{2} \frac{\partial \underset{\sim}{N}}{\partial \underset{\sim}{Z}}\left(\xi_{j-\frac{1}{2}}, \frac{\underset{\sim}{Z}_j+\underset{\sim}{Z}_{j-1}}{2}\right), \text{ (4Nx4N matrices)}$$

$$\frac{\partial \underset{\sim}{N}}{\partial \underset{\sim}{Z}}(\xi, \underset{\sim}{Z}) = \frac{1}{2} Re^{-\frac{1}{2}\xi} \sum_{m=1}^{N} \sum_{i=1}^{N} \begin{bmatrix} 0 & 0 & 0 & 0 \\ \left[\frac{\partial \gamma^j_{m,i}}{\partial u_k}\right] & \left[\frac{\partial \gamma^j_{m,i}}{\partial v_k}\right] & \left[\frac{\partial \gamma^j_{m,i}}{\partial f_k}\right] & \left[\frac{\partial \gamma^j_{m,i}}{\partial g_k}\right] \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

where

$$\frac{\partial \gamma^j_{m,i}}{\partial u_k} = \beta^j_{m,i}\, g_i\, \delta_{m,k} \qquad (j^{\underline{th}} \text{ row},\ k^{\underline{th}} \text{ column})$$

$$\frac{\partial \gamma^j_{m,i}}{\partial v_k} = \alpha^j_{m,i}\, f_m\, \delta_{i,k}$$

$$\frac{\partial \gamma^j_{m,i}}{\partial f_k} = \left\{ \alpha^j_{m,i} [v_i - g_i] + \tfrac{1}{2} \beta^j_{m,i}\, g_i \right\} \delta_{m,k} \qquad (3.2.12)$$

$$\frac{\partial \gamma^j_{m,i}}{\partial g_k} = \left\{ -\alpha^j_{m,i}\, f_m + \beta^j_{m,i} [u_m + \tfrac{1}{2} f_m] \right\} \delta_{i,k}$$

$$\delta_{i,k} = \begin{cases} 0 & , \quad i \neq k \\ 1 & , \quad i = k \end{cases} .$$

Equations (3.2.10) and (3.2.11) define a linear system to be solved for each successive Newton iterate. In order to solve this system, we note that (3.2.11) can be looked upon as a block tridiagonal matrix of 4N x 4N blocks grouped as follows:

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{\mathcal{Z}}} = \begin{bmatrix} B_0 & R_1 & & & 0 \\ L_1 & & R_2 & & \\ & L_2 & & \ddots & \\ & & \ddots & & R_J \\ 0 & & & L_J & B_1 \end{bmatrix} \qquad (3.2.13)$$

In more standard tridiagonal notation,

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}} = \begin{bmatrix} A_1 & C_1 & & & & & \\ B_2 & A_2 & C_2 & & & 0 & \\ & \ddots & \ddots & \ddots & & & \\ & & & B_j & A_j & C_j & \ddots \\ & & 0 & & \ddots & \ddots & \ddots \\ & & & & & B_{J+1} & A_{J+1} \end{bmatrix} \qquad (3.2.14)$$

or written in a shorthand notation

$$\frac{\partial \underset{\sim}{\Phi}}{\partial \underset{\sim}{F}} = [B_j, A_j, C_j] \quad . \qquad (3.2.15)$$

Note that the upper half (i.e., the first 2N rows) of each C_j and the lower half (i.e., the last 2N rows) of each B_j are all zeros; this will become important in determining the easiest and most efficient scheme for solving the linear system. The specific elements of the block matrices B_j, A_j, and C_j are given in Appendix A.

Keller [7] examines direct factorization methods for linear systems of the form

$$\mathcal{A} \underset{\sim}{\mathcal{X}} \equiv \underset{\sim}{\mathcal{R}} \qquad (3.2.16)$$

where $\mathcal{A}$ is the block tridiagonal matrix (3.2.14) and

$$\underset{\sim}{\mathcal{X}}^T \equiv (\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_{J+1}), \quad \underset{\sim}{\mathcal{R}}^T \equiv (\underset{\sim}{r}_1, \cdots, \underset{\sim}{r}_{J+1}) . \qquad (3.2.17)$$

We briefly summarize that work here, and then go further.

We seek factorizations of the form

$$\mathcal{A} = \mathcal{L}\mathcal{U} \ , \ \mathcal{L} \equiv [\beta_j, \delta_j, 0] \ , \ \mathcal{U} \equiv [0, \alpha_j, \gamma_j] \ , \quad 1 \leq j \leq J+1 \qquad (3.2.18)$$

(in our previous shorthand notation where β_j, δ_j, α_j, γ_j are $4N^{\underline{th}}$ order matrices) so that L and U are block lower and upper triangular, respectively. Solving (3.2.16) is equivalent to solving the two systems

$$L\underset{\sim}{Y} = \underset{\sim}{R} \quad , \quad U\underset{\sim}{X} = \underset{\sim}{Y} \; . \tag{3.2.19}$$

Then, merely carrying out the multiplication of LU gives the following recursion relations for the β_j, δ_j, α_j, γ_j:

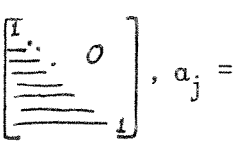
$$\begin{aligned} &\delta_1 \alpha_1 = A_1 \\ &\delta_j \gamma_j = C_j \qquad , \quad 1 \leq j \leq J \\ &\beta_j \alpha_{j-1} = B_j \qquad , \quad 2 \leq j \leq J+1 \\ &\delta_j \alpha_j = A_j - \beta_j \gamma_{j-1} \qquad , \quad 2 \leq j \leq J+1 \end{aligned} \tag{3.2.20}$$

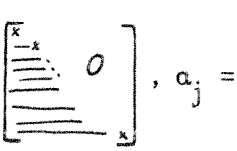
The solution of (3.2.19) is then obtained by solving the $4N^{\underline{th}}$ order systems

$$\delta_1 \underset{\sim}{Y}_1 = \underset{\sim}{r}_1 \qquad\qquad \alpha_{J+1} \underset{\sim}{X}_{J+1} = \underset{\sim}{Y}_{J+1} \tag{3.2.21}$$

$$\delta_j \underset{\sim}{Y}_j = \underset{\sim}{r}_j - \beta_j \underset{\sim}{Y}_{j-1}, \quad 2 \leq j \leq J+1 \qquad \alpha_j \underset{\sim}{X}_j = \underset{\sim}{Y}_j - \gamma_j \underset{\sim}{X}_{j+1}, \quad J \geq j \geq 1$$

Keller distinguishes four common choices of the δ_j and α_j as follows:

case i) $\delta_j = I$ case iii) $\delta_j =$  , $\alpha_j =$ 

case ii) $\alpha_j = I$ case iv) $\delta_j =$  , $\alpha_j =$ 

(3.2.22)

Operational counts for our problem using these methods are (to leading orders):

$$\text{case i)}\quad J\left(\frac{208}{3}N^3 + 32N^2 - \frac{4}{3}N\right)$$

$$\text{case ii)}\quad J\left(\frac{280}{3}N^3 + 44N^2 - \frac{4}{3}N\right) \qquad (3.2.23)$$

$$\text{cases iii) and iv)}\quad J\left(\frac{184}{3}N^3 + 32N^2 - \frac{4}{3}N\right) .$$

It is especially important to notice here that the work involved in solving for one Newton iterate increases with the cube of the number of series terms, N, and only linearly with the number of mesh intervals J. Hence, doubling the number of terms taken in the Legendre series involves eight times the computing time per Newton iterate.

In performing the actual computations, we began by using case (i) of (3.2.22) in order to ease the coding difficulty (since some subroutines needed for this method had been previously written for other calculations). Although good results were obtained for the lower Reynolds number ($R < 10$), poor convergence and even divergence was obtained for Newton's method at higher Reynolds numbers. Believing this to be a problem of ill-conditioning and accumulation of round-off error when computing (3.2.20) and (3.2.21), we tried iterative improvement techniques to refine our solution. This gave improved results at first, but as R and N were increased, the convergence of Newton's method became very bad once again. Convinced that Newton's method should be converging and that the problem was in the solution of the linear system (3.2.10), we tried a new factori-

zation scheme which worked quite well.

This new factorization was suggested by the "parallel shooting" method proposed by Keller [8] for numerically solving systems of ordinary differential equations with multi-point boundary conditions. The parallel shooting method is used when accumulation of truncation error in using a regular shooting method leads to too inaccurate a solution. The method splits the interval of integration into several subintervals and performs the integration over each subinterval. The solutions over the subintervals are patched together using appropriate continuity requirements. This results in an exponential decrease in truncation error accumulation. We use an analogous "parallel shooting" factorization in solving (3.2.10) to reduce the propagation of round-off error through the recursions (3.2.20), (3.2.21).

In this factorization, we consider methods which basically partition the block tridiagonal matrix (3.2.14) into two parts which are then factored into LU and UL decompositions. The factorization proceeds as follows, for some fixed k:

$$A = \begin{bmatrix} \alpha_1 & & & & & \\ \beta_2 & \alpha_2 & & & & \\ & \ddots & \ddots & & & \\ & & \beta_k & \alpha_k & 0 & \\ & & 0 & \delta_{k+1} & \gamma_{k+1} & \\ & & & & \ddots & \ddots \\ & & & & & \gamma_{n-1} \\ & & & & & \delta_n \end{bmatrix} \begin{bmatrix} \delta_1 & \gamma_1 & & & & \\ & \ddots & \ddots & & & \\ & & \delta_{k-1} & \gamma_{k-1} & & \\ & & 0 & \delta_k & \gamma_k & \\ & & & \beta_{k+1} & \alpha_{k+1} & 0 \\ & & & & \beta_{k+2} & \alpha_{k+2} \\ & & & & & \ddots \\ & & & & & \beta_{n+1} \alpha_{n+1} \end{bmatrix} \tag{3.2.24}$$

This factorization gives rise to the following recursion relations for the α_j, β_j, δ_j, γ_j:

$$\left.\begin{aligned} &\alpha_1\delta_1 = A_1 \\ &\alpha_j\gamma_j = C_j \\ &\beta_{j+1}\delta_j = B_{j+1} \\ &\alpha_{j+1}\delta_{j+1} = A_{j+1} - \beta_{j+1}\gamma_j \end{aligned}\right\} \quad j = 1, 2, \ldots, k-1$$

$$\delta_{J+1}\alpha_{J+1} = A_{J+1} \qquad (3.2.25)$$

$$\left.\begin{aligned} &\gamma_j\alpha_{j+1} = C_j \\ &\delta_{j+1}\beta_{j+1} = B_{j+1} \\ &\delta_j\alpha_j = A_j - \gamma_j\beta_{j+1} \end{aligned}\right\} \quad j = J, J-1, \ldots, k+1$$

$$\alpha_k\gamma_k = C_k$$

$$\delta_{k+1}\beta_{k+1} = B_{k+1}$$

As in the other direct factorizations, solving the system (3.2.16) is equivalent to solving the two systems (3.2.19). For the current factorization (3.2.24), the recursion relations for the $\underset{\sim}{x}_i$ and $\underset{\sim}{y}_i$ are

$$\alpha_1\underset{\sim}{y}_1 = \underset{\sim}{f}_1 \qquad\qquad \delta_{J+1}\underset{\sim}{y}_{J+1} = \underset{\sim}{f}_{J+1}$$

$$\alpha_j\underset{\sim}{y}_j = \underset{\sim}{f}_j - \beta_j\underset{\sim}{y}_{j-1} \quad j = 2, \ldots, k \qquad\qquad \delta_j\underset{\sim}{y}_j = \underset{\sim}{f}_j - \gamma_j\underset{\sim}{y}_{j+1}, \; j = J, J-1, \ldots, k+1$$

(3.2.26)

$$\delta_k \underset{\sim}{x}_k + \gamma_k \underset{\sim}{x}_{k+1} = \underset{\sim}{y}_k$$

$$\beta_{k+1} \underset{\sim}{x}_k + \alpha_{k+1} \underset{\sim}{x}_{k+1} = \underset{\sim}{y}_{k+1} \qquad (3.2.27)$$

$$\delta_{j-1} \underset{\sim}{x}_{j-1} = \underset{\sim}{y}_{j-1} - \gamma_{j-1} \underset{\sim}{x}_j \quad , \quad j = k, k-1, \ldots, 2$$

$$\alpha_{j+1} \underset{\sim}{x}_{j+1} = \underset{\sim}{y}_{j+1} - \beta_{j+1} \underset{\sim}{x}_j \quad , \quad j = k+1, k+2, \ldots, J \quad . \qquad (3.2.28)$$

Using the above relations, we solve the linear system in an "inward sweep," "link," and an "outward sweep." We first solve equations (3.2.25) for the α_j, δ_j, β_j, γ_j and at the same time solve equations (3.2.26) for the $\underset{\sim}{y}_j$ as j goes toward k from 1 and J+1. This constitutes the "inward sweep." Next we solve equations (3.2.27) for $\underset{\sim}{x}_k$ and $\underset{\sim}{x}_{k+1}$. This is the "link" between the two halves of the partitioning of $\mathcal{A}$. Then using these values of $\underset{\sim}{x}_k$ and $\underset{\sim}{x}_{k+1}$ to start, we perform the "outward sweep," solving equations (3.2.28) for the rest of the $\underset{\sim}{x}_j$.

The specific choice of α_j, δ_j we have used for computing has been

$$\alpha_j = I, \; j = 1, \ldots, k, \quad \delta_j = I, \; j = k+1, \ldots, J+1 \; . \qquad (3.2.29)$$

This "parallel shooting" decomposition has proven much more stable for appropriate values of k than the scheme (3.2.18) in terms of propagation of round-off error through the sweeps. That is, it gets much more accurate solutions to our linear system than did the other schemes. In general, for our system (3.2.10), the greatest accuracy was obtained for k approximately equal to J/2; for k near

J+1, the accuracy was degraded substantially. (Note that for k = J+1, the parallel shooting decomposition becomes the same as the straight LU decomposition (3.2.18).)

Having solved the linear system (3.2.10) for $\underset{\sim}{\delta F}^{(\nu)}$, we generate $\underset{\sim}{F}^{(\nu+1)}$ from (3.2.9) and we have a new Newton iterate. We continue generating Newton iterates until some error criterion is met. This error criterion and other sources of error are discussed in the next section.

3.3 Error Analysis

To estimate how close the last Newton iterate is to the exact solution, we observe that there are five sources of error:

(1) Newton's method error (in a finite number of iterations),

(2) Truncation error in the finite difference discretization,

(3) Series truncation error,

(4) Error from the outer boundary conditions,

(5) Accumulation of round-off error.

Of these, the first two can be analyzed theoretically and computationally, and the last three less thoroughly. We will discuss each one separately.

Keller [7] shows that under certain conditions (involving isolatedness of the solution to (3.2.5), (3.2.6) and closeness of the initial approximation of (3.2.7)) that Newton's method converges quadratically. By quadratic convergence, we mean that the error $\underset{\sim}{e}^{(\nu)}$ between the $\nu^{\underline{th}}$ Newton iterate $\underset{\sim}{F}^{(\nu)}$ and the actual root $\underset{\sim}{F}^*$ of (3.2.8) satisfies

$$\|\underset{\sim}{e}^{(\nu+1)}\| \leq k \|\underset{\sim}{e}^{(\nu)}\|^2 . \qquad (3.3.1)$$

In solving (3.2.8) by Newton's method, we observed the Newton corrections, $\delta\underset{\sim}{F}^{(\nu)}$, and observed "quadratic convergence" in the sense:

$$\| \delta\underset{\sim}{F}^{(\nu+1)} \| \leq K \| \delta\underset{\sim}{F}^{(\nu)} \|^2 \tag{3.3.2}$$

This can be shown to be equivalent to (3.3.1). Another measure of the convergence is in the change of various physical parameters, especially the drag coefficient (C_D), as calculated from successive Newton iterates. The drag coefficient as given by (3.1.14) can be calculated at each Newton iterate; we call it $C_D^{(\nu)}$ for the ν^{th} Newton iterate. Then if

$$\delta C_D^{(\nu)} = C_D^{(\nu)} - C_D^{(\nu-1)}, \quad \nu = 1, 2, \ldots \tag{3.3.3}$$

we observed that

$$|\delta C_D^{(\nu)}| \leq K \, |\delta C_D^{(\nu-1)}|^2 \tag{3.3.4}$$

both for the rigid sphere and the gas bubble for all Reynolds numbers. Table 1 shows these observations for various Reynolds numbers.

The generation of Newton iterates is normally terminated when

$$|\delta C_D^{(\nu)}| < 10^{-3} \tag{3.3.5}$$

(sometimes the iterates were terminated before or, more frequently, after this criterion has been met, but all results for which physical or theoretical results are cited satisfy at least (3.3.5)).

When computations were made for Reynolds number 40, we found that Newton's method would not converge quadratically after two iterations as was the usual case. Upon increasing the number

of net points, J, from 30 to 60, however, Newton's method converged quadratically once more. The mechanism behind this behavior is not definitely understood, but the behavior was also observed by Keller and Nieuwstadt [9]. In this connection, it should be recalled that the theoretical quadratic convergence proof for Newton's method in Keller [7] required that the net spacing be sufficiently small. We expect finer nets to be required for the same accuracy as R increases (since the solutions change faster near the body), and it may be this effect that is observed here.

The second source of error is the truncation error involved in the approximation of the differential equation (3.2.5) by the box scheme discretization (3.2.7). Keller [7] shows that if $\mathcal{N}(\xi, \mathcal{Z})$ and $\mathcal{Z}(\xi)$ are sufficiently smooth, then

$$\mathcal{Z}(\xi_j) - \mathcal{Z}_j = \sum_{\nu=1}^{m} \left(\frac{h}{2}\right)^{2\nu} \mathcal{e}_\nu(\xi_j) + O(h^{2m+2}), \quad 0 \leq j \leq J \ . \qquad (3.3.6)$$

Knowing this, Richardson extrapolation can be used to improve the accuracy by constructing

$$\overline{\mathcal{Z}}(\xi_j) = \frac{4\mathcal{Z}_{j,2} - \mathcal{Z}_{j,1}}{3} \qquad (3.3.7)$$

where $\mathcal{Z}_{j,1}$ is the approximation to $\mathcal{Z}(\xi_j)$ computed from (3.3.7) using mesh spacing h, and $\mathcal{Z}_{j,2}$ is the approximation to $\mathcal{Z}(\xi_j)$ using mesh spacing h/2. Then from (3.3.6) we have

$$\mathcal{Z}(\xi_j) - \overline{\mathcal{Z}}_j = O(h^4) \ .$$

One measure of the accuracy of a solution to (3.3.7) is obtained by computing the drag coefficient $C_D(\xi)$ from (3.1.13) at

different values of ξ. The more accurate the solution, the more nearly constant $C_D(\xi)$ should be. Keller and Takami [10] use this as a measurement of the accuracy of their solutions for the cylinder problem. More recently, Keller and Nieuwstadt [9] showed that the constancy of $C_D(\xi)$ improved remarkably with the application of Richardson extrapolation.

In our solutions we observed the same improvement in the constancy of $C_D(\xi)$. For example, as shown in Table 2, for Reynolds number 0.5 and J = 31 and $\xi_\infty = 4.9$, the calculated value of $C_D(\xi)$ for the gas bubble dropped 12.3% from $\xi = 0$ to $\xi = \xi_\infty$; however, using Richardson extrapolation, the corresponding change was only .27%. Similar improvement was noted for other Reynolds numbers. Although (due to the presence of other errors) this does not prove the validity of the extrapolation procedure for our problem, it does demonstrate a dramatic improvement in the computed solution.

The third source of error is that stemming from the replacement of the infinite series for ψ and ζ (3.1.1) by the finite series of N terms each. This error is much more difficult to treat than the two previously discussed errors because there is no expression for the remainder of the series. Since we cannot treat this error theoretically, we try to estimate the number of terms necessary for the desired accuracy empirically. We compute the solution for successively larger values of N until the solution is as accurate as is consistent with the computing budget. This same practice has been used by Keller and Nieuwstadt [9), Dennis and Walker [3], and others. In fact, our experiments with various numbers of terms led

us to agree with the final choices of N made by Dennis and Walker. Table 3 shows the effects on $C_D(0)$ of changing N for several Reynolds numbers. Figure 2 shows the effect of varying N on $\zeta(0,\theta)$ for R = 5 for the rigid sphere.

One interesting observation is that the f_n and g_n are so small for n greater than about four that these terms are negligible in summing the series for C_D. One might think, then, that in solving (3.2.2) a smaller value of N could be used. However, inclusion of these higher terms changes the values of the first few f_n and g_n substantially, which in turn changes the calculated value of C_D.

The error introduced by imposing an outer boundary at a finite distance from the sphere with asymptotic boundary conditions is also difficult to treat theoretically. Again we relied on an empirical evaluation of the appropriate distance and boundary conditions. We tried only two values of ξ_∞, $\xi_\infty = \pi$ and $\xi_\infty = 4.9$. The former is the distance used by Keller and Nieuwstadt and the latter is used by Dennis and Walker. For R = 0.1, $\xi_\infty = \pi$ gave $C_D(0) = 121.1$ while $\xi_\infty = 4.9$ gave $C_D(0) = 118.1$, both being computed with J = 31. Since the greater value of ξ_∞ is most likely the more accurate, and since the difference was substantial, we chose $\xi_\infty = 4.9$. Note that $\xi_\infty = \pi$ corresponds to the outer boundary's being taken at 23.14 body radii while $\xi_\infty = 4.9$ corresponds to 134.2 body radii or more than five times as far away.

As mentioned before, the conditions to be met by ψ and ζ on the outer boundary were the Oseen conditions given by Batchelor [2]. It was virtually no more expensive to use this condition than the free

stream condition. However after much computation using the Oseen conditions, a test was made to see how a solution computed using the free stream condition would differ. For R = 5, J = 31, $\xi_\infty = 4.9$, $C_D(0)$ was 2.09575 for the Oseen solution and 2.09581 for the free stream solution, a difference of only 0.6×10^{-4}, or less than the Newton iteration error. Also, the $g_n(\xi_\infty)$ were typically on the order of 10^{-6} and the $\frac{f_n(\xi_\infty)}{f_1(\xi_\infty)}$, $n > 1$ were on the order of 10^{-4}. Changes of this magnitude cannot influence the solution substantially and so, perhaps, the free stream conditions might have been used, but were not.

Accumulation of round-off, the last source of error, is dependent upon the computer used to calculate the solutions. Our calculations were performed on an IBM 370/155 which carries about seven significant digits in single precision. At about R = 10 and N = 14, we encountered a failure of Newton's method to converge using the standard decomposition schemes (3.2.18) to solve the linear systems. Residual correction methods as discussed in Isaacson and Keller [5] were applied to reduce the round-off error to the point where Newton's method would converge. While this worked satisfactorily for the rigid sphere solutions, it did not work for the gas bubble. When the parallel shooting decomposition (3.2.24) was tried, the results improved dramatically. Once again, Newton's method was seen to converge quadratically. No residual correction procedure was necessary for accurate solution of the linear systems when using the parallel shooting decomposition. Although round-off accumulation

(greatly aggravated by the ill-conditioned nature of the linear system) was a serious problem, it is believed that it is no longer a threat to the accuracy of the solutions obtained.

4. RESULTS AND ASSESSMENT

In this chapter we present and discuss the results and conclusions gained from our solution of the Navier-Stokes equations. We compare our results with theoretical and experimental results and the numerical results of others.

The following is a list of the Reynolds numbers for which solutions were computed, along with the largest used values of N and J, whether Richardson extrapolation was done and whether computations were performed for the rigid sphere, gas bubble or both:

R	N	J	Richardson	Rigid/Gas	
0.1	6	60	yes	both	
0.5	6	60	yes	both	
1	6	60	yes	both	
5	6	60	yes	gas	
5	14	60	yes	rigid	
10	16	30	no	both	
20	20	30	no	gas	
40	20	60	yes	gas	
60	20	90	yes	gas	(4.1)

Our computations for flow past a rigid sphere were largely in preparation for the work on the gas bubble. Because of this we only computed solutions for Reynolds numbers 0.1, 0.5, 1, 5, and 10 and went no higher. Our main purpose was to test out the numerical method and the computer code. Nevertheless, we did check our computations with the results of Dennis and Walker. We found very

close agreement with them on drag coefficients and pressure coefficients and vorticity values. The following table shows our values and those of Dennis and Walker for the drag coefficient C_D and the pressure coefficient at $\theta = 0$ and $\theta = \pi$, as well as the drag coefficients calculated by Le Clair, et al. [12].

R	Present $C_D(0)$	Dennis & Walker C_D	Le Clair et al. C_D	Present $k(0)$	Present $k(\pi)$
0.1	122.10	122.10	122.04	-60.07	62.03
0.5	25.74	25.85	-	-12.03	13.84
1.0	13.72	13.72	13.66	-5.889*	7.366*
5.0	3.594	3.605	3.515	-1.243	2.546
10.0	2.074*	2.212	2.144	-.6627*	1.726*

*computed without Richardson extrapolation. (4.2)

Dennis and Walker computed solutions for Reynolds numbers as high as 40 and extrapolated the onset of separated flow behind the sphere at Reynolds number 20.5. Since our calculations did not extend that far, we were unable to observe separated flow. Figures 3 and 4 show streamline and equivorticity line plots computed for Reynolds number 10. Figure 5 shows the development of vorticity at the sphere surface with Reynolds number. Figure 6 shows the development of pressure at the sphere surface with increasing Reynolds number.

Dennis and Walker stated that Reynolds number 40 was approximately the upper limit of their numerical treatment since as Reynolds number, R, and the number of series terms, N, increased, the relaxation parameter they are forced to use becomes so small

that their convergence criteria become meaningless. Also as this relaxation parameter becomes small, the number of iterations and hence the amount of computation become forbiddingly large.

Since we have consistently observed the quadratic convergence of Newton's method even for large R and N, we feel confident that our numerical method is not only more efficient than theirs, but also capable of going to much higher Reynolds numbers. In fact as we have mentioned and will discuss further, we carried our calculations to Reynolds number 60 for the gas bubble with no weakening of the method. Keller and Nieuwstadt [9] used the same type of numerical technique on the cylinder with up to 30 terms in the (Fourier) series with few difficulties. They were thus solving 120-component systems. Hence, we believe our method valid for Reynolds numbers of 100 or greater and 30 or more terms in the series.

Our main interest is in the calculation of solutions of fluid flow past a spherical gas bubble. To our knowledge calculations for this problem have never been performed before by any method. This problem does not lend itself easily to solution by methods like those of Dennis and Walker which iterate between solving first an equation for the vorticity, then the equation for the stream function, etc. This is because the no-stress boundary condition at the bubble surface (2.8) ties the stream function and the vorticity together there as is seen by the expression for the f_n' and g_n in (3.1.8b'). Our method can treat this problem just as naturally as the rigid sphere problem because of the ease with which it handles very general boundary conditions.

We now compare the results gained from our numerical calculations with the theoretical results of Levich [13] and Moore [15] for high Reynolds number and the standard results for low Reynolds number cited by Batchelor [2]. We will also mention briefly experimental results cited by Levich and their disagreement with both theory and our numerical results.

Batchelor derives an asymptotic formula for the drag coefficient at low Reynolds number:

$$C_D = \frac{8}{R} . \tag{4.3}$$

This value is exactly 2/3 the value for a rigid sphere. Our calculations agree with this formula very well for $R \leq 0.1$. The table below shows, for each Reynolds number, the theoretical value of C_D, our calculated value, and the difference as a percentage of the theoretical value.

R	$\frac{8}{R}$	Present $C_D(0)$	% diff.	
0.1	80	80.83	1.04	
0.5	16	16.85	5.31	(4.4)
1.0	8	8.795	9.95	
5.0	1.6	2.184	36.3	

Thus we see that (4.3) is not very accurate for $R \geq 0.5$.

The theory for high Reynolds number flow past a bubble is more interesting and has received more attention than low Reynolds number flow. Batchelor [2] derives the first asymptotic term for

the drag coefficient for high Reynolds number flow from the rate of energy dissipation in an irrotational flow and calculates this value as

$$C_D = \frac{24}{R} \quad . \tag{4.5}$$

(Batchelor, Moore, and Levich all define C_D to be $\frac{D}{\frac{1}{2}\pi\rho U^2 a^2}$ in contrast to our definition ($C_D = \frac{D}{\pi\rho U^2 a^2}$); hence, they actually give $C_D = \frac{48}{R}$.) Moore [15] goes further and calculates the energy dissipation rate in the boundary layer at the bubble and in the wake and arrives at the next term in the expansion for C_D, giving

$$C_D = \frac{24}{R}\left(1 - \frac{4\sqrt{2}(6\sqrt{3}+5\sqrt{2}-14)}{5\sqrt{\pi}\ R^{\frac{1}{2}}} + O(R^{-5/6})\right) \tag{4.6}$$

or evaluating the constant involved,

$$C_D = \frac{24}{R}\left(1 - \frac{2.2107}{R^{\frac{1}{2}}} + O(R^{-5/6})\right) \tag{4.7}$$

Our calculations bear out Moore's theoretical result very well. The table below shows, for each Reynolds number, our calculated value of C_D, Moore's value computed from (4.7), the value given by the first term in the expansion (4.5), and the difference between our value and Moore's value as a percentage of the latter.

R	Present $C_D(0)$	Moore C_D	$\frac{24}{R}$	% diff.
10	1.175	.7222	2.4	62.7
20	.6810	.6068	1.2	12.2
40	.4156	.3903	.6	6.47
60	.3001	.2858	.4	5.00

(4.8)

If we assume, as Moore shows, that the next term in the expansion is $O(R^{-5/6})$, and write

$$C_D = \frac{24}{R}\left(1 - 2.2107\, R^{-\frac{1}{2}} + cR^{-5/6}\right) \tag{4.9}$$

we may use our calculated results from (4.8) to determine possible values of c. Then for R = 40, we calculate c to be 0.9167 and for R = 60, we calculate c to be 1.016. Thus c would appear to be roughly 1, lending further support to the order of the third term $R^{-5/6}$.

Another physical parameter of interest in studying flow past drops and bubbles is the velocity of rise of a bubble of a certain size. Levich derives the terminal rate of rise for gas bubbles through a liquid in the presence of gravity by equating the drag on the bubble to the buoyant force acting on the bubble:

$$C_D \pi \rho U^2 a^2 = \frac{4}{3} \pi \rho a^3 g \ . \tag{4.10}$$

where g is the acceleration of gravity.

Levich uses the first term in the expansion so that

$$C_D = \frac{24}{R} = \frac{12\nu}{aU} \ , \tag{4.11}$$

giving for the terminal velocity

$$U = \frac{1}{9} \frac{a^2 g}{\nu} \ . \tag{4.12}$$

If Moore's results had been available at the time, Levich could have used the more accurate expansion (4.6) for C_D and computed

$$U = \frac{1}{9}\frac{a^2 g}{\nu} + \frac{k^2 \nu}{4a} + \frac{k}{2}\sqrt{\frac{\nu}{2a}}\sqrt{\frac{k^2 \nu}{2a} + \frac{4}{9}\frac{a^2 g}{\nu}} \qquad (4.13)$$

where

$$k = \frac{4\sqrt{2}(6\sqrt{3} + 5\sqrt{2} - 14)}{5\sqrt{\pi}} = 2.2107 \ldots \quad .$$

In determining terminal velocity, U, as a function of bubble radius, a, from our numerical solutions, we are faced with only knowing the drag for certain Reynolds numbers. Hence, for a given Reynolds number and drag coefficient, U and a must satisfy

$$\frac{2Ua}{\nu} = R \qquad (4.14)$$

and the terminal velocity condition (4.10). These two equations determine U and a in terms of R, C_D, g, and ν. The solution is

$$U = \tfrac{1}{2}\left[\frac{16Rg\nu}{3C_D}\right]^{\frac{1}{3}} \qquad (4.15)$$

$$a = \left[\frac{3C_D \nu^2 R^2}{16g}\right]^{\frac{1}{3}} \qquad (4.16)$$

Thus, we can, for any fluid whose kinematic viscosity ν is known, compute the radius and rise velocity for given R and C_D. Figure 7 shows the terminal velocity for various radii bubbles rising through pure water at 20° C. It also plots the values given by Levich's theory (4.12) and the modified theory (4.13). Finally, figure 7 shows the same curve when using the low Reynolds number value for C_D (4.5) which gives

$$U = \frac{1}{3}\frac{a^2 g}{\nu} \quad . \qquad (4.17)$$

Levich cites experimental results in support of his theoretical results, but finds large discrepancies between the work of most experimentalists and his theory. He is forced to consider (as are we since our results agree closely with his) the reasons for this discrepancy. There are two possible reasons: either the bubble is not spherical or the fluid is not pure. Although at higher Reynolds numbers the bubble does deform from the spherical, Levich shows that there is no significant deformation until the Reynolds number is greater than about 1600. Since we are dealing with much lower Reynolds numbers, this possibility cannot be regarded as the real reason for the discrepancy. Levich goes on to show that the presence of even a small amount of surface contaminants (or surfactants) can, through the attachment of large molecules to the surface of the bubble, cause the bubble to act as though it were a rigid sphere. Levich also cites experiments conducted by Gorodetskaya with doubly distilled water which agree closely with his theory.

Moore and Levich both agree that any separation of the wake behind the bubble is slight. Levich states that, for example, the separation zone extends only 2° on either side of the line of symmetry behind the bubble at R = 1250! This coincides with our observed lack of separation for $R \leq 60$.

Figure 8 shows the vorticity on the bubble surface with increasing Reynolds number. Figure 9 shows the same kind of plot for the pressure coefficient $k(\theta)$. Figures 10 and 11 show streamline and equi-vorticity line plots for R = 60.

Our computations were performed on an IBM 370/155. Single precision arithmetic (of about seven decimal digits accuracy) was used, except for certain minor calculations (such as computation of 3-J symbols). Since the number of terms N necessary increases with Reynolds number as does the number of mesh intervals J, the amount of computation increases with Reynolds number. The time required for one Newton iterate increases as

$$t_c \alpha N^3 J \quad . \tag{4.18}$$

As an example, the actual computation time for several values of N and J is given below.

N	J	t_c (in secs.)
6	30	22.4
6	60	47.7
10	30	103
20	30	1100
20	60	2100

. (4.19)

The chief advantages of the method of series truncation are due to the problem's being reduced to the solution of ordinary rather than partial differential equations. This eliminates differencing in the θ direction and thus avoids two difficulties - that of using very small θ steps in the wake, and that of mushrooming core requirements as ξ_∞ is increased.

The main disadvantage of the series truncation concept is that it uses series of analytic functions to approximate the vorticity which

behaves non-analytically as $R \to \infty$. This necessitates the use of greater and greater numbers of series terms which then greatly increases the amount of computation required by (4. 18).

For the future we would propose dividing the flow region angularly into sectors and approximating the stream function and vorticity by, say, cubic polynomials in θ whose coefficients are functions of ξ. Continuity requirements across sector boundaries would reduce the number of unknowns considerably. The resulting problem would be a system of ordinary differential equations in ξ. The advantage of this type of method is that the sector containing the wake could be made smaller as Reynolds number increased to more closely approximate the important part of the vorticity.

In conclusion, we believe that the method of series truncation coupled with the centered Euler scheme, Newton's method, and the parallel shooting factorization of the block tridiagonal system is an accurate, efficient method for solving fluid flow problems in separable regions.

APPENDIX A

The elements of the A_j, B_j, and C_j as given by (3.2.14) are given below. The symbol x is symbolic for $h/2$ times some nonzero element of the Jacobian matrix $\frac{\partial \underset{\sim}{N}}{\partial \underset{\sim}{Z}}\left(\xi_{j-\frac{1}{2}}, \frac{\underset{\sim}{Z}_j + \underset{\sim}{Z}_{j-1}}{2}\right)$, the actual value being obvious during the programming.

$$A_1 = \begin{bmatrix} 0 & 0 & I & 0 \\ I & 0 & 0 & \begin{matrix} a_1 & & 0 \\ & \ddots & \\ 0 & & a_N \end{matrix} \\ -I & 0 & \begin{matrix} -(\frac{3}{2})^2 h/2 & & 0 \\ & \ddots & \\ 0 & & -(N+\frac{1}{2})^2 h/2 \end{matrix} & \begin{matrix} -2\frac{h}{2}e^{5/2\xi_{\frac{1}{2}}} & & 0 \\ & \ddots & \\ 0 & & -N(N+1)\frac{h}{2}e^{5/2\xi_{\frac{1}{2}}} \end{matrix} \\ \begin{matrix} -x \cdots -x \\ \vdots \quad \vdots \\ -x \cdots -x \end{matrix} & \begin{matrix} -1+\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & -1+\frac{h}{2}-x \end{matrix} & \begin{matrix} -x \cdots -x \\ \vdots \quad \vdots \\ -x \cdots -x \end{matrix} & \begin{matrix} -2\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & -N(N+1)\frac{h}{2}-x \end{matrix} \end{bmatrix}$$

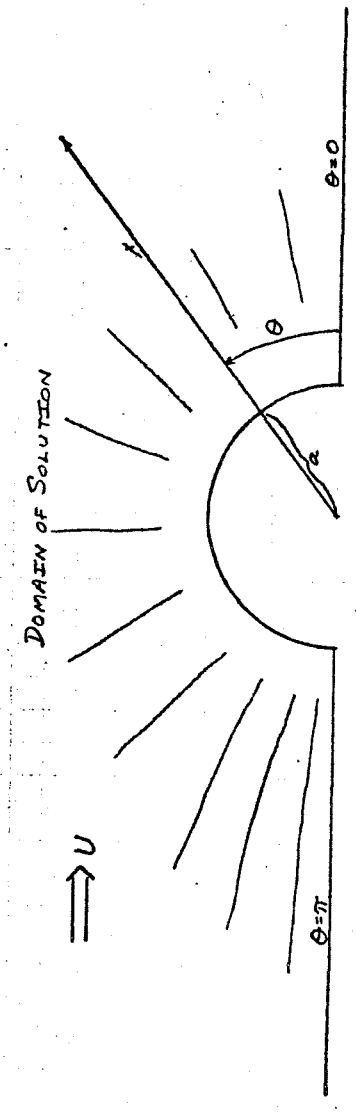
$$A_{J+1} = \begin{bmatrix} -\frac{h}{2}I & 0 & I & 0 \\ 0 & -\frac{h}{2}I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}$$

$$A_j = \begin{bmatrix} -h/2\, I & 0 & I & 0 \\ 0 & -\frac{h}{2} I & 0 & I \\ -I & 0 & \begin{matrix} -(\frac{3}{2})^2 \frac{h}{2} & & 0 \\ & \ddots & \\ 0 & & -(N+\frac{1}{2})^2 \frac{h}{2} \end{matrix} & \begin{matrix} -2\frac{h}{2} e^{5/2\xi_{j-\frac{1}{2}}} & & 0 \\ & \ddots & \\ 0 & & -N(N+1)\frac{h}{2} e^{5/2\xi_{j-\frac{1}{2}}} \end{matrix} \\ \begin{matrix} -x & \cdots & -x \\ \vdots & & \vdots \\ -x & \cdots & -x \end{matrix} & \begin{matrix} -1+\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & -1+\frac{h}{2}-x \end{matrix} & \begin{matrix} -x & \cdots & -x \\ \vdots & & \vdots \\ -x & \cdots & -x \end{matrix} & \begin{matrix} -2\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & -N(N+1)\frac{h}{2}-x \end{matrix} \end{bmatrix}$$

$$B_j = \begin{bmatrix} -\frac{h}{2} I & 0 & I & 0 \\ 0 & -\frac{h}{2} I & 0 & I \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

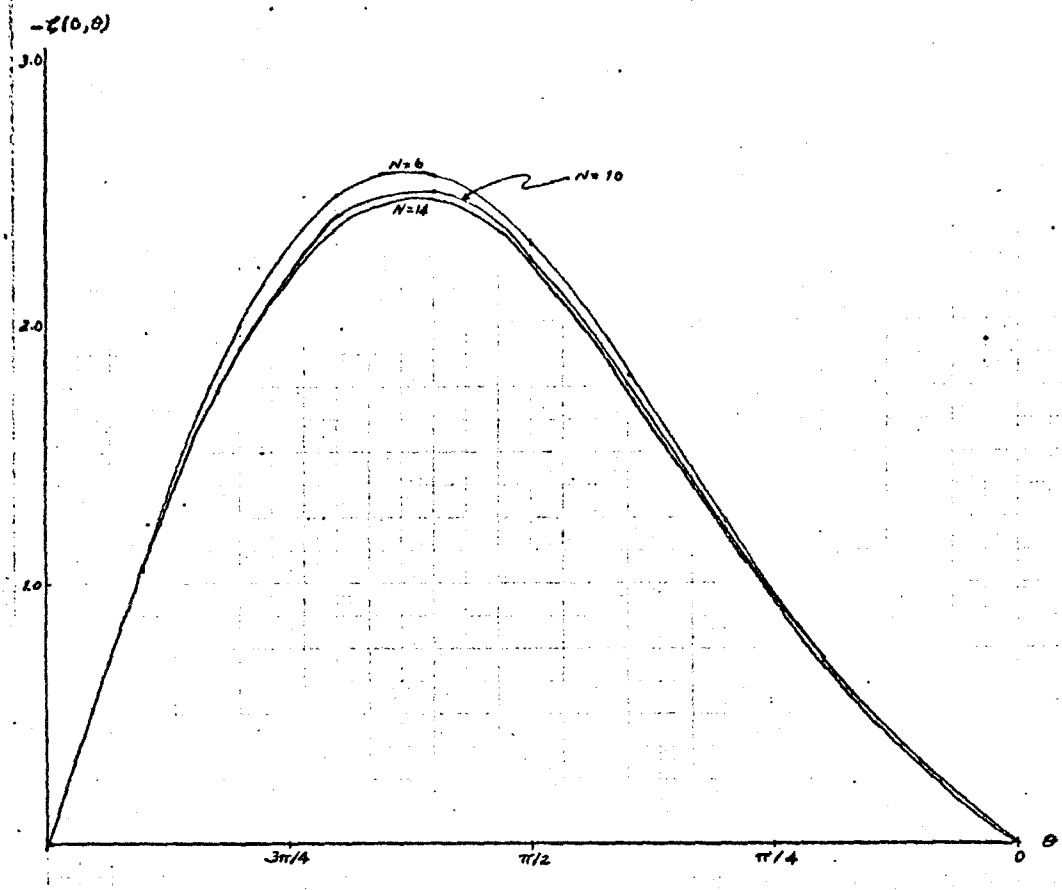
$$C_j = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ I & 0 & \begin{matrix} -(\frac{3}{2})^2 \frac{h}{2} & & 0 \\ & \ddots & \\ 0 & & -(N+\frac{1}{2})^2 \frac{h}{2} \end{matrix} & \begin{matrix} -2\frac{h}{2} e^{5/2\xi_{j-\frac{1}{2}}} & & 0 \\ & \ddots & \\ 0 & & -N(N+1)\frac{h}{2} e^{5/2\xi_{j-\frac{1}{2}}} \end{matrix} \\ \begin{matrix} -x & \cdots & -x \\ \vdots & & \vdots \\ -x & \cdots & -x \end{matrix} & \begin{matrix} 1+\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & 1+\frac{h}{2}-x \end{matrix} & \begin{matrix} -x & \cdots & -x \\ \vdots & & \vdots \\ -x & \cdots & -x \end{matrix} & \begin{matrix} -2\frac{h}{2}-x & & \\ & \ddots & -x \\ -x & & -N(N+1)\frac{h}{2}-x \end{matrix} \end{bmatrix}$$

I is the N x N identity matrix and 0 is the N x N zero matrix. a_j in A_1 is zero for the rigid sphere case and $\frac{1}{2}j(j+1)$ for the gas bubble. Note that B_j and C_j are each half zeros. In general the centered Euler scheme gives rise to this structure for separated end conditions; if there are p conditions at the left end, then B_j will have p rows of zeros and C_j will have n-p rows of zeros (where n is the size of each block).



SPHERICAL POLAR COORDINATE SYSTEM

FIG. 1



EFFECT OF N ON VORTICITY AT RIGID SPHERE SURFACE FOR $R=5$

FIG. 2

FIGURE 3

STREAMLINES FOR R = 10.0

NO = 16 JMAX = 31

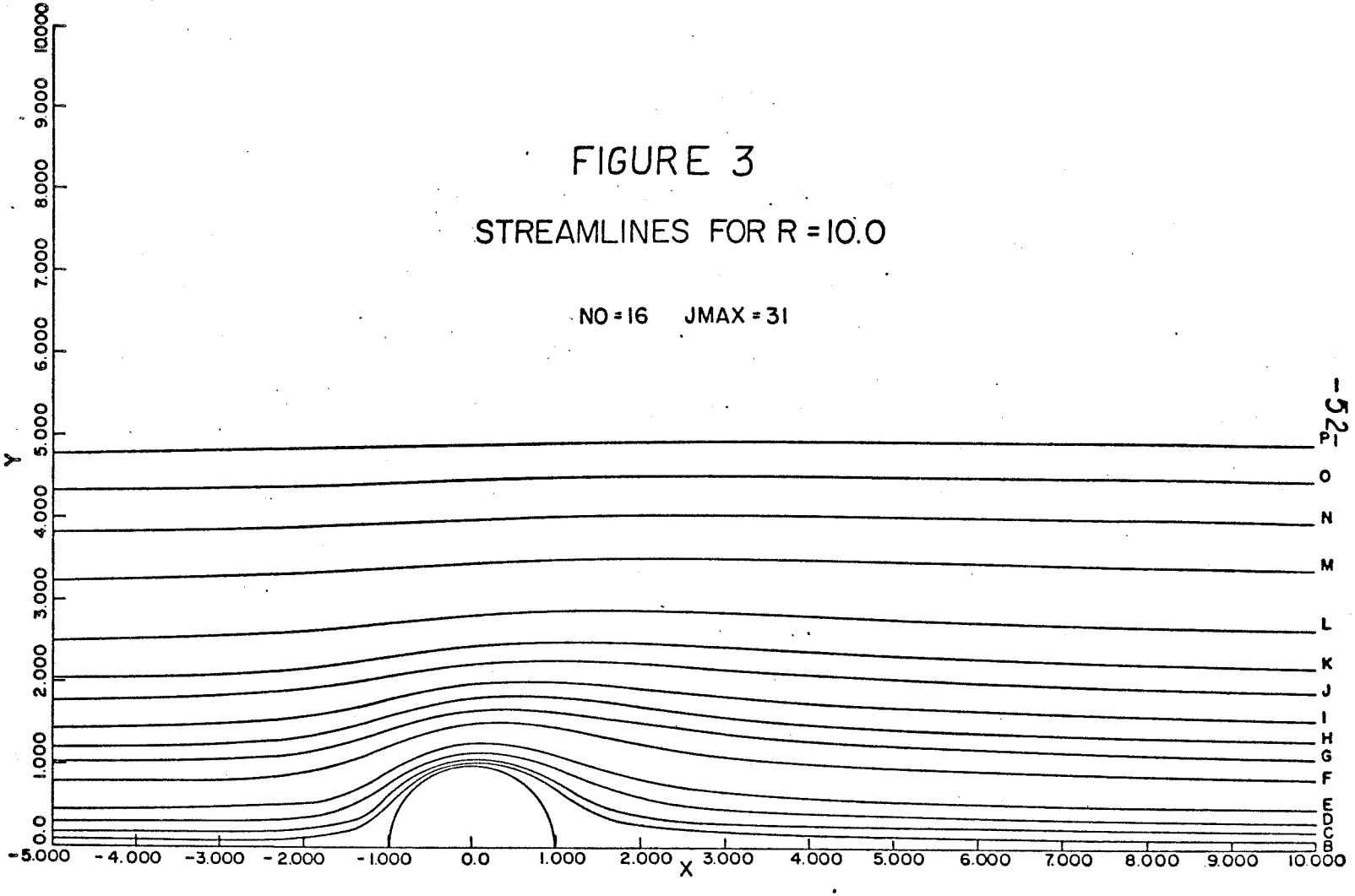
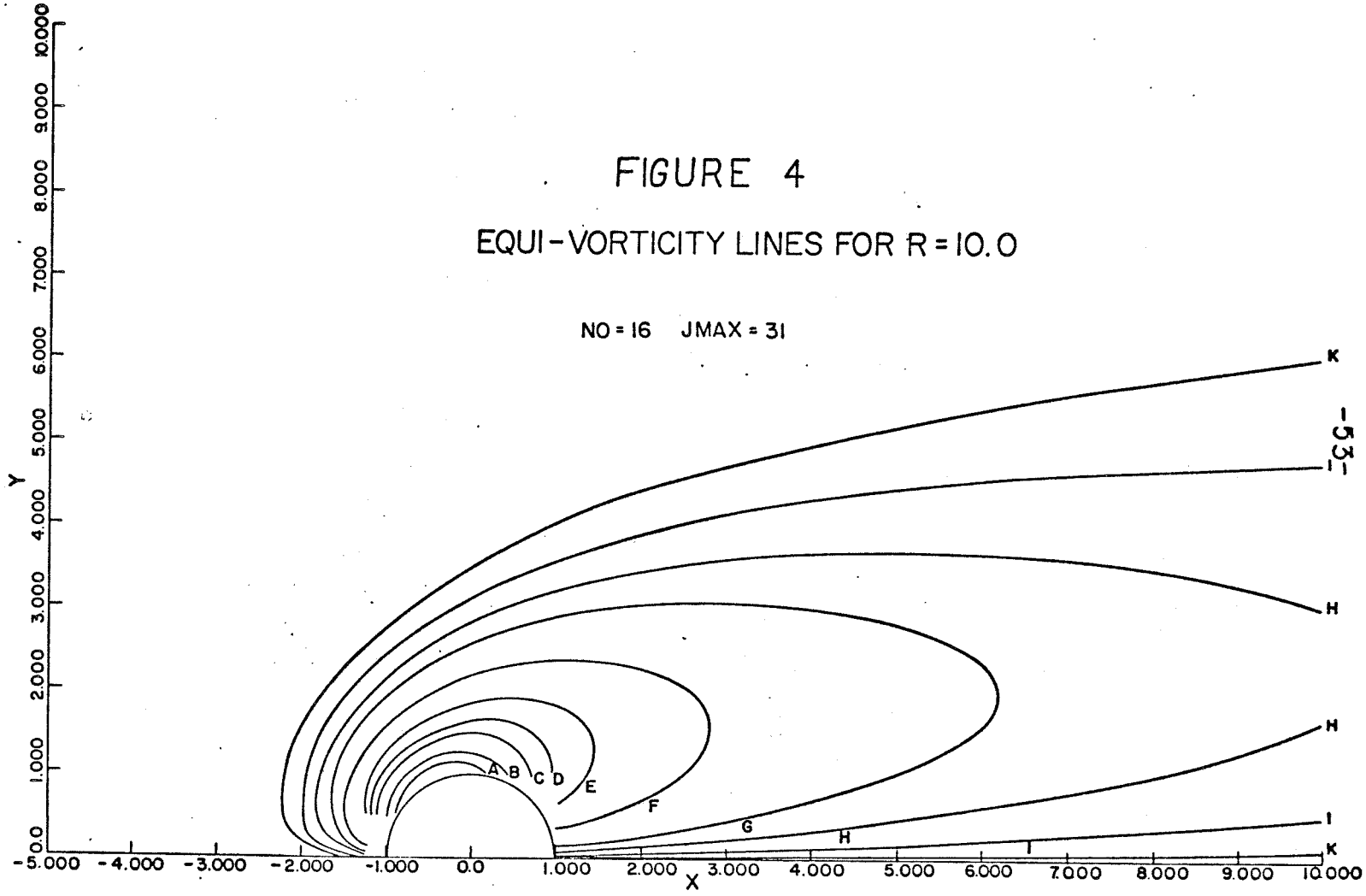
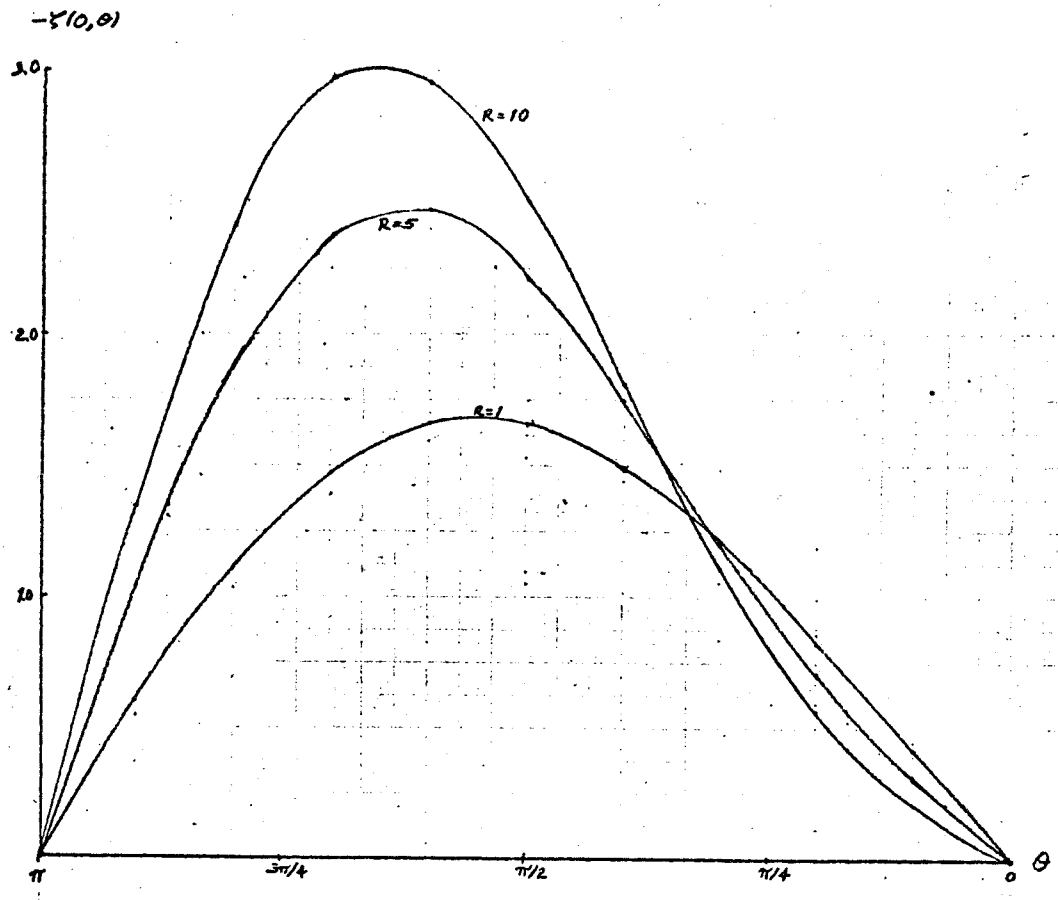


FIGURE 4

EQUI-VORTICITY LINES FOR R = 10.0





VORTICITY ON THE RIGID SPHERE SURFACE

FIG. 5

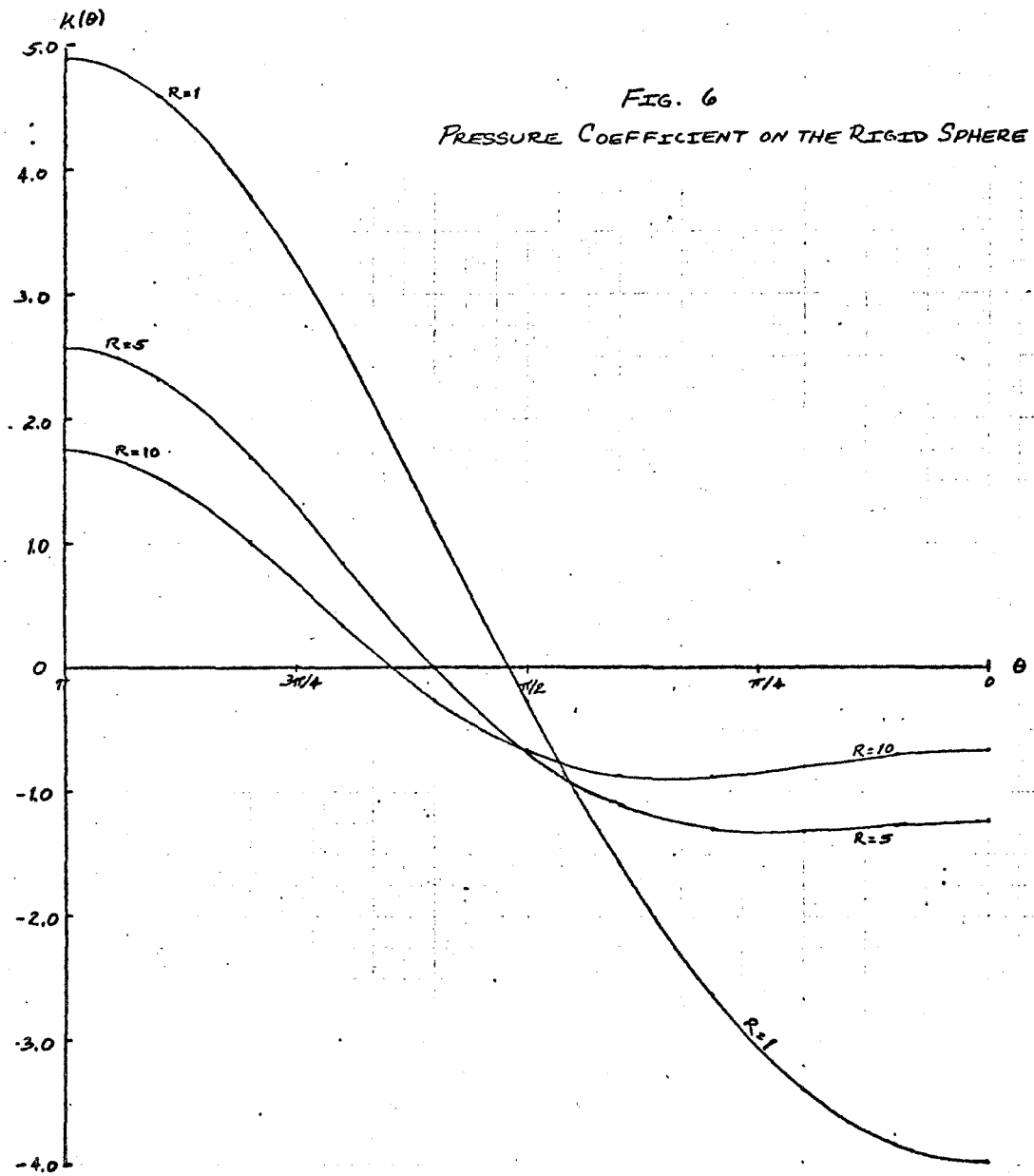


FIG. 6
PRESSURE COEFFICIENT ON THE RIGID SPHERE

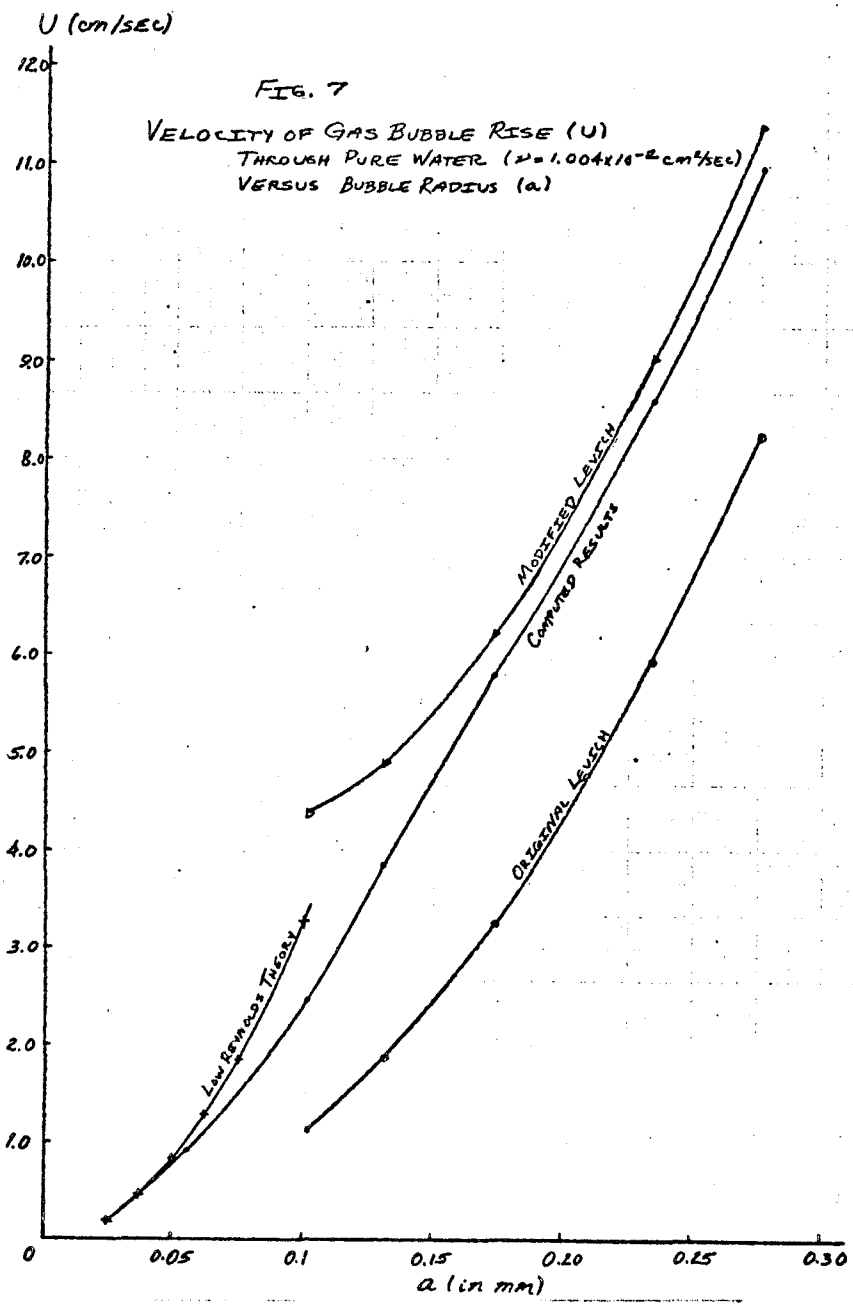
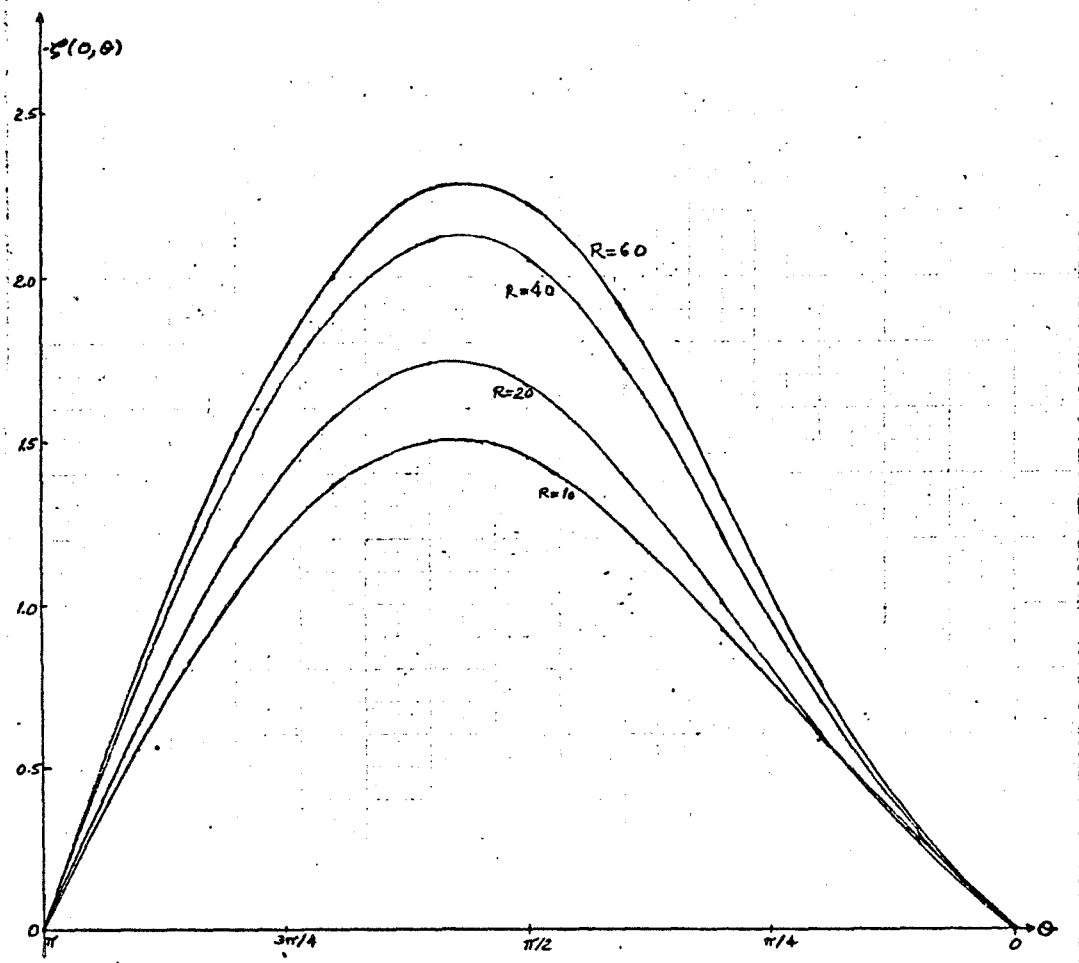


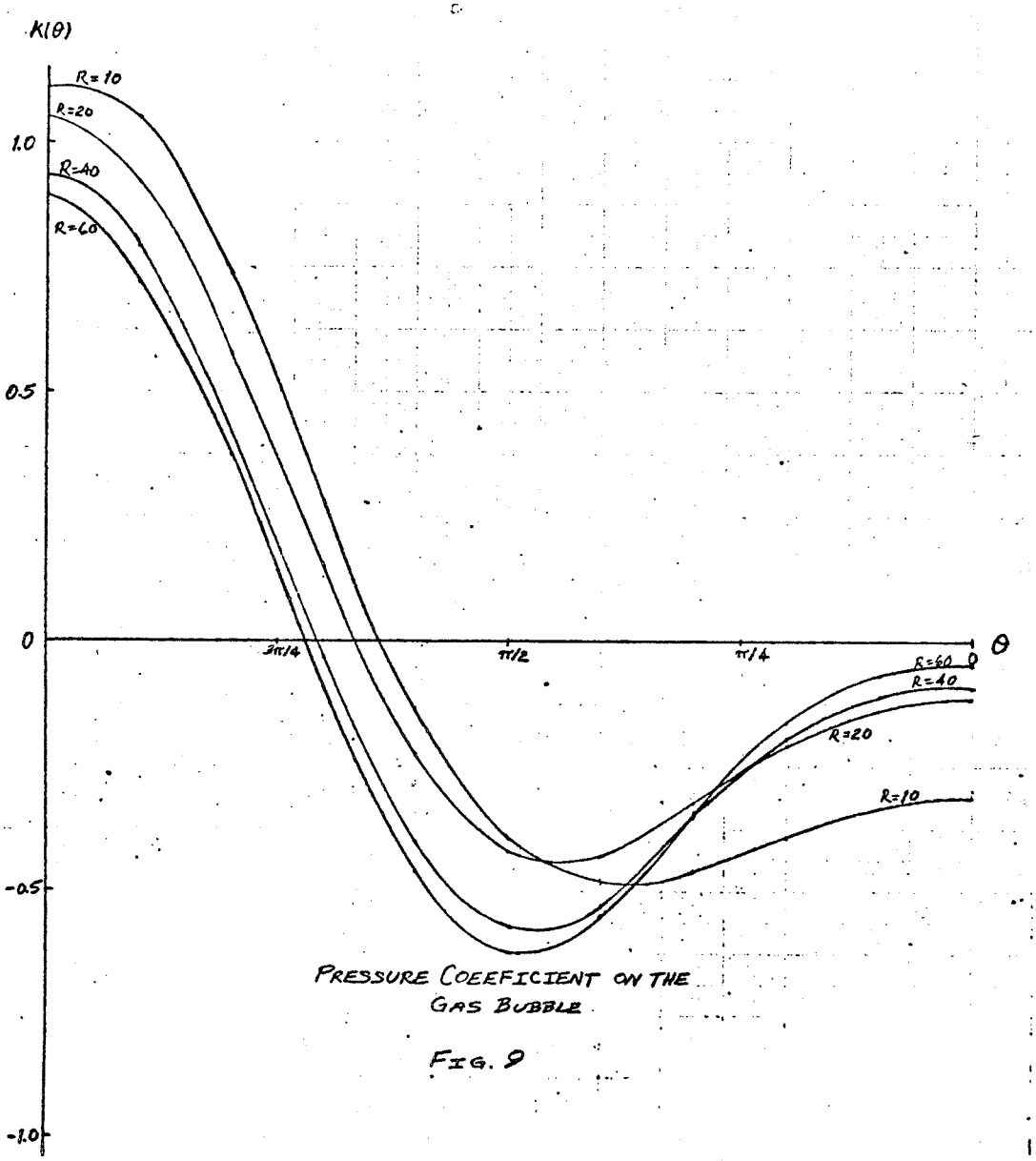
FIG. 7

VELOCITY OF GAS BUBBLE RISE (U)
THROUGH PURE WATER ($\nu = 1.004 \times 10^{-2}$ cm^2/sec)
VERSUS BUBBLE RADIUS (a)



VORTICITY ON THE GAS BUBBLE SURFACE

FIG. 8



PRESSURE COEEFICIENT ON THE GAS BUBBLE

FIG. 9

FIGURE 10

STREAMLINES FOR R=60.0

N = 20 JMAX = 31

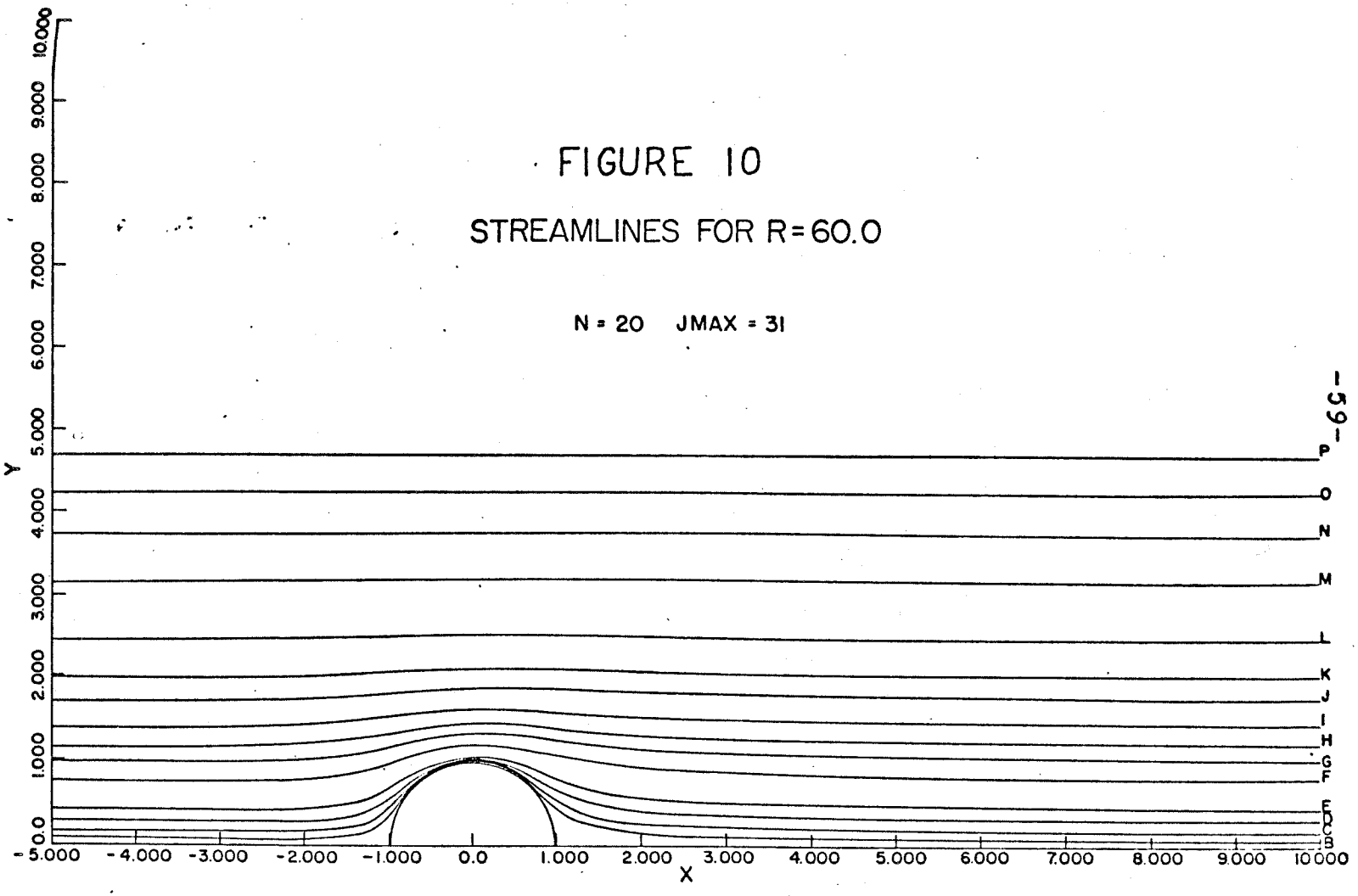
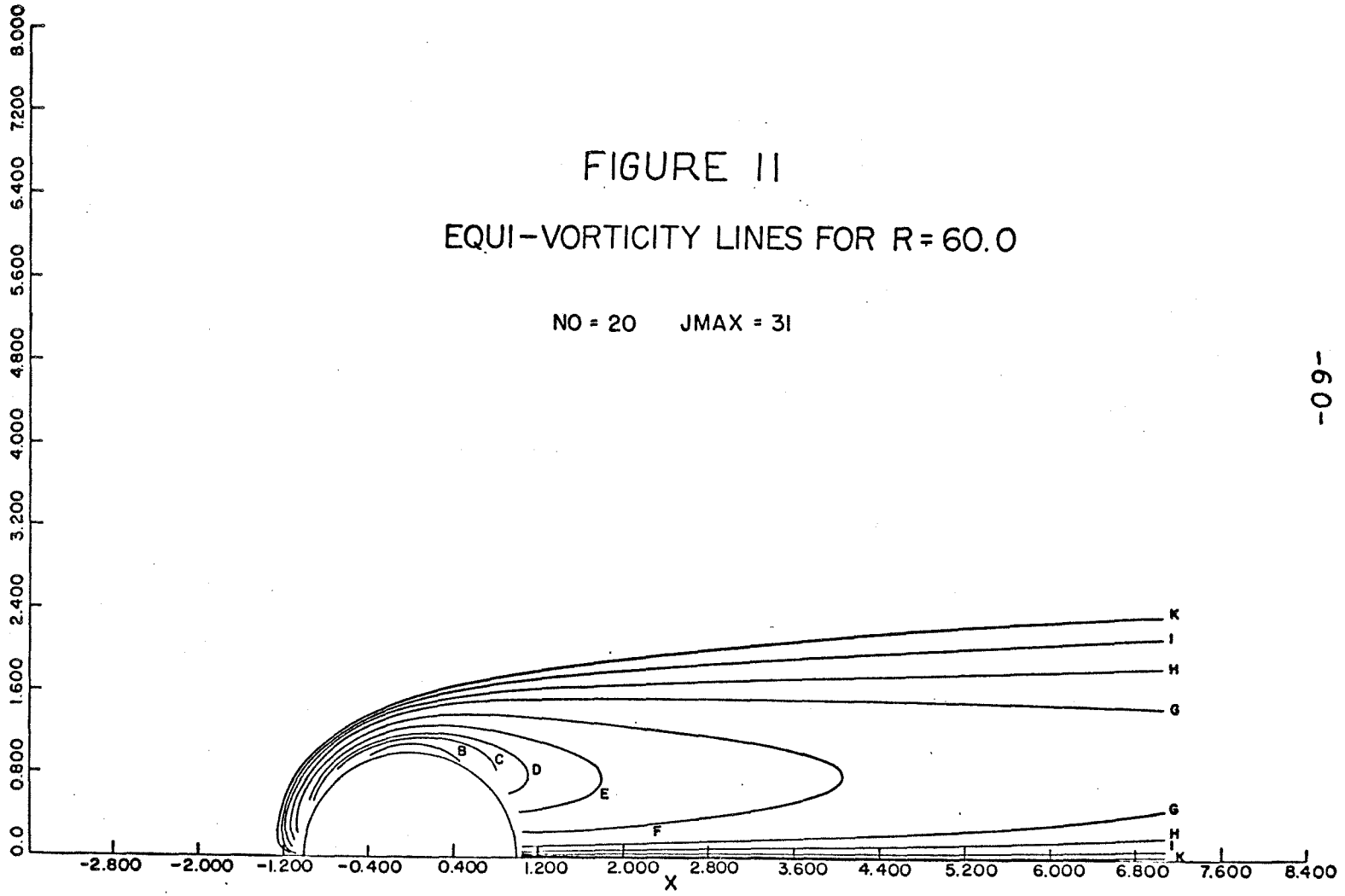


FIGURE 11

EQUI-VORTICITY LINES FOR R = 60.0

NO = 20 JMAX = 31



R	Rigid or Gas	N	J	$\delta C_D^{(1)}$ / $\|\delta \underset{\sim}{F}^{(1)}\|$	$\delta C_D^{(2)}$ / $\|\delta \underset{\sim}{F}^{(2)}\|$	$\delta C_D^{(3)}$ / $\|\delta \underset{\sim}{F}^{(3)}\|$
.1	Gas	6	31	77.61	.4575	.6256 (-3)
				2335.0	3.789	.3388 (-2)
.5	Gas	6	31	.6254	.1053 (-2)	0.0
				4.334	.2635 (-1)	.3812 (-3)
5	Gas	6	31	.3545	.4194 (-3)	.6986 (-5)
				4.375	.1238	.3738 (-3)
20	Gas	16	31	.1567	.1559 (-1)	.5717 (-3)
				2.340	1.161	.4582 (-1)
60	Gas	20	61	.7448 (-1)	.7151 (-2)	.9614 (-4)
				1.799	1.193	.3662
.5	Rigid	6	31	.1890	.1831 (-3)	
				16.19	.1782 (-1)	
10	Rigid	6	31	.1902(-1)	.2731 (-2)	.1717 (-4)
				7.223	.5153	.1661 (-2)

TABLE 1

Convergence of $\delta C_D^{(\nu)}$, $\|\delta \underset{\sim}{F}^{(\nu)}\|$

(.1234 (-5) means $.1234 \times 10^{-5}$)

	$C_D(\xi)$ R = 0.5 Gas Bubble N = 6		$C_D(\xi)$ R = 5 Rigid Sphere N = 14		$C_D(\xi)$ R = 40 Gas Bubble N = 20		$C_D(\xi)$ R = 60 Gas Bubble N = 20	
ξ	I	II	I	II	I	II	I	II
0.0	16.24	16.85	3.429	3.594	.3963	.4156	.2967	.3001
0.32667	16.14	16.85	3.414	3.594	.3993	.4153	.2972	.3000
0.65333	16.05	16.85	3.397	3.594	.3850	.4156	.2938	.3001
0.98000	15.95	16.84	3.373	3.594	.3729	.4158	.2914	.3001
1.3067	15.86	16.84	3.337	3.594	.3630	.4158	.2895	.3001
1.6333	15.76	16.84	3.290	3.593	.3546	.4157	.2879	.3001
1.9600	15.67	16.84	3.182	3.593	.3470	.4155	.2864	.3001
2.2867	15.56	16.84	3.182	3.592	.3401	.4153	.2850	.3001
2.6133	15.45	16.84	3.127	3.591	.3335	.4150	.2837	.3001
2.9400	15.33	16.84	3.070	3.589	.3273	.4146	.2825	.3000
3.2667	15.18	16.83	3.014	3.587	.3213	.4143	.2812	.3000
3.5933	15.01	16.83	2.959	3.585	.3162	.4136	.2800	.3000
3.9200	14.82	16.83	2.907	3.582	.3134	.4124	.2788	.2999
4.2467	14.61	16.82	2.858	3.581	.3087	.4115	.2776	.2987
4.5733	14.39	16.81	2.806	3.575	.3036	.4109	.2763	.2992
4.9000	14.23	16.80	2.755	3.571	.2982	.4103	.2750	.2996

TABLE 2

$C_D(\xi)$ calculated with and without Richardson Extrapolation
I - without extrapolation
II - with extrapolation

R	Rigid or Gas	N	$C_D(0)$	Difference	% of Higher N
20	Gas	16	.6858		
		20	.6810	.0048	.70
1	Rigid	6	13.16		
		8	13.19	.03	.23
5	Rigid	6	3.720		
		10	3.629	.091	2.5
		14	3.594	.035	.98
10	Rigid	14	2.085		
		16	2.074	.011	.53

TABLE 3

Effect of N on $C_D(0)$

REFERENCES

[1] Abramowitz, M., and Stegun, I., Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, National Bureau of Standards Applied Mathematics Series 55, 1966.

[2] Batchelor, G. K., An Introduction to Fluid Dynamics, Cambridge University Press, London, 1967.

[3] Dennis, S. C. R., and Walker, J. D. A., "Calculation of the Steady Flow Past a Sphere at Low and Moderate Reynolds Numbers," Journal of Fluid Mechanics, 1971, Vol. 48, part 4.

[4] Dennis, S. C. R., and Walker, J. D. A., "Numerical Solutions for the Time-Dependent Flow Past an Impulsively Started Sphere," The Physics of Fluids, 1972, Vol. 15, No. 4.

[5] Isaacson, E., and Keller, H. B., Analysis of Numerical Methods, John Wiley and Sons, Inc., New York, 1966.

[6] Jenson, V. G., "Viscous Flow Round a Sphere at Low Reynolds Numbers (< 40)," Proceedings of the Royal Society of London, 1959, Vol. A249.

[7] Keller, H. B., "Accurate Difference Methods for Nonlinear Two-Point Boundary Value Problems," SIAM Journal of Numerical Analysis, to appear, 1974.

[8] Keller, H. B., Numerical Methods for Two-Point Boundary Value Problems, Ginn-Blaisdell, 1968.

[9] Keller, H. B., and Nieuwstadt, F., "Viscous Flow Past Circular Cylinders," Computers and Fluids, 1973, Vol. 1.

REFERENCES (Cont'd)

[10] Keller, H. B., and Takami, H., "Numerical Studies of Steady Viscous Flow Past Cylinders," Numerical Solutions of Nonlinear Differential Equations (Edited by D. Greenspan), Wiley, New York, 1966.

[11] Landau, L. D., and Lifshitz, E. M., Fluid Mechanics, Pergamon Press, New York, 1966.

[12] Le Clair, B. P., Hamielec, A. E., and Pruppacher, H. R., "A Numerical Study of the Drag on a Sphere at Low and Intermediate Reynolds Numbers," Journal of the Atmospheric Sciences, 1970, Vol. 27.

[13] Levich, V. G., Physicochemical Hydrodynamics, Prentice-Hall, Englewood Cliffs, N.J., 1962.

[14] Milne-Thompson, L. M., Theoretical Hydrodynamics, The Macmillan Co., New York, 1966.

[15] Moore, D. W., "The Boundary Layer on a Spherical Gas Bubble," Journal of Fluid Mechanics, 1959, Vol. 6.

[16] Rimon, Y., and Cheng, S. I., "Numerical Solution of a Uniform Flow over a Sphere at Intermediate Reynolds Numbers," The Physics of Fluids, 1969, Vol. 12, No. 5.

[17] Rotenberg, M., Bivins, R., Metropolis, N., and Wooten, J. K., The 3-j and 6-j Symbols, MIT Press, Cambridge, Mass., 1959.

[18] Talman, J. D., Special Functions, Benjamin, New York, 1968.

REFERENCES (Cont'd)

[19] Underwood, R. L., "Calculation of Incompressible Flow Past a Circular Cylinder at Moderate Reynolds Numbers, Journal of Fluid Mechanics, 1969, Vol. 37.

PART II

NUMERICAL SOLUTION OF SINGULAR ENDPOINT BOUNDARY VALUE PROBLEMS

1. INTRODUCTION

In this part, we consider a system of linear ordinary differential equations on a finite (or infinite) interval with a singularity of the first kind at one endpoint, subject to a linear system of two-point boundary conditions. We determine conditions under which unique solutions exist. We study a numerical method of high order accuracy for these problems. We also treat the same problem with singularities at both endpoints and with a singularity on the interior of the interval.

Previous work on numerical methods for these problems has been done by Gustafsson [2], Natterer [7], Jamet [3], and Shampine [8]. Gustafsson uses a numerical method similar to ours, but treats only scalar problems, not systems, and does not deal at all with existence or uniqueness of solutions. Natterer treats systems, using a projection method and gets $O(h^2[\ln h]^r)$ accuracy. He also treats existence and uniqueness of solutions, but uses unnatural looking boundary conditions, and does not state when the problem has a solution, only when the operator is Fredholm with index zero (not when the operator's inverse exists). Jamet also treats only scalar equations and uses three-point finite difference schemes, which, for a model problem, gives $O(h^{1-\sigma})$ accurate solutions ($\sigma \epsilon (0, 1)$ is a parameter of the problem). Shampine treats a class of nonlinear second order scalar equations, all with the same linear differential

operator. He proves existence and uniqueness of solutions of this equation for certain boundary value problems and the convergence of collocation and finite difference methods.

In Chapter 2, we define what we mean by a solution to the boundary value problem and examine the question of existence and uniqueness. We show that this is equivalent to a certain linear algebra-calculus problem, and we analyze this problem. We give examples of the application of this theory, and also discuss the theory for singularities at both endpoints, and on the interval's interior, and the infinite interval case.

In Chapter 3, we describe a numerical method based on the nature of the fundamental solution matrix near the singularity as described in Coddington and Levinson [1]. We use this to reduce the problem to a regular one away from the singularity. Then we briefly describe a class of finite difference schemes for solving the regular problem and show how Richardson extrapolation may be used to improve the accuracy.

Finally, in Chapter 4, we give numerical examples demonstrating the theory. We employ the box scheme (with and without Richardson extrapolation) and a new one step implicit finite difference scheme of accuracy $O(h^4)$.

2. EXISTENCE AND UNIQUENESS

2.1 The Problem

The system of linear differential equations we consider has a singular point of the first kind, as defined by Coddington and Levinson[1]. This system may be written:

$$y' = A(x)y + b(x), \quad x \in (0,1) \tag{2.1.1}$$

where

$$A(x) = \frac{1}{x} R + \tilde{A}(x) \ . \tag{2.1.2}$$

Here $y(x)$, $b(x)$ are n component vectors, and $A(x)$, R, $\tilde{A}(x)$ are $n \times n$ matrices. R is a constant matrix, $\tilde{A}(x)$ is analytic on $[0,1]$, and $b(x) \in C(0,1]$ ($u(x) \in C^k(a,b]$ means $u(x)$ has k continuous derivatives on $(a,b]$). For any solution $y(x)$ of (2.1.1), we require

$$y(x) \in C^1(0,1] \ . \tag{2.1.3}$$

We also impose a linear system of two-point boundary conditions written as

$$\lim_{x \to 0^+} \{B_o y(x)\} + B_1 y(1) = \beta \ . \tag{2.1.4}$$

Note that we cannot merely write

$$B_o y(0) + B_1 y(1) = \beta \tag{2.1.5}$$

because $y(x)$ is not even necessarily defined at $x = 0$. Notice also that (2.1.4) implies that $\lim_{x \to 0^+} B_o y(x)$ is bounded.

Let Y(x) be a fundamental solution matrix for the homogeneous equation for (2.1.1). That is, Y(x) satisfies

$$Y' = A(x)Y,\ x\epsilon(0,1],\ Y(x_o) = I,\ x_o\epsilon(0,1]. \tag{2.1.6}$$

Then any solution to (2.1.1) can be written

$$y(x) = Y(x)c + y_p(x),\ x\epsilon(0,1] \tag{2.1.7}$$

where c is a constant vector and $y_p(x)$ is any particular solution of (2.1.1). By the variation of parameters method, a formula for $y_p(x)$ can be easily determined as

$$y_p(x) = Y(x)\int_{\epsilon}^{x} Y^{-1}(t)b(t)dt,\ \epsilon \geq 0\ . \tag{2.1.8}$$

For a regular equation (2.1.1) (i.e., for A(x) analytic on $[0,1]$), it can be easily shown that a necessary and sufficient condition for a solution to exist for (2.1.1), (2.1.5) for every β is that $B_oY(0) + B_1Y(1)$ be nonsingular. We now derive similar conditions for the singular problem.

2.2 The Theory

Requiring y(x) to be a unique solution for (2.1.1), (2.1.4) is obviously equivalent to requiring that there exists a unique c in (2.1.7) such that

$$\lim_{x\to 0+}\{B_o(Y(x)c+y_p(x))\} + B_1(Y(1)c+y_p(1)) = \beta\ . \tag{2.2.1}$$

Then we are really faced with the problem of solving a system of the form:

$$\lim_{x\to 0+}\{B(x)c+g(x)\} = \gamma\ , \tag{2.2.2}$$

where B(x) and g(x) may be singular (i.e., $\|B(x)\|$, $\|g(x)\|$ may $\to \infty$ as x approaches zero). It is possible that singularities in the g(x) may exactly cancel singularities in B(x)c for specific choices of c, and that if β (and hence γ) is chosen appropriately, (2.2.2) may have a unique solution c. Among the results we will now prove is that there exists a unique solution c for every choice of γ if and only if $\lim_{x\to 0^+} B(x)$ exists and is nonsingular and $\lim_{x\to 0^+} g(x)$ exists.

We now discuss the existence of solutions c to (2.2.2). Let $b_{ij}(x)$ be the ij^{th} component of B(x), and let c_i, $g_i(x)$, and γ_i be the i^{th} components of c, g(x), and γ, respectively. Then the i^{th} equation of (2.2.2) is just

$$\lim_{x\to 0^+}\left\{\sum_{j=1}^{n} b_{ij}(x)c_j + g_i(x)\right\} = \gamma_i . \qquad (2.2.3)$$

Suppose $g_i(x)$ remains bounded as x approaches zero. Then the sum in (2.2.3) must also be bounded. Now suppose that all $b_{ij}(x)$ are bounded, except one $b_{ij'}(x)$ which is not bounded. Then, obviously, $c_{j'}$ must be zero, and precisely one extra condition is imposed on c by the singularity. If two components, $b_{ij'}(x)$ and $b_{ij''}(x)$, are unbounded (assuming still that $g_i(x)$ is bounded) then either there is one condition between $c_{j'}$ and $c_{j''}$, namely $b_{ij'}(x)c_{j'}+b_{ij''}(x)c_{j''} = O(1)$ as $x \to 0$, or else $c_{j'} = c_{j''} = 0$. Thus, there are either one or two extra conditions imposed on c by the singularity. If $g_i(x)$ is unbounded at $x = 0$, there must be at least one $b_{ij}(x)$ unbounded at $x = 0$, and there will be at least one extra condition imposed on c by the singularity in order to cancel the singularity of $g_i(x)$. Thus we see that any singularity in B(x) or g(x) imposes additional (linear) constraints on the n

unknowns c_i. Hence we make the following definition:

Definition 2.2.4. The number of independent linear constraints imposed on c by the singularity of B(x) and g(x) is called the singularity index, s, of the system (2.2.2).

If we require a solution for every γ in (2.2.2), then we can show that the singularity index, s, is zero, and that B(x) and g(x) are bounded as x approaches zero. To see this, suppose that one component of g(x), say $g_1(x)$, is unbounded at x = 0. Then, as we have seen, this imposes at least one extra condition on c. This condition must be linearly independent from the system (2.2.2) since it does not involve γ. Also, the system (2.2.2) prescribes n linearly independent conditions. This is true since if the left hand side of the i^{th} equation of (2.2.2) is a linear combination of the others, the right hand side, γ_i, must be the same linear combination of the other γ_j; but since we seek a solution for every γ, γ_i is not a fixed linear combination of the other γ_j. Hence, we have at least n + 1 independent linear conditions (n from the equations of (2.2.2), and at least one imposed by the singularity) on the n components of c. Thus, there can be no solution c of (2.2.2) for every γ if any component of g(x) is unbounded as x approaches zero. The same argument can be applied now to show that all the $b_{ij}(x)$ are bounded. So B(x) and g(x) must be bounded as x goes to zero, and the singularity index, s, is zero. Then B(0) and g(0) exist, and B(0) must be nonsingular. We have

$$c = B(0)^{-1}(\gamma - g(0)) . \qquad (2.2.5)$$

In terms of our original problem, we have proved the following:

Theorem 2.2.6

A necessary and sufficient condition for a solution of (2.1.1), (2.1.4) to exist for every β is that $\lim_{x\to 0+} B_o Y(x)$ and $\lim_{x\to 0+} B_o y_p(x)$ exist and that

$$Q \equiv \lim_{x\to 0+} B_o Y(x) + B_1 Y(1) \tag{2.2.7}$$

must be nonsingular. This solution is given by (2.1.7) where

$$c = Q^{-1}(\beta - \lim_{x\to 0+} B_o y_p(x) - B_1 y_p(1)) \quad . \tag{2.2.8}$$

If $n > s > 0$, there can clearly be at most n-s independent equations in (2.2.2) and hence γ lies in an at most n-s dimensional subspace of E^n.

To be more specific, suppose B(x) has the form

$$B(x) = \sum_{\ell=1}^{p} M_\ell q_\ell(x) + M_o(x) \text{ as } x \to 0, \tag{2.2.9}$$

where the M_ℓ are constant n x n matrices, the $q_\ell(x)$ are scalar functions, and

(a) $M_o(x) = O(1)$ as $x \to 0$ (2.2.10)

(b) $q_{\ell'}(x) \neq O(q_\ell(x))$ as $x \to 0$, if $\ell \neq \ell'$

(c) $|q_\ell(x)| \to \infty$ as $x \to 0$.

$$g(x) = \sum_{\ell=1}^{p} g_\ell q_\ell(x) + g_o(x) \quad \text{as } x \to 0 \tag{2.2.11}$$

where the g_ℓ are constant vectors, the $q_\ell(x)$ are the same functions described above, and

$$g_o(x) = O(1) \text{ as } x \to 0 \ . \tag{2.2.12}$$

(Although (2.2.9)-(2.2.12) are little restriction on B(x) and g(x), we will show shortly that for our boundary value problem, these forms occur.) We define the pn x n matrix $\mathcal{M}$ and the pn component vector $\mathcal{G}$ by

$$\mathcal{M} \equiv \begin{bmatrix} M_1 \\ M_2 \\ \cdot \\ \cdot \\ \cdot \\ M_p \end{bmatrix} \ , \qquad \mathcal{G} \equiv \begin{bmatrix} g_1 \\ g_2 \\ \cdot \\ \cdot \\ \cdot \\ g_p \end{bmatrix} \tag{2.2.13}$$

Some of the M_ℓ and g_ℓ may be zero.

Given the above structure, we can now rewrite (2.2.2) as

$$\lim_{x \to 0^+} \left\{ \sum_{\ell=1}^{p} q_\ell(x) \left[M_\ell c + g_\ell \right] + M_o(x)c + g_o(x) \right\} = \gamma \ .$$

From (2.2.10), (2.2.12) we must have

$$\begin{aligned} &\text{(a)} \quad M_\ell c = -g_\ell \ , \quad \ell = 1, \ldots, p \\ &\text{(b)} \quad M_o(0)c = \gamma - g_o(0) \ . \end{aligned} \tag{2.2.14}$$

In matrix form, (2.2.14a) can also be written as

$$\mathcal{M}c = -\mathcal{G}. \qquad (2.2.15)$$

Hence, $\mathcal{G}$ must be in the range of $\mathcal{M}$ ($\mathcal{G} \in R(\mathcal{M})$). Also, the number of independent equations of (2.2.15) is just the singularity index, s, and is also the rank of $\mathcal{M}$. Thus we have

$$s = \text{rk}\ (\mathcal{M}) = \text{rk}\ ([\mathcal{M}, \mathcal{G}]) \qquad (2.2.16)$$

where $[\mathcal{M}, \mathcal{G}]$ is the augmented matrix for the system (2.2.15). Since c must satisfy the (effectively) s conditions (2.2.15), we have that there are exactly n-s free parameters of c to satisfy (2.2.14b), and

$$n-s \geq \text{rk}(M_o(0)) = \text{rk}\ ([M_o(0),\ \gamma - g_o(0)])\ , \qquad (2.2.17)$$

and thus that $\gamma - g_o(0)$ lies in an at most n-s dimensional subspace of E^n (and, therefore, that γ, too, lies in an at most n-s dimensional subspace).

To summarize these results we have the following:

Theorem 2.2.18

The system (2.2.2) with B(x) and g(x) of the forms (2.2.9-12) has a solution c if and only if

(a) $\overline{\mathcal{M}}c = -\overline{\mathcal{G}}$

has a solution where (2.2.19)

(b) $$\overline{\mathcal{M}} = \begin{bmatrix} \mathcal{M} \\ M_o(0) \end{bmatrix}\ ,\quad \overline{\mathcal{G}} = \begin{bmatrix} \mathcal{G} \\ g_o(0) - \gamma \end{bmatrix}\ .$$

An equivalent condition is that

$$rk(\overline{\mathcal{M}}) = rk([\overline{\mathcal{M}}, \overline{\mathcal{G}}]) \quad . \tag{2.2.20}$$

The solution c will be unique if and only if this common rank is n.

Three extreme cases illustrate the development thus far. The first case, s = 0, we have already discussed. $s = rk(\mathcal{M}) = 0$ implies $\mathcal{M} \equiv 0$ and hence $B(x) = M_o(x) = O(1)$ as $x \to 0$ which in turn implies $\mathcal{G} \equiv 0$ and $g(x) = g_o(x) = O(1)$ as $x \to 0$. If a solution is required for every γ, then since $\gamma - g_o(0) \in \mathcal{R}(M_o(0))$, $M_o(0) = B(0)$ must be nonsingular as was shown before. The second case is s = n. The solution c is uniquely determined from the singularity by (2.2.15), and there is exactly one γ, determined by (2.2.14b), consistent with c. Hence, this one γ is the only right hand side for which the original system (2.2.2) has a solution. If γ is not this value, then (2.2.2) has no solution. The third case is $\mathcal{G} \notin \mathcal{R}(\mathcal{M})$, and in this case there is no solution of (2.2.2). This is the case when the singularities of g(x) are such that they cannot be cancelled by the singularities of B(x).

To return to our original problem and show that for the boundary value problem the restrictions on the form of B(x) and g(x) are not unreasonable, we will examine the structure of the fundamental solution matrix Y(x) and the particular solution $y_p(x)$. Coddington and Levinson [1] show that if the constant matrix $\mathcal{R}$ in (2.1.2) has no eigenvalues which differ by a positive integer, the fundamental solution matrix for (2.1.1) has the form

$$Y(x) = P(x)\, x^{R}, \quad x\epsilon(0, \delta] \tag{2.2.21}$$

for some $\delta > 0$, where $P(x)$ (an n x n matrix) is analytic in $(0, \delta]$ and x^{R} is defined as $\exp(R \ln x)$. Then since $\tilde{A}(x)$ in (2.1.2) is analytic, it can be expanded as

$$\tilde{A}(x) = \sum_{k=0}^{\infty} A_k x^k, \quad x\epsilon(0, \delta') \tag{2.2.22}$$

for some $\delta' > 0$, where the A_k are constant matrices. $P(x)$ can also be expanded as

$$P(x) = \sum_{k=0}^{\infty} P_k x^k, \quad x\epsilon(0, \delta') \tag{2.2.23}$$

where the P_k are constant matrices, and the radius of convergence of the series is the same as that for $\tilde{A}(x)$. P_o can be chosen to be the identity matrix, and the P_k are defined recursively in terms of the A_k. The condition that R has no eigenvalues separated by a positive integer is no real restriction since, as they show, the equation (2.1.1) can always be transformed easily and reversibly into one where this condition is satisfied.

If R is a Jordan block, i.e.

$$R = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & 0 \\ & & \ddots & 1 \\ 0 & & & \lambda \end{bmatrix}_{nxn} \tag{2.2.24}$$

then

$$x^R = x^\lambda \begin{bmatrix} 1 & \log x & \frac{\log^2 x}{2!} & \cdots \\ & \ddots & & \\ & & \log x & \\ & & & 1 \end{bmatrix} = x^\lambda Z(x) \qquad (2.2.25)$$

and

$$Y(x) = P(x)x^R = \sum_{k=0}^{\infty} P_k x^k x^R \ . \qquad (2.2.26)$$

If $\mathrm{Re}\lambda > 0$, there are no singular terms here so the singularity index will be 0. If $\mathrm{Re}\lambda < 0$, then writing $k_{max} = [\![-\mathrm{Re}\lambda]\!]$, (2.2.26) becomes

$$Y(x) = \sum_{k=0}^{k_{max}} P_k x^{k+\lambda} Z(x) + \sum_{k=k_{max}+1}^{\infty} P_k x^{k+\lambda} Z(x) \qquad (2.2.27)$$

where the terms in the first sum may be unbounded as $x \to 0$ and those in the second sum are bounded. Thus since $B(x) = B_o Y(x) + B_1 Y(1)$, the $q_\ell(x)$ in (2.2.9) have the form $x^{k+\lambda}(\ln x)^i$, $0 \leq k \leq k_{max}$, $0 \leq i \leq n-1$. Suppose b(x) has the form

$$b(x) = x^\alpha \sum_{k=m_b}^{\infty} b_k x^k \ , \quad x \in (0, \delta') \qquad (2.2.28)$$

where α is a scalar constant, the b_k are constant vectors, and m_b is a (possibly negative) integer. Then from (2.1.8)

$$y_p(x) = \sum_{k=0}^{\infty} \sum_{\ell=m_b}^{\infty} P_k x^{kI+R} \int_\epsilon^x t^{(\alpha+\ell)I-R} \, dt \, b_\ell \ . \qquad (2.2.29)$$

Evaluating this integral shows that if $\alpha + m_b > -1$, then $g(x) = B_o y_p(x)$ has exactly the same type singularities as B(x). If $\alpha + m_b \leq -1$, then there may be additional $q_\ell(x)$ of the form $x^{\alpha+k}(\ln x)^i$, $0 \leq k \leq [\![-\alpha]\!]$, $0 \leq i \leq n$. So the number of singularities, p, is at most

$n([-\alpha]+[-\mathrm{Re}\lambda]+2)$. If R is not a Jordan block or if α is a constant matrix instead of a scalar, the problem is more complicated. But obviously it can still be shown that B(x) and g(x) have the form (2.2.9-12).

In summary we have the following:

Theorem 2.2.30

For the problem (2.1.1), (2.1.4) with b(x) of the form (2.2.28), B(x) and g(x) are of the forms (2.2.9-12). Furthermore, this problem has a solution if and only if the rank condition (2.2.20) holds. This solution is unique if and only if the common rank in (2.2.20) is n.

2.3 Example

We now consider the equation

$$u'' + \frac{\sigma}{x} u' = -x^{1-\sigma}\cos x-(2-\sigma)x^{-\sigma}\sin x, \quad 0<x\leq 1 \qquad (2.3.1)$$

studied by Gustafsson [2]. This same differential operator is also treated by Natterer [7], Jamet [3], and Shampine [8]. Equation (2.3.1) has the general solution

$$u(x) = \alpha_1+\alpha_2 x^{1-\sigma} + x^{1-\sigma}\cos x, \quad 0 < x \leq 1. \qquad (2.3.2)$$

We rewrite (2.3.1) as a two component first order system:

$$\text{(a)} \qquad y' = A(x)y+b(x), \quad x\in(0,1] \qquad (2.3.3)$$

where

(b) $$A(x) = \begin{bmatrix} 0 & 1 \\ 0 & -\frac{\sigma}{x} \end{bmatrix}, \quad b(x) = \begin{bmatrix} 0 \\ -x^{1-\sigma}\cos x-(2-\sigma)x^{-\sigma}\sin x \end{bmatrix}$$

$$y = \begin{bmatrix} u \\ u' \end{bmatrix} . \qquad (2.3.3)$$

In this case,

$$R = \begin{bmatrix} 0 & 0 \\ 0 & -\sigma \end{bmatrix}, \quad \tilde{A}(x) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad (2.3.4)$$

and

$$Y(x) = \begin{bmatrix} 1 & \frac{1}{1-\sigma}x^{1-\sigma} \\ 0 & x^{-\sigma} \end{bmatrix}, \quad y_p(x) = \begin{bmatrix} x^{1-\sigma}\cos x \\ (1-\sigma)x^{-\sigma}\cos x-x^{1-\sigma}\sin x \end{bmatrix} . \qquad (2.3.5)$$

We use the boundary conditions (2.1.4) where

$$B_o \equiv \begin{bmatrix} b_{11}^0 & b_{12}^0 \\ b_{21}^0 & b_{22}^0 \end{bmatrix}, \quad B_1 \equiv \begin{bmatrix} b_{11}^1 & b_{12}^1 \\ b_{21}^1 & b_{22}^1 \end{bmatrix}, \quad \beta \equiv \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} . \qquad (2.3.6)$$

We investigate the conditions on B_o, B_1, and β for which solutions exist for $\sigma > 0$. We have

$$\text{(a) } B(x)=B_o Y(x)+B_1 Y(1) = \begin{bmatrix} b_{11}^0+b_{11}^1 & \frac{1}{1-\sigma}x^{1-\sigma}b_{11}^0+b_{12}^0x^{-\sigma}+\frac{1}{1-\sigma}b_{11}^1+b_{12}^1 \\ b_{21}^0+b_{21}^1 & \frac{1}{1-\sigma}x^{1-\sigma}b_{21}^0+b_{22}^0x^{-\sigma}+\frac{1}{1-\sigma}b_{21}^1+b_{22}^1 \end{bmatrix}$$

(2.3.7)

$$\text{(b) } g(x)=B_o y_p(x)+B_1 y_p(1) = \begin{bmatrix} b_{11}^0x^{1-\sigma}\cos x+b_{12}^0(1-\sigma)x^{-\sigma}\cos x-b_{12}^0x^{1-\sigma}\sin x \\ \quad +b_{11}^1\cos 1+b_{12}^1[(1-\sigma)\cos 1-\sin 1] \\ b_{21}^0x^{1-\sigma}\cos x+b_{22}^0(1-\sigma)x^{-\sigma}\cos x-b_{22}^0x^{1-\sigma}\sin x \\ \quad +b_{21}^1\cos 1+b_{22}^1[(1-\sigma)\cos 1-\sin 1] \end{bmatrix}$$

For $0 < \sigma < 1$, we have

$$B(x) = \begin{bmatrix} 0 & b_{12}^0 \\ 0 & b_{22}^0 \end{bmatrix} x^{-\sigma} + M_o(x) \ ,$$

$$g(x) = \begin{bmatrix} b_{12}^0(1-\sigma) \\ b_{22}^0(1-\sigma) \end{bmatrix} x^{-\sigma} + g_o(x) \ . \qquad (2.3.8)$$

Taking $q_1(x) = x^{-\sigma}$, the singularity conditions (2.2.13) become

$$\mathcal{M}c = \begin{bmatrix} 0 & b_{12}^0 \\ 0 & b_{22}^0 \end{bmatrix} c = -\begin{bmatrix} b_{12}^0(1-\sigma) \\ b_{22}^0(1-\sigma) \end{bmatrix} \quad \text{where } c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \qquad (2.3.9)$$

If a solution is required for every β, we have seen that $s = rk(\mathcal{M}) = 0$, which implies $b_{12}^0 = b_{22}^0 = 0$. In terms of the scalar equation (2.3.1), this means that we cannot include the derivative of the solution at the

origin in the boundary conditions. The other alternative is $s = rk(M) = 1$, and either $b_{12}^0 \neq 0$ or $b_{22}^0 \neq 0$, or both. Then c_2 must be $-(1-\sigma)$, and the other conditions on c (2.2.14b) become

$$M_o(0)c = \begin{bmatrix} b_{11}^0+b_{11}^1 & \frac{1}{1-\sigma}\, b_{11}^1+b_{12}^1 \\ b_{21}^0+b_{21}^1 & \frac{1}{1-\sigma}\, b_{21}^1+b_{22}^1 \end{bmatrix} \begin{bmatrix} c_1 \\ \sigma-1 \end{bmatrix} = \begin{bmatrix} \beta_1 - b_{11}^1\cos 1 - b_{12}^1[(1-\sigma)\cos 1 - \sin 1] \\ \beta_2 - b_{21}^1\cos 1 - b_{22}^1[(1-\sigma)\cos 1 - \sin 1] \end{bmatrix} \tag{2.3.10}$$

or,

$$c_1 \begin{bmatrix} b_{11}^0+b_{11}^1 \\ b_{21}^0+b_{21}^1 \end{bmatrix} = \begin{bmatrix} \beta_1 - (b_{11}^1+(\sigma-1)b_{12}^1)(1-\cos 1)+b_{12}^1\sin 1 \\ \beta_2 - (b_{21}^1+(\sigma-1)b_{22}^1)(1-\cos 1)+b_{22}^1\sin 1 \end{bmatrix}, \tag{2.3.11}$$

which implies that a necessary and sufficient condition for a solution to exist if either of b_{12}^0 or b_{22}^0 is nonzero is that

(a) $$\frac{\beta_1 - (b_{11}^1+(\sigma-1)b_{12}^1)(1-\cos 1)+b_{12}^1\sin 1}{b_{11}^0+b_{11}^1} = \frac{\beta_2 - (b_{21}^1+(\sigma-1)b_{22}^1)(1-\cos 1)+b_{22}^1\sin 1}{b_{21}^0+b_{21}^1},$$

$$\text{if } b_{11}^0+b_{11}^1 \neq 0 \text{ and } b_{21}^0+b_{21}^1 \neq 0 ,$$

(b) $\beta_1 = (b_{11}^1+(\sigma-1)b_{12}^1)(1-\cos 1) - b_{12}^1\sin 1$ if $b_{11}^0+b_{11}^1 = 0$ and $b_{21}^0+b_{21}^1 \neq 0$,

(c) $\beta_2 = (b_{21}^1+(\sigma-1)b_{22}^1)(1-\cos 1) - b_{22}^1\sin 1$ if $b_{11}^0+b_{11}^1 \neq 0$ and $b_{21}^0+b_{21}^1 = 0$,

or

(d) both (b) and (c) if $b_{11}^0+b_{11}^1 = 0$ and $b_{21}^0+b_{21}^1 = 0$. (2.3.12)

In (a, b, c) here, the solution for c (and hence for y(x)) will be unique, but will not in (d). For example, with the scalar equation (2.3.1), we

can pose the following boundary condition (which includes the derivative of the solution at the origin):

$$u(0) = 1$$

$$\lim_{x \to 0} u'(x) + u'(1) = (\sigma-1)(1-\cos 1)-\sin 1 \quad . \tag{2.3.13}$$

This problem has the unique solution

$$u(x) = 1-x^{1-\sigma} + x^{1-\sigma}\cos x \quad . \tag{2.3.14}$$

For $1 < \sigma < 3$, a similar analysis may be carried out, showing that if a solution is required for every β, then B_o must be the zero matrix, and the problem must be posed as an initial value problem from $x = 1$. If a solution is not required for every β, conditions similar to (2.3.12) must hold.

For $\sigma > 3$, the analysis easily yields the fact that B_o must always be the zero matrix, so the only problem which can reasonably be posed is the initial value problem from $x = 1$.

In treating this problem, Gustafsson [2] makes no mention of conditions like (2.3.12), but only treats cases where $u'(0)$ does not appear in the boundary conditions for $0 < \sigma < 1$. For other ranges of σ, our results are also more general than his treatment. Also, Guatafsson requires bounded solutions $u(x)$ on $[0, 1]$, a special case of our theory. In Chapter 4, we treat a numerical example which Gustafsson's treatment does not allow.

2.4 Other Singular Problems

We now consider three extensions of the theory developed

before. The first of these is the case of an equation with a singularity at both ends of the interval $[0,1]$:

$$y' = (\frac{1}{x} R_o + \frac{1}{1-x} R_1 + \tilde{A}(x))\, y+b(x) , \quad x\epsilon(0,1) \tag{2.4.1}$$

where R_o and R_1 are constant n x n matrices, $\tilde{A}(x)$ is analytic on $[0,1]$, and $b(x)\epsilon C(0,1)$. We use the boundary conditions

$$\lim_{x \to 0+} B_o y(x) + \lim_{x \to 1-} B_1 y(1) = \beta \ . \tag{2.4.2}$$

Substituting the form of y(x) (2.1.7) into this boundary condition we have

$$\lim_{x \to 0+} [B_o Y(x)c+B_o\, y_p(x)] + \lim_{x \to 1-} [B_1 Y(x)c+B_1 y_p(x)] = \beta \tag{2.4.3}$$

and since the limits are independent, this implies

$$\begin{aligned} &\text{(a)} \quad B_o Y(x)c+B_o y_p(x) = O(1) \text{ as } x \to 0 \ , \\ &\text{(b)} \quad B_1 Y(x)c+B_1 y_p(x) = O(1) \text{ as } x \to 1 \ . \end{aligned} \tag{2.4.4}$$

Then if we define $\mathcal{M}_o$, $\mathcal{M}_1$, $\mathcal{G}_o$, $\mathcal{G}_1$, $M_o(x)$, $M_1(x)$, $g_o(x)$, and $g_1(x)$ in analogy to (2.2.9-12), we have

$$\begin{aligned} &\text{(a)} \quad B_o Y(x) = \sum_{\ell=1}^{p_o} M_\ell^0 q_\ell^0(x) + M_o(x) \text{ as } x \to 0 \\ &\text{(b)} \quad B_1 Y(x) = \sum_{\ell=1}^{p_1} M_\ell^1 q_\ell^1(x) + M_1(x) \text{ as } x \to 1 \\ &\text{(c)} \quad B_o y_p(x) = \sum_{\ell=1}^{p_o} g_\ell^0 q_\ell^0(x) + g_o(x) \text{ as } x \to 0 \\ &\text{(d)} \quad B_1 y_p(x) = \sum_{\ell=1}^{p_1} g_\ell^1 q_\ell^1(x) + g_1(x) \text{ as } x \to 1 \end{aligned} \tag{2.4.5}$$

(e) $M_o(x), g_o(x) = O(1)$ as $x \to 0$

(f) $M_1(x), g_1(x) = O(1)$ as $x \to 1$

(g) $|q_\ell^0(x)| \to \infty$, $q_\ell^0(x) \neq O(q_{\ell'}^0(x))$ as $x \to 0$ if $\ell \neq \ell'$

(h) $|q_\ell^1(x)| \to \infty$, $q_\ell^1(x) \neq O(q_{\ell'}^1(x))$ as $x \to 1$ if $\ell \neq \ell'$ (2.4.5) Cont'd

M_ℓ^0, M_ℓ^1 are constant matrices, and g_ℓ^0, g_ℓ^1 are constant vectors. As before, we define

$$\mathcal{M}_o = \begin{bmatrix} M_1^0 \\ \vdots \\ M_{p_o}^0 \end{bmatrix}, \quad \mathcal{M}_1 = \begin{bmatrix} M_1^1 \\ \vdots \\ M_{p_1}^1 \end{bmatrix}, \quad \mathcal{G}_o = \begin{bmatrix} g_1^0 \\ \vdots \\ g_{p_o}^0 \end{bmatrix}, \quad \mathcal{G}_1 = \begin{bmatrix} g_1^1 \\ \vdots \\ g_{p_1}^1 \end{bmatrix}. \tag{2.4.6}$$

If b(x) has the form of (2.4.5c) near x = 0 and (2.4.5d) near x = 1, we have

Theorem 2.4.7

The boundary value problem (2.4.1-2) has a solution if and only if the system

(a) $$\mathcal{M}c = \begin{bmatrix} \mathcal{M}_o \\ \mathcal{M}_1 \end{bmatrix} c = - \begin{bmatrix} \mathcal{G}_o \\ \mathcal{G}_1 \end{bmatrix} = - \mathcal{G} \tag{2.4.8}$$

(b) $$[M_o(0) + M_1(1)]\, c = \beta - g_o(0) - g_1(1)$$

has a solution c. An equivalent condition is

$$\operatorname{rk}\begin{bmatrix} M \\ M_o(0)+M_1(1) \end{bmatrix} = \operatorname{rk}\begin{bmatrix} M & , & G \\ M_0(0)+M_1(1), & \beta-g_o(0)-g_1(1) \end{bmatrix} . \tag{2.4.9}$$

This solution is unique if and only if this common rank is n.

Natterer [7] derives matrix rank conditions similar to these, but only for the equation (2.4.1) with $\tilde{A}(x) \equiv 0$ and b(x) bounded on $[0,1]$. In this case our results reduce to his. For $\tilde{A}(x) \not\equiv 0$, he allows $\tilde{A}(x)$ to have singularities which are weaker than $\frac{1}{x}R_o + \frac{1}{1-x}R_1$. The boundary conditions he treats are

$$\lim_{x \to 0} B_o x^{-R_o} y(x) + \lim_{x \to 1} B_1 (1-x)^{-R_1} y(x) = 0 \tag{2.4.10}$$

instead of (2.4.2).

The second extension of the theory is the case of a singularity in the interior of the interval. The equation is the same as our original equation (2.1.1) but on the interval $x \in [-1,0) \cup (0,1]$:

$$y' = (\frac{1}{x} R + \tilde{A}(x))\, y + b(x), \quad x \in [-1,0) \cup (0,1]. \tag{2.4.11}$$

We use a system of boundary conditions at -1 and 1:

$$B_{-1} y(-1) + B_1 y(1) = \beta \ . \tag{2.4.12}$$

Coddington and Levinson's statement [1] about the fundamental solution matrix is that

$$Y(z) = P(z)z^{R} , \quad 0< |z|< \delta , \tag{2.4.13}$$

where R has no eigenvalues separated by a positive integer, and P(z) is analytic (and therefore single-valued) in the punctured disk about the origin. Also, $y_p(x)$ is analytic on $(0,1]$. Hence, the solution $y_p(x)$ can be analytically continued onto the negative real axis from the positive real axis. By a solution to (2.4.11), (2.4.12), we mean any of the functions

$$y(x) = Y(x)c+y_p(x) , \quad x\epsilon [-1,0)\cup(0,1] \tag{2.4.14}$$

which satisfies (2.4.12). In satisfying (2.4.12), we must have

$$[B_{-1}Y(-1) + B_1Y(1)]\ c = \beta-B_{-1}y_p(-1)-B_1y_p(1) \tag{2.4.15}$$

since Y(-1), Y(1), $y_p(-1)$, and $y_p(1)$ exist with no singularities. Then here the singularity index is zero, so that if a solution is required for every β, $B_{-1}Y(-1)+B_1Y(1)$ must be nonsingular. If $B_{-1}Y(-1)+B_1Y(1)$ is singular, then $\beta-B_{-1}y_p(-1)-B_1y_p(1)$ must lie in its range and hence be orthogonal to the nullspace of its conjugate transpose. The numerical treatment of interior singularities is discussed in section 3.1.

The third and final extension to the theory is simply treating the case of a regular differential equation on an infinite interval. We will illustrate this case for a semi-infinite interval, treating the problem

$$\begin{aligned} &\text{(a)} \quad y' = A(x)y+b(x), \quad x\epsilon [0,\infty) \\ &\text{(b)} \quad \lim_{x\to\infty} B_\infty y(x)+B_0y(0) = \beta . \end{aligned} \tag{2.4.16}$$

If we make the change of variable

$$t = \frac{1}{x+1} \quad , \quad \text{or} \quad x = \frac{1}{t} - 1 \tag{2.4.17}$$

we map $x \in [0, \infty)$ into $t \in [1, 0)$. Letting $\hat{y}(t) = y(\frac{1}{t} - 1)$, $\hat{A}(t) = A(\frac{1}{t} - 1)$, $\hat{b}(t) = b(\frac{1}{t} - 1)$, the problem is transformed into

$$\begin{aligned} &\text{(a)} \quad \hat{y}'(t) = -\frac{1}{t^2}\hat{A}(t)\hat{y}(t) + \frac{1}{t^2}\hat{b}(t), \quad t \in [1, 0) \\ &\text{(b)} \quad \lim_{t \to 0^+} B_\infty \hat{y}(t) + B_o \hat{y}(1) = \beta \ . \end{aligned} \tag{2.4.18}$$

Then a necessary and sufficient condition for (2.4.18) to have at most a singularity of the first kind at $t = 0$ is that $A(x)$ is analytic at ∞ and $A(\infty) = 0$. This result is proved by Coddington and Levinson [1]. This statement implies that if

$$A(x) = \frac{1}{x} R + O(\frac{1}{x^2}) \quad \text{as } x \to \infty \ , \tag{2.4.19}$$

(2.4.18) will have exactly a singularity of the first kind if R is not the zero matrix. Hence, for any problem (2.4.16) for which (2.4.19) holds, we can apply our regular theory.

Having discussed existence and uniqueness for problems with a singularity of the first kind, we now turn to their numerical solution. In the next chapter we discuss a numerical method for solving these problems with arbitrary order accuracy, and in Chapter 4 we give numerical examples of this method.

3. NUMERICAL SOLUTION

In this chapter we assume that (2.1.1), (2.1.4) has a unique solution, and we develop a numerical method for computing an approximation to it. This method, unlike the pure finite difference methods of Jamet [3] or the projection method of Natterer [7], has arbitrary order accuracy, depending upon the scheme used to implement the method. The basic idea used here was also used by Gustafsson in [2] for scalar equations.

Essentially, our method consists in using the series form (2.2.21) of the fundamental solution matrix $Y(x)$ to get away from the singular point. With this we define an equivalent boundary value problem on $[\delta, 1]$ instead of $(0, 1]$. The problem defined on $[\delta, 1]$ is regular, and any number of finite difference schemes may be applied to compute its solution. Then knowing an approximation to the fundamental solution matrix on $(0, \delta]$, we can compute an approximation to the particular solution $y_p(x)$ from the variation of parameters formula (2.1.8), and hence an approximation to $y(x)$ on $(0, \delta]$. This combined with the finite difference solution on $[\delta, 1]$ gives a numerical solution on $(0, 1]$. We now examine this method in more detail, especially the formulation of the boundary value problem on $[\delta, 1]$.

3.1 Numerical Method

In order to formulate a boundary value problem on $[\delta, 1]$ (for fixed δ) which is equivalent to our original problem, we use the same differential equation (2.1.1) and need only examine the boundary conditions. From (2.2.9-12), we have

$$B_o Y(x) = \sum_{\ell=1}^{p} M_\ell q_\ell(x) + Q_o(x) \quad \text{as} \quad x \to 0$$

$$B_o y_p(x) = \sum_{\ell=1}^{p} g_\ell q_\ell(x) + r_o(x) \quad \text{as} \quad x \to 0 \qquad (3.1.1)$$

$$Q_o(x),\ r_o(x) = O(1) \quad \text{as } x \to 0 .$$

Then, from the original boundary conditions (2.1.4), we have

$$\text{(a)} \quad \mathcal{M} c = -\mathcal{G}$$

$$\text{(b)} \quad Q_o(0)\, c + B_1 y(1) = \beta - r_o(0) \ . \qquad (3.1.2)$$

Since, from (2.1.7)

$$c = Y^{-1}(\delta)(y(\delta) - y_p(\delta)) \qquad (3.1.3)$$

we have

$$B_{o\delta} y(\delta) + B_{1\delta} y(1) = \beta_\delta \qquad (3.1.4)$$

$$B_{0\delta} \equiv \begin{bmatrix} Q_o(0) \\ \mathcal{M} \end{bmatrix} Y^{-1}(\delta), \quad B_{1\delta} \equiv \begin{bmatrix} B_1 \\ 0 \end{bmatrix}, \quad \beta_\delta \equiv \begin{bmatrix} \beta - r_o(0) \\ -\mathcal{G} \end{bmatrix} + B_{o\delta} y_p(\delta).$$

Then (2.1.1) and (3.1.4) form a regular two-point boundary value problem on $[\delta, 1]$. It is easily seen that this problem has a solution if and only if

$$\text{rk}\left(\begin{bmatrix} Q_o(0) + B_1 Y(1) \\ \mathcal{M} \end{bmatrix} \right) = \text{rk} \left(\begin{bmatrix} Q_o(0) + B_1 Y(1), & \beta - r_o(0) - B_1 y_p(1) \\ \mathcal{M} \ , & -\mathcal{G} \end{bmatrix} \right) \qquad (3.1.5)$$

and that this solution is unique if and only if this common rank is n.

Using this information, it is easy to show that the regular problem has a unique solution if and only if the original, singular problem has a unique solution, which we have assumed. Since these solutions are identical, we have the following:

Theorem 3.1.6

The unique solution to the original problem (2.1.1), (2.1.4) is identical on $[\delta, 1]$ to the unique solution of (2.1.1), (3.1.4) and is identical on $(0, \delta]$ to y(x) given by (2.1.7) where c is given by (3.1.3). Thus the original problem is equivalent to the new regular problem on $[\delta, 1]$ and an "initial value" problem on $(0, \delta]$.

Note that although there are (p+1) n equations in (3.1.4), there are in fact only n independent ones. Thus when actually solving this problem, we only use n independent equations from (3.1.4). If the singularity index of the system is zero, the boundary conditions reduce to

$$[Q_o(0)Y^{-1}(\delta)]\, y(\delta) + B_1 y(1) = \beta - r_o(0) + Q_o(0)Y^{-1}(\delta)y_p(\delta) .$$

Our numerical method is based on this equivalence theorem. We solve the regular problem (2.1.1), (3.1.4) on $[\delta, 1]$ by a finite difference scheme, thus giving a numerical approximation to $y(\delta)$ from which c can be computed by (3.1.3) and then compute an approximation to y(x) on $(0, \delta]$ from (2.1.7).

The only difficulty with this method is that we do not know Y(x) and $y_p(x)$ on $(0, \delta]$. These are necessary not only to compute y(x) on $(0, \delta]$, but also for the computation of the boundary conditions

(3.1.4). However, we can compute approximations $Y^N(x)$ and $y_p^N(x)$ to them by truncating the series for $P(x)$ in (2.2.23), giving

$$Y^N(x) \equiv P^N(x)x^R = \sum_{k=0}^{N} P_k x^k x^R, \quad x\epsilon(0,\delta]$$
$$y_p^N(x) \equiv Y^N(x) \int_{\epsilon}^{x} [Y^N(t)]^{-1} b(t)dt, \quad x\epsilon(0,\delta]. \tag{3.1.7}$$

In reality, it does not matter whether we use these definitions of $Y^N(x)$ and $y_p^N(x)$ or some other, so long as $Y^N(x) \to Y(x)$ and $y_p^N(x) \to y_p(x)$ as $N \to \infty$ uniformly for $x\epsilon(0,\delta]$. Using $Y^N(x)$ and $y_p^N(x)$, we can obtain a "truncated" regular problem with the same differential equation (2.1.1) and boundary conditions given by

$$B_{o\delta}^N y(\delta) + B_{1\delta} y(1) = \beta_\delta^N$$
$$B_{o\delta}^N \equiv \begin{bmatrix} Q_o^N(0) \\ \mathcal{M} \end{bmatrix} [Y^N(\delta)]^{-1}, \quad \beta_\delta^N \equiv \begin{bmatrix} \beta - r_o^N(0) \\ -\mathcal{G} \end{bmatrix} + B_{o\delta}^N y_p^N(\delta) . \tag{3.1.8}$$

Note that we assume here that N is large enough that we have computed the singularities in $B_o Y(x)$ and $B_o y_p(x)$ exactly, so that we have $\mathcal{M}$ and $\mathcal{G}$, rather than some $\mathcal{M}^N$ and $\mathcal{G}^N$.

The differential equation (2.1.1) and boundary conditions (3.1.8) form a "truncated" regular problem. We will show in the course of our error analysis in the next section that for large enough N this problem has a unique solution; call it $y^N(x)$. We wish to compute a numerical approximation to $y^N(x)$ on $[\delta, 1]$.

We impose a net of mesh points on $[\delta, 1]$. Let $\{x_j\}_{j=0}^{J}$ be the set of net points with $x_o = \delta$, $x_J = 1$. (The net may be allowed to

"spill over" the ends of $[\delta, 1]$; that is, there may be some x_j with $j < 0$ such that $0 < x_j < \delta$ and others with $j > J$ such that $x_j > 1$. All we are interested in for the moment is those points in $[\delta, 1]$). Let u_j be the approximation to $y^N(x_j)$ computed by an unspecified scheme. We now define

$$C_n^N \equiv [Y^N(\delta)]^{-1} (u_o - y_p^N(\delta)) \tag{3.1.9}$$

and

$$y_h^N(x) \equiv \begin{cases} u_j, & x = x_j, j = 0, 1, \ldots, J \\ \\ Y^N(x) C_h^N + y_p^N(x), & 0 < x \leq \delta \end{cases} \tag{3.1.10}$$

Then $y_h^N(x)$ is our computed approximation to the solution $y(x)$ of our original problem (2.1.1), (2.1.4).

Any number of schemes may be used to compute the approximations u_j on the net $\{x_j\}_{j=0}^J$. Kreiss [6] has developed a complete theory for a broad class of finite difference schemes, which Gustafsson [2] uses. Keller [4] analyzes the centered Euler or "box" scheme with Richardson extrapolation for solving two-point boundary value problems. White [9] studies a more general class of one step implicit finite difference schemes. In section three of this chapter, we discuss a one step implicit scheme similar to those studied by White. In Chapter 4, we apply that scheme and Keller's box scheme with and without extrapolation to a specific problem.

In the problem with a singularity at each end of the interval, we proceed simultaneously at both ends as we did here for one end.

That is, we use truncated approximations to $Y(x)$ and $y_p(x)$ near both ends and formulate a regular boundary value problem on $[\delta_o, 1-\delta_1]$. Then compute a solution on this interval and use the computed solution to obtain two c's, c_0 and c_1 thus enabling the solution to be computed on $(0, \delta_o]$ and $[1-\delta_1, 1)$. The analysis goes through in a similar fashion to that done before.

For a singularity on the interior of the interval (2.4.11-12) we can impose a mesh on $[-1, -\delta] \cup [\delta, 1]$ and use a matching condition derivable from the fact that $y(x) = Y(x)c + y_p(x)$ must have the same constant c on both sides of the singularity. Then we can compute $y(x)$ on $[-\delta, 0) \cup (0, \delta]$ from the knowledge of c. Another way is to write the problem as a 2n component system on $[\delta, 1]$ as follows: define

$$\text{(a)} \qquad \hat{y}(x) \equiv \begin{bmatrix} y(x) \\ y(-x) \end{bmatrix}, \quad x \in [\delta, 1] \tag{3.1.11}$$

Then $\hat{y}(x)$ satisfies

$$\text{(b)} \qquad \hat{y}'(x) = \hat{A}(x)\hat{y}(x) + \hat{b}(x), \quad x \in [\delta, 1] \tag{3.1.11}$$

$$\text{(c)} \qquad \hat{A}(x) \equiv \begin{bmatrix} A(x) & 0 \\ 0 & -A(-x) \end{bmatrix}, \quad \hat{b}(x) \equiv \begin{bmatrix} b(x) \\ -b(-x) \end{bmatrix}$$

with boundary conditions

$$\text{(d)} \qquad \hat{B}_\delta \hat{y}(\delta) = \hat{\beta}_\delta, \qquad \hat{B}_1 \hat{y}(1) = \beta \tag{3.1.11}$$

$$\text{(e)} \qquad \hat{B}_\delta \equiv [Y^{-1}(\delta) \cdot \; -Y^{-1}(-\delta)], \quad \hat{\beta}_\delta \equiv Y^{-1}(\delta)y_p(\delta) - Y^{-1}(-\delta)y_p(-\delta)$$

$$\hat{B}_1 \equiv [B_1 \quad B_{-1}] .$$

The first of the boundary conditions is just the matching condition alluded to before. Then (3.1.11) is just a two-point boundary value problem to be solved in the usual way. It is easily shown that this problem has a unique solution if and only if the original problem (2.4.11-12) has a unique solution. Once the solution $\hat{y}(x)$ has been obtained, the constant vector c can be calculated thus giving y(x) on $[-\delta, 0) \cup (0, \delta]$. The idea of doubling the system order was discussed in Keller [4] for solving multipoint boundary value problems.

The remaining question is: how close is our numerical solution to the exact solution? We answer this question in the next section.

3.2 Error Analysis

In determining the accuracy of the numerical solution, we must estimate

$$e(x) \equiv \begin{cases} y(x_j)-u_j\,, & x = x_j,\ j = 0, 1, \ldots, J \\ y(x)-y_h^N(x)\,, & x \in (0, \delta] \end{cases} \qquad (3.2.1)$$

The error in estimating $y(x_j)-u_j$ is composed of two parts: i) the error caused by approximating the boundary conditions (3.1.4) of the exact problem by the "truncated" conditions (3.1.8), and ii) the error caused by solving the truncated problem approximately by a finite difference scheme. This second source of error is determined by the choice of δ and the properties of the finite difference scheme. We assume the finite difference scheme has accuracy $O(h^r)$, so that

$$\max_{0 \leq j \leq J} \| y^N(x_j) - u_j \| \leq K_1(\delta) h^r \text{ as } h \to 0 \text{ for fixed } N \text{ and } \delta. \quad (3.2.2)$$

The first source of error, $\| y(x_j) - y^N(x_j) \|$ is more difficult. If we define

$$\Delta_N(\delta) \equiv \max \left\{ \max_{x \in [0, \delta]} \| Y(x) - Y^N(x) \|, \max_{x \in [0, \delta]} \| y_p(x) - y_p^N(x) \| \right\}, \quad (3.2.3)$$

then $\Delta_N(\delta) \to 0$ as $N \to \infty$ for fixed δ. Using this and repeated applications of the triangle inequality, it can be shown that

$$\| B_{o\delta}^N - B_{o\delta} \| \leq K_2(\delta) \Delta_N(\delta)$$

$$\| \beta_\delta^N - \beta_\delta \| \leq K_3(\delta) \Delta_N(\delta) \qquad \text{as } N \to \infty \quad (3.2.4)$$

Since $y(x)$ and $y^N(x)$ satisfy the same differential equation (2.1.1), we have

$$y^N(x) = Y(x) c^N + y_p(x)$$

$$y(x) = Y(x)\, c + y_p(x) \ . \quad (3.2.5)$$

Subtracting the first from the second gives the estimate

$$\| y(x) - y^N(x) \| \leq \| Y(x) \| \ \| c - c^N \| \leq \left[\max_{\delta \leq x \leq 1} \| Y(x) \| \right] \| c - c^N \| \ . \quad (3.2.6)$$

Using the boundary conditions (3.1.4) and (3.1.8) and the inequalities (3.2.4), it can be shown that

$$\|c-c^N\| \leq K_4(\delta)\Delta_N(\delta) \text{ as } N \to \infty \text{ for fixed } \delta . \qquad (3.2.7)$$

Using (3.2.3), (3.2.6), and (3.2.7), we obtain

$$\|y(x)-y^N(x)\| \leq K_5(\delta)\Delta_N(\delta) \text{ as } N \to \infty \qquad (3.2.8)$$

and thus

$$\|e(x_j)\| \leq K_1(\delta)h^r+K_5(\delta)\Delta_N(\delta) \text{ as } h \to 0,\ N \to \infty \text{ for fixed } \delta. \qquad (3.2.9)$$

To estimate e(x) for $0< x< \delta$, we write

$$e(x) = y(x)-y_h^N(x) = Y(x)c+y_p(x) - Y^N(x)c_h^N-y_p^N(x)$$

and hence, using the triangle inequality and (3.2.3), we obtain

$$\|e(x)\| \leq \|Y(x)\,(c-c_h^N)\| + K_6(\delta)\Delta_N(\delta) . \qquad (3.2.10)$$

For sufficiently large N, the Banach lemma implies that $Y^N(\delta)$ is nonsingular and

$$\|Y^{-1}(\delta) - Y^{N^{-1}}(\delta)\| \leq K_7(\delta)\Delta_N(\delta) .$$

Then we can obtain the following estimate:

$$\|c-c_h^N\| \leq K_8(\delta)h^r+K_9(\delta)\Delta_N(\delta) \text{ as } h \to 0,\ N \to \infty, \text{ for fixed } \delta. \qquad (3.2.11)$$

From (3.2.10) we then have

$$\|e(x)\| \leq \|Y(x)\| [K_8(\delta)h^r + K_9(\delta)\Delta_N(\delta)] + K_{10}(\delta)\Delta_N(\delta)$$

$$\text{as } h \to 0,\ N \to \infty, \qquad (3.2.12)$$

for $0 < x < \delta$, with δ fixed.

Since $Y(x)$ may be unbounded as $x \to 0$, this estimate states that the error may also become unbounded. In general, this is all that we can hope to achieve and (3.2.12) is the best estimate that can be given. But if we consider each component of the error, we find that

$$|e_i(x)| \leq \|Y_i(x)\| [K_8(\delta)h^r + K_9(\delta)\Delta_N(\delta)] + K_{10}(\delta)\Delta_N(\delta),\ i = 1, \ldots, n$$

$$\text{as } h \to 0,\ N \to \infty,\ \delta \text{ fixed.} \qquad (3.2.13)$$

Here $e_i(x)$ is the i^{th} component of $e(x)$ and $Y_i(x)$ is the i^{th} row of $Y(x)$. Hence if all components of any row of $Y(x)$ are bounded as $x \to 0$, the error in the corresponding component of the solution will be $O(h^r + \Delta_N(\delta))$ for fixed δ on $[0, \delta]$. This situation will occur if our system is derived from a scalar equation all of whose homogeneous solutions are bounded. Then the error in computing the scalar equation's solution will be $O(h^r + \Delta_N(\delta))$ although the error in the derivatives may blow up. We study such a case in our numerical examples in Chapter 4.

We would like to obtain an a priori bound on $\Delta_N(\delta)$. In Appendix A we show that if $b(x)$ has the form (2.2.28) where α is a scalar constant, then we can construct a series representation of $y_p(x)$ which can be truncated to give

$$\Delta_N(\delta) \leq K_{11}\delta^{N-\lambda} \text{ for a certain } \lambda. \tag{3.2.14}$$

Combining (3.2.9) and (3.2.12), we have in general

$$\|e(x)\| \leq \begin{cases} K_7(\delta)\Delta_N(\delta)+K_1(\delta)h^r, \; x = x_j, \; j = 0, 1, 2, \ldots, J \\ \\ \|Y(x)\| \left[K_8(\delta)h^r+K_9(\delta)\Delta_N(\delta)\right] +K_{10}(\delta)\Delta_N(\delta), \; x\epsilon(0, \delta] \end{cases}$$

$$\text{as } h \to 0, \; N \to \infty, \text{ for } \delta \text{ fixed.} \tag{3.2.15}$$

Improvements on this can be made in specific cases as stated by (3.2.13) and (3.2.14).

In actually performing computations, we first fix δ, then choose an N large enough so that $\Delta_N(\delta)$ is of the same size as h^r for the range of h desired. In the numerical examples in Chapter 4, we examine $\Delta_N(\delta)$ for a specific problem. If we use Richardson extrapolation with our finite difference scheme, we must take account of the increased $O(h^{2r})$ accuracy in choosing N.

3.3 Finite Difference Schemes

We now look at two one step implicit finite difference schemes for solving the "truncated" regular problem (2.1.1), (3.1.8) on $[\delta, 1]$.

The first scheme is the centered Euler or "box" scheme, which has been thoroughly analyzed by Keller [4]. Here we will only state the scheme and the results of Keller's error analysis. We impose the mesh of net points $\{x_j\}_{j=0}^{J}$ on $[\delta, 1]$ with $x_o = \delta$, $x_J = 1$ and variable net spacing h_j defined by

$$h_j \equiv x_j - x_{j-1}, \quad j = 1, \ldots, J. \tag{3.3.1}$$

The box scheme simply approximates the derivative by a centered difference quotient so that the difference scheme is

$$L_h u_j \equiv \frac{u_j - u_{j-1}}{h_j} - A(x_{j-\frac{1}{2}}) \left(\frac{u_j + u_{j-1}}{2}\right) - b(x_{j-\frac{1}{2}}) = 0 , \tag{3.3.2}$$

where $x_{j-\frac{1}{2}} = x_j - h_j/2$. We also impose the boundary condition

$$B_{o\delta}^{N} u_o + B_{1\delta} u_J = \beta_\delta^N . \tag{3.3.3}$$

We can rewrite (3.3.2) as a linear system for the u_j's as follows:

$$-L_j u_{j-1} + R_j u_j = h_j b(x_{j-\frac{1}{2}})$$

$$L_j \equiv I + \frac{h_j}{2} A(x_{j-\frac{1}{2}}) \qquad (3.3.4)$$

$$R_j \equiv I - \frac{h_j}{2} A(x_{j-\frac{1}{2}}) \ .$$

This can be written as a large linear system

$$\begin{bmatrix} B_{o\delta}^N & & & & B_{1\delta} \\ -L_1 & R_1 & & & \\ & -L_2 & R_2 & 0 & \\ & & \ddots & \ddots & \\ & 0 & & & -L_J R_J \end{bmatrix} \begin{bmatrix} u_o \\ u_1 \\ \cdot \\ \cdot \\ \cdot \\ u_J \end{bmatrix} = \begin{bmatrix} \beta_\delta^N \\ h_1 b(x_{\frac{1}{2}}) \\ \cdot \\ \cdot \\ \cdot \\ h_J b(x_{J-\frac{1}{2}}) \end{bmatrix} \qquad (3.3.5)$$

which can be solved for the u_j. Here again we assume that the boundary conditions have been reduced so that they contain exactly n equations. If these boundary conditions are of separated endpoint type, they can be written:

$$D_o u_o = \gamma_o, \ D_1 u_J = \gamma_1$$

where D_o is a p x n matrix, γ_o is a p component vector, D_1 is an (n-p) x n matrix, and γ_1 is an (n-p) component vector. Then the linear system can be written

$$\begin{bmatrix} D_o & & & & \\ -L_1 & R_1 & & & \\ & -L_2 & R_2 & & \\ & & & -L_J R_J & \\ & & & D_1 \end{bmatrix} \begin{bmatrix} u_o \\ u_1 \\ \vdots \\ \\ u_J \end{bmatrix} = \begin{bmatrix} \gamma_o \\ h_1 b(x_{\frac{1}{2}}) \\ \vdots \\ h_J b(x_{J-\frac{1}{2}}) \\ \gamma_1 \end{bmatrix} \qquad (3.3.6)$$

and can be solved by direct factorization methods as Keller [4] points out.

Keller proves that (for $h \equiv \max_j h_j$ sufficiently small) the following error expansion holds:

$$u_j = y^N(x_j) + h_j^2 e_1(x_j) + O(h^4) \quad . \qquad (3.3.7)$$

This allows us to perform an extrapolation process by computing $\bar{u}_j$ as the numerical approximation to $y^N(x_j)$ using step sizes $h_j/2$. Then (3.3.7) implies

$$\bar{u}_{2j} = y^N(x_{2j}) + (\frac{h_j}{2})^2 e_1(x_{2j}) + O(h^4) \, .$$

These two expansions can be combined to yield

$$\frac{4\bar{u}_{2j} - u_j}{3} = y(x_j) + O(h^4)$$

as an $O(h^4)$ approximation to $y(x_j)$. In fact this extrapolation can be continued to give an arbitrary order of accuracy, $O(h^{2r})$.

The second finite difference scheme is arrived at in a slightly different way. First we integrate the differential equation (2.1.1) to obtain

$$y(x_j)-y(x_{j-1}) = \int_{x_{j-1}}^{x_j} [A(t)y(t)+b(t)]\,dt \quad . \tag{3.3.8}$$

We now approximate $y(x)$ on $[x_{j-1}, x_j]$ by a cubic polynomial $H_j(x)$ so that $H_j(x_{j-1}) = y(x_{j-1})$, $H_j(x_j) = y(x_j)$, $H_j'(x_{j-1}) = y'(x_{j-1})$, and $H_j'(x_j) = y'(x_j)$. The coefficients in $H_j(x)$ depend on $y(x_j)$, $y(x_{j-1})$, $y'(x_j)$, and $y'(x_{j-1})$. We use the original differential equation (2.1.1) as an expression for $y'(x)$ to substitute for $y'(x_j)$ and $y'(x_{j-1})$ in $H_j(x)$ so that $H_j(x)$ now depends only on A, b, $y(x_j)$, and $y(x_{j-1})$ and depends on them only linearly. When this approximation to $y(t)$ on $[x_{j-1}, x_j]$ is substituted into the integral above, we get a linear relation between $y(x_j)$ and $y(x_{j-1})$ and thus a one step implicit scheme which can be written

$$\text{(a)} \qquad L_h u_j \equiv \frac{u_j-u_{j-1}}{h_j} - \mathcal{J}_{1,j}\, u_j - \mathcal{J}_{2,j} u_{j-1} - r_j = 0$$

$$\text{(b)} \qquad B_{o\delta}^{N}\, u_o + B_{1\delta} u_J = \beta_\delta^N \, . \tag{3.3.9}$$

This system can also be written as was the box scheme in (3.3.4) and solved by the same methods as described before. Appendix B contains the actual structure of the L_j and R_j in (3.3.4). Note that the $\mathcal{J}_{1,j}$, $\mathcal{J}_{2,j}$, and r_j involve integrals of $A(t)t^k$ where $k = 0, 1, 2,$ or 3 and of $b(t)$. If these integrals can be performed symbolically then there is no problem; if not, then the integrals themselves must be evaluated by a method of the same order accuracy as this overall

scheme.

The truncation error of a finite difference scheme is defined to be

$$\tau_j[y] \equiv L_h y(x_j) - [y'(x_j) - A(x_j)y(x_j) - b(x_j)] \qquad (3.3.10)$$

where y(x) has four continuous derivatives. For our case, assuming y^N has four continuous derivatives on $[\delta, 1]$, a Taylor series expansion gives for the truncation error

$$\tau_j[y^N] = \frac{1}{h_j}\int_{x_{j-1}}^{x_j} A(t)\,\frac{(t-x_{j-1})^2(t-x_j)^2}{24}\,y^{N(iv)}(\xi(t))dt + O(h^5)$$

$$\text{as } h \to 0, \qquad (3.3.11)$$

where $\xi(t)$ is determined by t, x_{j-1}, and x_j. We can thus establish as a bound for the truncation error:

$$\|\tau_j[y^N]\| \leq \max_{\xi\in[x_{j-1},x_j]} \|y^{N(iv)}(\xi)\| \max_{x\in[x_{j-1},x_j]} \|A(x)\| \frac{h_j^4}{720} + O(h^5)\,. \qquad (3.3.12)$$

White [9] proves that if the original problem has a unique solution and the one step scheme is consistent, then it is stable. (Consistency and stability are defined in Keller [5].) Since we have assumed that our problem has a unique solution and since the scheme is obviously consistent (since $\tau_h[y^N] \to 0$ as $h \to 0$), we have that

$$\|y^N(x_j) - u_j\| = O(h^4)\,. \qquad (3.3.13)$$

Extrapolation is not really worthwhile here since the next term in the truncation error is $O(h^5)$. Extrapolation here would only give an

improvement of one in the order of accuracy, whereas for the box scheme, it gives an improvement of two.

In the next chapter, both of these methods are used, and extrapolation is used with the box scheme.

4. NUMERICAL EXAMPLES

We now consider the equation studied in section 2.3

$$u'' + \frac{\sigma}{x} u' = -x^{1-\sigma} \cos x - (2-\sigma) x^{-\sigma} \sin x, \quad x \in (0, 1]. \tag{4.1}$$

We consider two values of σ and one appropriate system of boundary conditions for each value in accordance with section 2.3 as follows:

$$\begin{aligned} &\text{(a)} \quad \text{for } \sigma = 0.5, \ u(0) = 1, \ u(1) = 2 \\ &\text{(b)} \quad \text{for } \sigma = 1.5, \ u(0) = 1, \ u(1) = \cos 1 . \end{aligned} \tag{4.2}$$

From the analysis in section 2.3, it can be shown that (4.2) satisfies the conditions of Theorem 2.2.30 for a unique solution in each case.

We can rewrite the scalar equation as the following system:

$$y' = A(x)y + b(x), \quad x \in (0, 1]$$

$$y(x) \equiv \begin{bmatrix} u(x) \\ \\ u'(x) \end{bmatrix}, \quad A(x) \equiv \begin{bmatrix} 0 & 1 \\ \\ 0 & -\frac{\sigma}{x} \end{bmatrix}, \quad b(x) \equiv \begin{bmatrix} 0 \\ \\ -x^{1-\sigma} \cos x - (2-\sigma) x^{-\sigma} \sin x \end{bmatrix}. \tag{4.3}$$

For boundary conditions, we have

$$\text{(a)} \quad \begin{bmatrix} 1 & 0 \\ \\ 0 & 0 \end{bmatrix} y(0) + \begin{bmatrix} 0 & 0 \\ \\ 1 & 0 \end{bmatrix} y(1) = \beta \tag{4.4}$$

where,

$$\text{(b)} \quad \text{for } \sigma = 0.5, \ \beta = \begin{bmatrix} 1 \\ \\ 2 \end{bmatrix}, \ \text{for } \sigma = 1.5, \ \beta = \begin{bmatrix} 1 \\ \\ \cos 1 \end{bmatrix}.$$

The series solution for Y(x) for (4.3) gives

$$Y(x) = P(x)x^R = \left\{ I + \begin{bmatrix} 0 & 1/(1-\sigma) \\ 0 & 0 \end{bmatrix} x \right\} \begin{bmatrix} 1 & 0 \\ 0 & x^{-\sigma} \end{bmatrix} = \begin{bmatrix} 1 & \frac{1}{1-\sigma} x^{1-\sigma} \\ 0 & x^{-\sigma} \end{bmatrix} \tag{4.5}$$

so that

$$Y^0(x) = \begin{bmatrix} 1 & 0 \\ 0 & x^{-\sigma} \end{bmatrix} \quad , \quad Y^N(x) = Y(x) \text{ , for } N \geqslant 1 \text{ .} \tag{4.6}$$

We also have

$$\text{(a)} \qquad y_p^N(x) = - \sum_{k=0}^{N} \frac{(-1)^k x^{2k+2-\sigma}}{(2k+2)!} \begin{bmatrix} x \\ \\ 2k+3-\sigma \end{bmatrix}$$

$$\text{(b)} \qquad y_p(x) = - \begin{bmatrix} x^{1-\sigma}(1-\cos x) \\ \\ x^{1-\sigma}\sin x + (1-\sigma)x^{-\sigma}(1-\cos x) \end{bmatrix}. \tag{4.7}$$

Since (4.7a) is an alternating series, we have (without exact knowledge of $y_p(x)$):

$$\max_{0 \leqslant x \leqslant \sigma} \| y_p^N(x) - y_p(x) \| \leqslant \frac{(2N+5-\sigma)\delta^{2N+3-\sigma}}{(2N+4)!}$$

and $$\| Y^N(x) - Y(x) \| \leqslant \begin{cases} \frac{1}{1-\sigma} \delta^{1-\sigma} & , \quad N = 0 \ (\sigma < 1) \\ 0 & \quad N \geqslant 1 \end{cases}$$

so that

$$\Delta_N(\delta) \leq \begin{cases} \frac{1}{1-\sigma}\,\delta^{1-\sigma} & , N = 0 \ (\sigma < 1) \\ \frac{(2N+5-\sigma)\delta^{2N+3-\sigma}}{(2N+4)!} & , N \geq 1 . \end{cases} \tag{4.8}$$

For $\sigma = 0.5$, the boundary conditions for the regular problem on $[\delta, 1]$ are

$$B^N_{o\delta}\, y(\delta) + B_{1\delta} y(1) = \beta^N_\delta$$

$$B^N_{o\delta} \equiv \begin{bmatrix} 1 & -\frac{\delta}{1-\sigma} \\ & \\ 0 & 0 \end{bmatrix}, \quad B_{1\delta} \equiv B_1, \quad \beta^N_\delta \equiv \begin{bmatrix} 1+\frac{1}{1-\sigma}\sum_{k=0}^{N} \frac{(-1)^k}{(2k+1)!}\delta^{2k+3-\sigma} \\ \\ 2 \end{bmatrix} \tag{4.9}$$

since the singularity index of the system is zero. For $\sigma = 1.5$, the boundary conditions for the regular problem are

$$B^N_{o\delta}\, y(\delta) + B_{1\delta} y(1) = \beta^N_\delta$$

$$B^N_{o\delta} \equiv \begin{bmatrix} 1 & -\frac{\delta}{1-\sigma} \\ 0 & 0 \\ 0 & \frac{1}{1-\sigma} \\ 0 & 0 \end{bmatrix}, \quad B_{1\delta} \equiv \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\beta_\delta^N \equiv \begin{bmatrix} 1 - \sum_{k=0}^{N} \frac{(-1)^k \delta^{2k+3-\sigma}}{(2k+2)!} + \frac{\delta}{1-\sigma} \sum_{k=0}^{N} \frac{(-1)^k \delta^{2k+2-\sigma}(2k+3-\sigma)}{(2k+2)!} \\ \cos 1 \\ -\frac{1}{1-\sigma} \sum_{k=0}^{N} \frac{(-1)^k \delta^{2k+2-\sigma}(2k+3-\sigma)}{(2k+2)!} \\ 0 \end{bmatrix}$$

These four equations reduce easily to the equivalent system:

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} y(\delta) + \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} y(1) = \begin{bmatrix} 1 - \sum_{k=0}^{N} \frac{(-1)^k \delta^{2k+3-\sigma}}{(2k+2)!} \\ \cos 1 \end{bmatrix} . \quad (4.10)$$

We employ three schemes to solve this regular problem:

I) Keller's box scheme - $O(h^2)$ accurate;

II) Keller's box scheme with Richardson extrapolation - $O(h^4)$ accurate;

III) Hermite polynomial interpolation - $O(h^4)$ accurate.

In using the third scheme, we evaluate the integrals $\int A(t)t^k dt$ in closed form and evaluate $\int b(t)dt$ as an infinite series which we sum to 10^{-12} accuracy. Our computations were performed on an IBM 370/158 using double precision which gives approximately sixteen significant digits accuracy, so that round-off error was not a factor in the computations.

Table (4.11) shows the error $\|c-c_h^N\|$ for $\sigma = 0.5$ for the three schemes for various values of J. Here we have fixed $\delta = 0.1$, $N = 6$ so that $\Delta_N(\delta)$ is much smaller than h^4 so that the error in the

difference scheme is shown.

$\|c-c_h^N\|^*$ $(\sigma = 0.5)$

J	I**	II	III
10	.813(-2)	.218(-3)	.916(-4)
20	.220(-2)	.173(-4)	.808(-5)
40	.562(-3)	.117(-5)	.565(-6)
60	.251(-3)	.235(-6)	.114(-6)
80	.141(-3)	.748(-7)	.365(-7)

(4.11)

* $\|a\| = \max \{|a_1|, |a_2|\}$

** (.123(-4) means $.123 \times 10^{-4}$).

Table (4.12) shows the same data for $\sigma = 1.5$.

$\|c-c_h^N\|$ $(\sigma = 1.5)$

J	I	II	III
20	.929(-4)	.785(-6)	.577(-6)
40	.238(-4)	.511(-7)	.374(-7)
60	.106(-4)	.102(-7)	.744(-8)
80	.599(-5)	.323(-8)	.236(-8)

(4.12)

The error in $\|c-c_h^N\|$ from using N = 0, 1, 2, and 3 terms in the truncated series is shown in Table (4.13) for $\sigma = 0.5$, $\delta = 0.1$, J = 80, and schemes II and III.

$\|c-c_h^N\|$ $(\sigma = 0.5)$

N	II	III
0	.134(-4)	.134(-4)
1	.812(-7)	.489(-7)
2	.748(-7)	.365(-7)
3	.748(-7)	.365(-7)

(4.13)

The error incurred by varying δ is reflected in the finite difference error and the error in truncating the series. Table (4.14) shows the error $\|c-c_h^N\|$ for various δ for $\sigma = 0.5$, $J = 80$, $N = 6$ (for which the effects of truncating the series are insignificant) for scheme I.

	$\delta = 0.05$	$\delta = 0.1$	$\delta = 0.2$	$\delta = 0.4$
$\|c-c_h^N\|$	.636(-3)	.141(-3)	.258(-4)	.279(-5)

(4.14)

In section 3.2 we pointed out that for a system whose fundamental solution matrix has a bounded row as $x \to 0$, the error in the corresponding component of the solution will remain bounded (in fact, will be $O(h^r+\Delta_N(\delta))$. For our problem here with $0 < \sigma < 1$, we have the first row of $Y(x)$ bounded as $x \to 0$ while the second row blows up like $x^{-\sigma}$. For $1< \sigma < 3$, both rows blow up. Table (4.15) shows the error in the solution, $|u(x)-y_h^{N(1)}(x)|$, and derivative, $|u'(x)-y_h^{N(2)}(x)|$ as $x \to 0$ for $\sigma = 0.5$, as computed by scheme I with $N = 6$, $J = 80$, and $\delta = 0.1$. Table (4.16) shows the same data as computed for $\sigma = 1.5$. The error in the solution remains bounded for $\sigma = 0.5$, but

blows up in the derivative (and for the solution and derivative for $\sigma = 1.5$) as predicted.

$\sigma = 0.5$

x	$\lvert u(x)-y_h^{N(1)}(x)\rvert$	$\lvert u'(x)-y_h^{N(2)}(x)\rvert$
.1	.894(-4)	.447(-3)
.075	.775(-4)	.515(-3)
.051	.635(-4)	.629(-3)
.016	.356(-4)	.112(-2)
.006	.218(-4)	.183(-2)
.001	.894(-5)	.447(-2)

(4.15)

$\sigma = 1.5$

x	$\lvert u(x)-y_h^{N(1)}(x)\rvert$	$\lvert u'(x)-y_h^{N(2)}(x)\rvert$
.1	.139(-15)	.300(-4)
.075	.915(-6)	.459(-4)
.051	.244(-5)	.835(-4)
.016	.906(-5)	.475(-3)
.006	.186(-4)	.206(-2)
.001	.539(-4)	.300(-1)

(4.16)

Provided N is sufficiently large, the error on $[\delta, 1]$ decreases as δ increases. But as δ is increased, N must also be increased to retain the same accuracy in $Y^N(\delta)$ and $y_p^N(\delta)$. Hence, in order to achieve maximum efficiency, we must balance our choices of δ, N,

and h in keeping with the order of accuracy of our finite difference scheme. For our sample problem here, this was not difficult, but for harder problems, it would be of more concern.

The computations were timed to get a measure of the relative efficiency of the two $O(h^4)$ schemes (II and III). Table (4.17) shows these times for various values of J. The box scheme with Richardson extrapolation is more efficient for our current problem. It was also much easier to analyze and program. Also the box scheme permits going to higher order accuracy very easily and efficiently whereas the Hermite scheme does not.

J	scheme II time*	scheme III time*
20	131	167
40	242	336
60	350	483
80	476	633

*all times given in milliseconds (10^{-3} seconds) (4.17)

In conclusion, we have developed a theory to tell us when a singular endpoint boundary value problem has a unique solution. We have also developed a method of arbitrarily high order of accuracy. We have found this method to be efficient and accurate and easily implemented.

APPENDIX A

In obtaining an a priori bound on $\Delta_N(\delta)$, we must estimate $\|Y(x)-Y^N(x)\|$ and $\|y_p(x)-y_p^N(x)\|$. The first of these is easily estimated since the definitions of $Y(x)$ and $Y^N(x)$ (2.2.26), (3.1.7) give

$$Y(x)-Y^N(x)=\left[\sum_{K=N+1}^{\infty} P_k x^k\right] x^R = \left[\sum_{k=0}^{\infty} P_{k+N+1} x^k\right] x^{(N+1)I+R}$$

so that

$$\max_{x\in[0,\delta]} \|Y(x)-Y^N(x)\| \leq K_1 \delta^{N+1-\gamma} \tag{A.1}$$

for fixed δ and N sufficiently large. Here $\gamma = -\mathrm{Re}(\lambda) > 0$ where λ is the eigenvalue of R with most negative real part. (If all eigenvalues of R are nonnegative, we take $\gamma = 0$.)

The estimation of $\|y_p(x)-y_p^N(x)\|$ is more difficult. We assume that b(x) in (2.1.1) has the form

$$b(x) = x^{\alpha} \sum_{k=0}^{\infty} x^k b_k \tag{A.2}$$

where α is a scalar constant. We take $y_p(x)$ to be the following:

$$y_p(x) = Y(x) \int^x Y^{-1}(t)b(t)dt \ . \tag{A.3}$$

The indefinite integral is not useful unless it can be computed in "closed form." But for our case, it can be evaluated in the following manner. Suppose $\alpha + k$ is not an eigenvalue of R for any integer $k \geq 1$. (We will show how this restriction may be removed later.) Then we have

$$\text{(a)} \qquad y_p(x) = Y(x) \sum_{k=0}^{\infty} \left[(k+\alpha+1)I-R\right]^{-1} x^{(k+\alpha+1)I-R} c_k$$

$$\text{(b)} \qquad c_k \equiv \sum_{\ell=0}^{k} Q_\ell b_{k-\ell} \tag{A.4}$$

$$\text{(c)} \qquad P^{-1}(x) = \sum_{k=0}^{\infty} Q_k x^k \ .$$

We now take (A. 4) as our definition of $y_p(x)$ and define

$$y_p^N(x) = Y^N(x) \sum_{k=0}^{N} \left[(k+\alpha+1)I-R\right]^{-1} x^{(k+\alpha+1)I-R} c_k \ . \tag{A.5}$$

For $k \leq N$, the c_k involve only b_ℓ and Q_ℓ for $0 \leq \ell \leq N$. Then it is easy to show that

$$\max_{x \in [0,\, \delta]} \left\| y_p(x) - y_p^N(x) \right\| \leq K_2 \delta^{N+2+\alpha} \tag{A.6}$$

If $\alpha + k'$ is an eigenvalue of R then the $k'-1$ term of the series (A. 4a) will not just be a simple matrix power of x but will also involve log terms. (In Appendix C we show the general form of $\int^x t^A dt$ where A is a constant matrix.) Hence in order to compute these singular terms exactly, we must have $N \geq k'$. In section 2. 2 we assumed that R has no eigenvalues separated by a positive integer so that there is only one k' for which $\alpha + k'$ is an eigenvalue of R. Hence for N sufficiently large, we have

$$\Delta_N(\delta) \leq K \max \left\{\delta^{N+1-\gamma},\ \delta^{N+2+\alpha}\right\}, \quad \text{or}$$

$$\Delta_N(\delta) \leq K\, \delta^{N+\lambda}, \quad \lambda = \min \left\{1-\gamma,\ \alpha+2\right\} . \tag{A.7}$$

APPENDIX B

In the interpolation method of section 3.3, we approximate the function $y(x)$ on $[x_{j-1}, x_j]$ by a cubic polynomial $H_j(x)$ so that

$$H_j(x_j) = y(x_j),\ H_j'(x_j) = y'(x_j)$$

$$H_j(x_{j-1}) = y(x_{j-1}),\ H_j'(x_{j-1}) = y'(x_{j-1})\ . \tag{B.1}$$

To do this, we define four cubic polynomials $h_i(x)$ on $[x_{j-1}, x_j]$ so that

$$\begin{aligned}
&h_1(x_{j-1}) = 1,\ h_1'(x_{j-1}) = h_1(x_j) = h_1'(x_j) = 0\\
&h_2'(x_{j-1}) = 1, h_2(x_{j-1}) = h_2(x_j) = h_2'(x_j) = 0\\
&h_3'(x_j) = 1, h_3(x_{j-1}) = h_3'(x_{j-1}) = h_3(x_j) = 0\\
&h_4(x_j) = 1, h_4(x_{j-1}) = h_4'(x_{j-1}) = h_4'(x_j) = 0\ .
\end{aligned} \tag{B.2}$$

These polynomials turn out to be

$$\begin{aligned}
h_1(x) &= 1 - \frac{3}{h^2}(x-x_{j-1})^2 + \frac{2}{h^3}(x-x_{j-1})^3\\
h_2(x) &= (x-x_j) - \frac{2}{h}(x-x_{j-1})^2 + \frac{1}{h^2}(x-x_{j-1})^3\\
h_3(x) &= -\frac{1}{h}(x-x_{j-1})^2 + \frac{1}{h^2}(x-x_{j-1})^3\\
h_4(x) &= \frac{3}{h^2}(x-x_{j-1})^2 - \frac{2}{h^3}(x-x_{j-1})^3\ .
\end{aligned} \tag{B.3}$$

The scheme is then

$$-L_j u_{j-1} + R_j u_j = h_j r_j$$

$$L_j \equiv I + \int_{x_{j-1}}^{x_j} A(t)h_2(t)dt\, A(x_{j-1}) + \int_{x_{j-1}}^{x_j} A(t)h_1(t)dt$$

$$R_j \equiv I - \int_{x_{j-1}}^{x_j} A(t)h_3(t)dt\, A(x_j) - \int_{x_{j-1}}^{x_j} A(t)h_4(t)dt \qquad \text{(B. 4)}$$

$$h_j r_j \equiv \int_{x_{j-1}}^{x_j} A(t)h_2(t)dt\, b(x_{j-1}) + \int_{x_{j-1}}^{x_j} A(t)h_3(t)dt\, b(x_j) + \int_{x_{j-1}}^{x_j} b(t)dt.$$

APPENDIX C

We wish to evaluate the indefinite integral $\int t^A dt$ for A an n x n constant matrix. We define

$$t^A \equiv \exp\{A \ln t\} \equiv \sum_{k=0}^{\infty} \frac{(A \ln t)^k}{k!} \tag{C.1}$$

From this we can show easily that

$$\frac{d}{dt}(t^A) = At^{A-I} \ . \tag{C.2}$$

Here I is the n x n identity matrix. Hence, if I + A is nonsingular, we have

$$\int t^A \, dt = [A+I]^{-1} \, t^{A+I} \tag{C.3}$$

If I + A is singular, then -1 is an eigenvalue of A, and there is a matrix T such that TAT^{-1} is in Jordan normal form:

$$\text{(a)} \qquad TAT^{-1} = J \equiv \begin{bmatrix} J_s & 0 \\ 0 & J_{n-s} \end{bmatrix} \tag{C.4}$$

$$\text{(b)} \qquad J_s = \begin{bmatrix} -1 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ 0 & & & -1 \end{bmatrix}$$

Here s is the multiplicity of the eigenvalue -1, J_s is an s x s matrix, and J_{n-s} is an (n-s) x (n-s) matrix. Then it is obvious that

$$t^A = T^{-1} \, t^J \, T = T^{-1} \begin{bmatrix} t^{J_s} & \\ & t^{J_{n-s}} \end{bmatrix} T \tag{C.5}$$

and

$$\int t^{A}dt = T^{-1}\begin{bmatrix} \int t^{J_s}dt & & \\ & & \\ & & [I+J_{n-s}]^{-1}\, t^{I+J_{n-s}} \end{bmatrix} T \quad . \qquad \text{(C. 6)}$$

It is well known that

$$t^{J_s} = t^{-1}\begin{bmatrix} 1 & \ln t & \frac{(\ln t)^2}{2!} & \cdots \\ & 1 & \ln t & \ddots \\ & & \ddots & \ddots \\ 0 & & & 1 \end{bmatrix} \qquad \text{(C. 7)}$$

so $$\int_{x}^{x} t^{J_s}dt = \int_{x}^{x} t^{-1}\begin{bmatrix} 1 & \ln t & & \\ & \ddots & \ddots & \\ & & \ddots & \\ & & & 1 \end{bmatrix}dt = \int^{\ln x} \begin{bmatrix} 1 & u & \frac{u^2}{2!} & \cdots \\ & \ddots & \ddots & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} du$$

$$\int t^{J_s}\, dt = \begin{bmatrix} \ln t & \frac{(\ln t)^2}{2!} & \frac{(\ln t)^3}{3!} & \cdots \\ & \ln t & \ddots & \\ & & \ddots & \\ & & & \ln t \end{bmatrix} . \qquad \text{(C. 8)}$$

Combining (C. 3), (C. 6), and (C. 8) we have $\int t^{A}\, dt$.

REFERENCES

[1] Coddington, E. A., and Levinson, N., Theory of Ordinary Differential Equations, McGraw-Hill, New York, 1955.

[2] Gustafsson, B., "A Numerical Method for Solving Singular Boundary Value Problems," Uppsala University, Dept. of Computer Sciences, Report No. 45, March 1973.

[3] Jamet, P., "On the Convergence of Finite-Difference Approximations to One-Dimensional Singular Boundary-Value Problems," Numerische Mathematik, 1970, Vol. 14, pp. 355-378.

[4] Keller, H. B., "Accurate Difference Methods for Linear Ordinary Differential Systems Subject to Linear Constraints," SIAM Journal on Numerical Analysis, 1969, Vol. 6, pp. 8-30.

[5] Keller, H. B., Numerical Methods for Two-Point Boundary-Value Problems, Blaisdell Publishing Co., Waltham, Mass., 1968.

[6] Kreiss, H. O., "Difference Approximations for Ordinary Differential Equations," Uppsala University, Dept. of Computer Sciences, Report No. 35, September 1971.

[7] Natterer, F., "A Generalized Spline Method for Singular Boundary Value Problems of Ordinary Differential Equations," Linear Algebra and Its Applications, 1973, Vol. 7, pp. 189-216.

[8] Shampine, L. F., "Numerical Methods for Singular Boundary Value Problems," SIAM Journal on Numerical Analysis (to appear).

REFERENCES (Cont'd)

[9] White, A. B., Ph.D. Thesis, Dept. of Applied Mathematics, California Institute of Technology, Pasadena, California, 1974.