ANALYSIS OF COMPUTATIONAL TECHNIQUES

FOR CIRCUIT THEORY

by

H. Haneda

Memorandum No. ERL-323

February 16, 1972

ANALYSIS OF COMPUTATIONAL TECHNIQUES FOR CIRCUIT THEORY

by

H. Haneda

Memorandum No. ERL-M323

February 16, 1972

ELECTRONICS RESEARCH LABORATORY

College of Engineering
University of California, Berkeley
94720

ANALYSIS OF COMPUTATIONAL TECHNIQUES FOR CIRCUIT THEORY

by

Hiromasa Haneda

ABSTRACT

This technical memorandum uses the measure of a matrix to unify and generalize the analysis of some numerical techniques useful in circuit theory.

An existence and uniqueness theorem for D. C. operating point is given given; a convergence region for the Newton-Raphson method is determined and its quadratic convergence is established. The effect of local round-off error is also discussed.

An estimate for the upper and lower bounds on the solutions of an important class of ordinary differential equations is given. This estimate is sharper than that obtained by using norms.

An estimate is given for the bounds on computed solutions of ordinary differential equations obtained by the backward Euler method and its modifications. A bound on the accumulated truncation error incurred by the backward Euler method is also given.

The effect of the step size in the implicit equation obtained by the backward Euler method on the existence and uniqueness of the solution as well as on the convergence of the Newton-Raphson method is discussed.

The research was sponsored by The National Science Foundation, Grant GK-10656X1 and The Joint Services Electronics Program, Contract F44620-71-C-0087. The results will be presented at the ISCT'72 in April, 1972.

ACKNOWLEDGEMENT

This technical memorandum describes the essential contents of the thesis written by the author under the guidance of Professor C. A. Desoer The constructive comments of Professor L. O. Chua are gratefully acknowledged and so is the teaching of Professor W. Kahan. Professor H. Lewy was the external reader for the thesis.

CONTENTS

CHAPTER 0.

INTRODUCTION

In this introductory chapter, we give an overall view of this thesis. Second, we give a few motivating examples from several different engineering fields. Third, we describe the contributions of this thesis as they are related to previous work. Finally we list notational conventions used in later chapters.

1. Introduction

This thesis is concerned with the analysis of some numerical techniques useful in circuit theory. The principal motivation of this thesis is to illuminate and give insight into a number of problems that are encountered in the implementation of computer aided design methods for electrical circuits in particular. The main thread throughout this thesis is the use of the measure of a matrix. Thanks to this approach a number of previous results are generalized and clarified (see Sec. 3 below). The organization of the thesis is as follows:

In Chapter I we define the measure of a matrix which was discussed by Dahlquist [1] and was used to investigate the stability of ordinary differential equations by Dahlquist [1] and Coppel [2]. We prove its properties in detail, some of which are new. We give interpretations of the measure of a matrix in terms of well-known classes of matrices. For the record we state a first-order implicit integration formula and the Newton-Raphson method to make our discussion precise in later chapters.

In Chapter II, we develop properties of D. C. equations which are encountered in analyzing electric circuits for their D. C. operating points and also in the use of implicit integration methods for computing their transient response. We prove an existence and uniqueness theorem; determine a guaranteed convergence region and the rate of convergence of the Newton-Raphson method for both the infinite and finite precision arithmetic computations.

In Chapter III, we estimate the upper and lower bounds on the solution of ordinary differential equations (O.D.E.'s):

$$\begin{cases} \dot{x} = f(x,t) + u(t) \\ x(0) = x_0 \end{cases} \tag{0.1}$$

where $x(t)$ and $u(t)$ are d dimensional vector for each time $t \geq 0$ and $f(\cdot,\cdot)$ is a function from $\mathbb{R}^d x \mathbb{R}_+$ into $\mathbb{R}^d$. These estimates are essentially due to Dahlquist [1] and Coppel [2], but theorems are stated in a more convenient and slightly extended manner. In view of our purposes we give these estimates for stable cases only. In electrical networks as well as chemical kinetics, the derivative(Jacobian) $D_1 f(x,t)$ of $f(\cdot,t)$ in (0.1) often has very widely spread eigenvalues for each $x(t) \in \mathbb{R}^d$, for each $t \in \mathbb{R}_+$, Sandberg & Shichman [17], Sandberg [15], Desoer & Shensa [21], Chua & Alexander [22], Gear [20].

Such O. D. E.'s are called stiff differential equations. Roughly speaking, the upper bound that we obtain is determined by the slowest time constant and the lower bound, by the fastest time constant.

In Chapter IV, we estimate bounds on computed solutions of O. D. E.'s when infinite precision arithmetic is used. We also estimate bounds on errors between the computed sequence by the backward Euler method and those obtained by its modifications, and a bound on the accumulated truncation error incurred by the backward Euler method. For the computation of the solution of stiff differential equations by standard explicit methods we are forced to choose very small step sizes to avoid numerical instability; the accumulation of local round-off errors and the computation time will become intolerable, [17], [15], [20]. A class of methods to allow dramatic step-size increases is that of implicit methods and its modifications, [17], [15], [20],[19]. In Chapter IV, we consider the backward(implicit) Euler method and its modifications. We estimate for any given step size bounds on computed solutions and errors incurred; show desirable properties of the effect of the initial error, the input error, the local truncation error and the step sizes. Finally, we extend and relate the results of Ch.II to the implicit equation obtained by the backward Euler method. The effect of the step size on the existence and uniqueness of the D. C. solution as well as on the convergence region of the Newton-Raphson method is evident from our formulas.

Some of the results in this thesis are being presented at 1972 IEEE International Symposium on Circuit Theory, [29].

2. Motivating Examples

The O. D. E.'s of the form (0.1) are encountered in many engineering problems. Motivating examples are given in important classes of O. D. E.'s of the form (0.1).

Class ND First, we show examples in a class of O. D. E.'s of the form (0.1) satisfying the following condition: there exists a dxd constant nonsingular matrix P such that $-PD_1 f(x,t)P^{-1}$ is uniformly positive definite, more precisely there exists a nonsingular matrix $P \in \mathbb{R}^{d\times d}$ and a positive constant $m > 0$ such that

$$\langle y, -PD_1 f(x,t)P^{-1}y \rangle \geq m|y|^2 \quad \text{for all } x \in \mathbb{R}^d, \text{ for all}$$

$$t \in \mathbb{R}_+, \text{ for all } y \in \mathbb{R}^d. \tag{0.2}$$

Example 0-1. RLC network (Fig.1).

Consider an RLC network consisting of independent sources, m linear time-invariant capacitors and n linear time-invariant inductors, (m+n) nonlinear resistors, and a linear time-invariant resistive (m+n)-port. We assume:

(i) m nonlinear voltage-controlled resistors are connected parallel to the m capacitors, and the n nonlinear current-controlled resistors are connected in series to the n inductors.

(ii) The m independent current sources are connected parallel to the m capacitors, and the n independent voltage sources are connected in series to the n inductors.

(iii) The (m+n)-port has a hybrid matrix H such that

$$\begin{bmatrix} i \\ v \end{bmatrix} = -H \begin{bmatrix} v_C \\ i_L \end{bmatrix} \tag{0.3}$$

where $i = (i_1, \cdots, i_m)^T$, $i_L = (i_{m+1}, \cdots, i_{m+n})^T$,
$v_C = (v_1, \cdots, v_m)^T$, $v = (v_{m+1}, \cdots, v_{m+n})^T$.

From Fig.1, we obtain:

$$\begin{cases} C\dot{v}_C = i - \hat{i}(v_C,t) + i_s \\ L\dot{i}_L = v - \hat{v}(i_L,t) + v_s \end{cases} \tag{0.4}$$

where $C = \mathrm{diag}(C_1, \cdots, C_m)$ with $C_i > 0$, the capacitance of the i-th capacitor; $L = \mathrm{diag}(L_1, \cdots, L_n)$ with $L_i > 0$, the inductance of the i-th inductor; $v_C \mapsto \hat{i}(v_C,t)$ represents the characteristics at time t of the m voltage-controlled resistors; $i_L \mapsto \hat{v}(i_L,t)$ represents the characteristics at time t of the n current-controlled resistors; i_s represents the m independent current sources; and v_s represents the n independent voltage sources. From (0.3) and (0.4), we obtain:

$$\left[\begin{array}{c|c} C & 0 \\ \hline 0 & L \end{array}\right] \begin{bmatrix} \dot{v}_C \\ \dot{i}_L \end{bmatrix} = -H \begin{bmatrix} v_C \\ i_L \end{bmatrix} - \begin{bmatrix} \hat{i}(v_C,t) \\ \hat{v}(i_L,t) \end{bmatrix} + \begin{bmatrix} i_s(t) \\ v_s(t) \end{bmatrix} . \tag{0.5}$$

Note that eq.(0.5) is not restricted to have its sources located as in Fig.1; indeed if there were sources inside the (m+n)-port we could extract the Norton equivalent current sources and the Thevenin equivalent voltage sources. Equation (0.5) is of the form (0.1) and

$$-D_1 f(x,t) = \left[\begin{array}{c|c} C & 0 \\ \hline 0 & L \end{array}\right]^{-1} \left\{ H + \left[\begin{array}{c|c} D_1 i(v_C,t) & 0 \\ \hline 0 & D_1 v(i_L,t) \end{array}\right] \right\}. \qquad (0.6)$$

Furthermore we assume (iv) $D_1\hat{i}(v_C,t)$ and $D_1\hat{v}(i_L,t)$ are both positive semidefinite for all $v_C \in \mathbb{R}^m$, for all $i_L \in \mathbb{R}^n$, for all $t \in \mathbb{R}_+$; and (v) H is positive definite (not necessarily symmetric).

Observe that $\left[\begin{array}{c|c} C & 0 \\ \hline 0 & L \end{array}\right]^{-1}$ is diagonal and positive, and that

$\left\{ H + \left[\begin{array}{c|c} D_1\hat{i}(v_C,t) & 0 \\ \hline 0 & D_1\hat{v}(i_L,t) \end{array}\right] \right\}$ is uniformly positive definite

in (v_C, i_L) and in t. By Lemma A-1 (Appendix), the condition of class ND is satisfied.

Example 0-2. Three-phase synchronous machine model, [24].

The next O. D. E. (0.7):

$$\frac{d}{dt}\left[L(t)\cdot i(t)\right] = -R\cdot i(t) + v(t) \qquad (0.7)$$

represents a model of a three-phase synchronous machine, where $i(t) \in \mathbb{R}^4$ and represents the currents through the three armature windings and through the field winding; $v(t) \in \mathbb{R}^4$ and represents the four terminal voltages to the ground; R is a 4x4 positive diagonal constant matrix which represents the resistance of the windings; and L(t) is a 4x4 inductance matrix which is time-varying. We assume that L(t) is symmetric for each time t; uniformly positive definite, i.e., there exists a positive constant $\varepsilon > 0$ such that

$$\langle y, L(t)y \rangle \geq \varepsilon |y|^2, \text{ for all } t \in \mathbb{R}_+, \text{ for all } y \in \mathbb{R}^4; \quad (0.8)$$

and L(t) is bounded on $\mathbb{R}_+$ (It is usually assumed periodic). Choose $\phi(t) = L(t) \cdot i(t)$ and v(t) as a state variable and input, respectively. Then, from (0.7) we obtain:

$$\dot{\phi}(t) = -R \cdot L^{-1}(t)\phi + v(t). \quad (0.9)$$

Observe that eq.(0.9) is of the form (0.1) and that the condition (0.2) is satisfied by choosing $P = R^{\frac{1}{2}}$.

Class NCSD Second, we show examples in a class of O. D. E.'s of the form (0.1) satisfying the following condition: there exists a dxd real constant nonsingular matrix P such that $-PD_1f(x,t)P^{-1}$ is uniformly column-sum dominant, i.e., there exists a nonsingular matrix $P \in \mathbb{R}^{dxd}$ and a positive constant $m > 0$ such that

$$a_{jj}(x,t) - \sum_{\substack{i=1 \\ (i \neq j)}}^{d} |a_{ij}(x,t)| \geq m, \text{ for all } x \in \mathbb{R}^d, \text{ for all}$$

$$t \in \mathbb{R}_+, \text{ for all } j = 1, \cdots, d. \quad (0.10)$$

Example 0-3. Nonlinear networks containing transistors and diodes, Sandberg [15] , [3] (Fig. 2).

The next O. D. E.:

$$\frac{d}{dt}u(t) + TF\left[C^{-1}(u)\right] + GC^{-1}(u) = B(t),\ t \geq 0 \qquad (0.11)$$

(where $u(t) \in \mathbb{R}^{2p+q}$) represents a network containing linear passive time-invariant resistors, p nonlinear transistors, q nonlinear diodes, and independent sources. The Gummel & Koehler type model is used for semiconductor elements. We assume that:

(i) G is the short-circuit conductance matrix of the (2p+q)-port and its Norton equivalent circuit characterization is

$$i = -Gv + B(t) \qquad (0.12)$$

where $v(t), i(t) \in \mathbb{R}^{2p+q}$ are the port-voltage and port-current at time t, respectively.

(ii) $T \triangleq T_1 \oplus T_2 \oplus \cdots\ T_p \oplus I_q$. (0.13)

$$T_k = \begin{bmatrix} 1 & -\alpha_r^{(k)} \\ -\alpha_f^{(k)} & 1 \end{bmatrix} \text{ with } 0 < \alpha_r^{(k)} < 1 \text{ and}$$

$$0 < \alpha_f^{(k)} < 1 \quad \text{for all } k = 1, \cdots, p. \qquad (0.14)$$

I_q is the qxq identity matrix.

(iii) $F(\cdot): \mathbb{R}^{2p+q} \to \mathbb{R}^{2p+q}$.

$$F(v) = (f_1(v_1), f_2(v_2), \cdots, f_{2p+q}(v_{2p+q}))^T, \text{ for all } v \in \mathbb{R}^{2p+q}; \qquad (0.15)$$

and $f_j(\cdot):\mathbb{R}\to\mathbb{R}$, $f_j(0) = 0$ and $f_j'(\alpha) \geq 0$ for all

$\alpha \in \mathbb{R}$, for all $j = 1,\cdots,2p+q$. (0.16)

(iv) $C^{-1}(\cdot)$ is the inverse of the mapping $C(\cdot):\mathbb{R}^{2p+q} \to \mathbb{R}^{2p+q}$, defined by:

$$C(v) \triangleq cv + \tau F(v) \quad \text{for all } v \in \mathbb{R}^{2p+q}, \tag{0.17}$$

where c and τ are both (2p+q)x(2p+q) constant positive diagonal matrices (v denotes the (2p+q)-dimensional port voltage).

(v) There exists a positive diagonal matrix $P > 0$ such that both PT and PG are strongly column-sum dominant. We can interpret eq.(0.11) as representing a nonlinear time-invariant RC network (see Fig. 3) containing dependent sources and driven by independent current sources. Equation (0.11) is of the form (0.1). Now, we want to show that the condition (0.10) is satisfied. Let $v(t) \triangleq C^{-1}(u(t))$ for all $t \in \mathbb{R}_+$. Observe that from assumption (iv): (a) v always exists (because C^{-1} is well-defined on $\mathbb{R}^{2p+q}$); (b) the derivative $Dv(u(t))$ is a (2p+q)x(2p+q) diagonal and uniformly positive matrix; and (c) $DF(v(u(t))) \in \mathbb{R}^{(2p+q)x(2p+q)}$ is diagonal and nonnegative for all $u(t) \in \mathbb{R}^{2p+q}$. Then, eq.(0.11) is rewritten as:

$$\frac{d}{dt} u(t) + TF\circ v(u) + Gv(u) = B(t),\ t \geq 0. \tag{0.18}$$

Then, using the chain rule and commutativity of diagonal matrices we obtain:

$$-PD_1f(x,t)P^{-1} = P\left\{T\cdot DF(v(\cdot))\cdot Dv(\cdot) + G\cdot Dv(\cdot)\right\}P^{-1}$$

$$= PTP^{-1}\cdot DF(v(\cdot))\cdot Dv(\cdot) + PGP^{-1}\cdot Dv(\cdot) \tag{0.19}$$

where $v(\cdot)$ is everywhere evaluated at $u(t)$.

First, observe that $-PD_1f(x,t)P^{-1}$ is column-sum dominant for all $u \in \mathbb{R}^{2p+q}$, since both PT and PG are column-sum dominant; the right multiplication by any positive diagonal matrix preserves the column-sum dominance property; and the sum of two column-sum dominant matrices is again column-sum dominant. To show that $-PD_1f(x,t)P^{-1}$ is uniformly column-sum dominant, observe that if $DF(v(\cdot))$ is bounded for all $u(t) \in \mathbb{R}^{2p+q}$, $Dv(\cdot)$ is positive for all $u(t) \in \mathbb{R}^{2p+q}$ and that if $DF(v(\cdot))$ is not bounded for some $u^*(t) \in \mathbb{R}^{2p+q}$, $Dv(u^*(t))$ is no longer positive, but $D_1F(v(u^*(t)))\cdot D_1v(u^*(t))$ is strictly positive. For more detailed calculation, see the literature [3] (pp. 1766-1767).

Example 0-4. The Xenon poisoning equation of a nuclear reactor is written as, [27], [14]:

$$\begin{cases} \dot{X}(t) = -\mu_1 X(t) + \mu_2 I(t) + af(t) - bX(t)f(t) \\ \dot{I}(t) = -\mu_2 I(t) + cf(t) \end{cases} \tag{0.20}$$

where $X(t)$ and $I(t)$ are the concentration of Xenon X^{135} and Iodine I^{135} at time t, respectively; μ_1 and μ_2 are positive constants called decay constants of X^{135} and I^{135}, respectively; $f(t)$ is the neutron flux at time t; a, b, and c are positive constants. The first equation shows that the net accumulation rate of $X(t)$ is the algebraic sum of formation term $\mu_2 I(t) + af(t)$ and removal term $-\mu_1 X(t) - bX(t)f(t)$. The term $\mu_2 I(t)$ is due to the decay of I^{135} and the term $af(t)$ is due to the fission. The term $-\mu_1 X(t)$ is due to the decay of X^{135} itself and the term $-bX(t)f(t)$ is due to the capture reaction. The second equation shows that the net accumulation rate of I^{135} is the sum of the formation term $cf(t)$ due to the fission and the removal term $-\mu_2 I(t)$ due to the decay of I^{135} itself.

We assume:

(i) There exists a constant $\alpha > 0$ such that $0 < \alpha \leq \mu_1 + bf(t)$, for all $t \in \mathbb{R}_+$ and

(ii) f is continuous on $\mathbb{R}_+$.

Equation (0.20) is of the form (0.1). By choosing $P = \mathrm{diag}(1, 2)$, we obtain:

$$-PD_1 f(x,t)P^{-1} = P \begin{bmatrix} \mu_1 + bf(t) & -\mu_2 \\ 0 & \mu_2 \end{bmatrix} P^{-1}$$

$$= \begin{bmatrix} \mu_1 + bf(t) & -\frac{1}{2}\mu_2 \\ 0 & \mu_2 \end{bmatrix} \tag{0.21}$$

Hence, $-PD_1 f(x,t)P^{-1}$ with $P = \mathrm{diag}(1,2)$ is uniformly column-sum dominant from the assumption.

<u>Example 0-5</u>. Plate-type distillation column model having only a reboiler, vapor space and condenser, Rosenbrock [26], Gould [28] (Fig. 4).

Referring to Fig.4, the equation of the mass balance at each plate follows as:

$$\begin{cases} \dfrac{d(H_0x_0)}{dt} = -V_0'y_0' - P_0x_0 + L_1x_1 + F_0z_0 \\ \dfrac{d(h_0y_0)}{dt} = V_0'y_0' - (V_0+Q_0)\cdot y_0 + F_0'z_0' \\ \dfrac{d(H_1x_1)}{dt} = V_0y_0 - (L_1+P_1)\cdot x_1 + F_1z_1 \end{cases} \tag{0.22}$$

where $V_0(V_0')$ is the vapor flow from vapor space above zeroth plate (from liquid on zeroth plate to vapor space above zeroth plate) of composition $y_0(y_0')$; $H_r(h_r)$, $r=0,1$ is the liquid (vapor) holdup on (above) r-th plate; p_0 is the pressure above the zeroth plate; P_r, $r=0,1$ is the liquid withdrawal of composition x_r; L_1 is the liquid flow from the first plate of composition x_1; F_r (F_r'), $r=0,1$ is the liquid (vapor) feed of composition $z_r(z_r')$; Q_0 is the vapor withdrawal of composition y_0; $y_0' = f(x_0,p_0)$ is the vapor-liquid equilibrium characteristic of the zeroth plate. The first and the second equations of (0.22) represent the dynamics of the reboiler, and the third represents that of the condenser. Let $\xi_0(t) \triangleq H_0x_0(t)$, $\xi_1(t) \triangleq h_0y_0(t)$ and $\xi_2(t) \triangleq H_1x_1(t)$. Then, eq.(0.22) becomes:

$$\begin{cases} \dot{\xi}_0 = -V_0'f(\xi_0H_0^{-1}, p_0) - P_0H_0^{-1}\xi_0 + L_1H_1^{-1}\xi_2 + F_0z_0 \\ \dot{\xi}_1 = V_0'f(\xi_0H_0^{-1}, p_0) - (V_0+Q_0)h_0^{-1}\xi_1 + F_0'z_0' \\ \xi_2 = V_0h_0^{-1}\xi_1 - (L_1+P_1)H_1^{-1}\xi_2 + F_1z_1 \end{cases} \tag{0.23}$$

where $(\xi_0(t),\ \xi_1(t),\ \xi_2(t))^T$ is the state variable and $(z_0,\ z_0',\ z_1)^T$ is the constant input. We assume that

$$f_{x_0} \triangleq \frac{\partial}{\partial x_0} f(x_0, p_0) \geqq 0, \text{ for all } x_0 \geqq 0 \text{ and that} \tag{0.24}$$

V_r, V_r', H_r, h_r, P_r, p_r, Q_r, L_r, F_r, F_r', r=0,1 are all positive constants. Then, eq.(0.23) is of the form (0.1) and observe that

$$-D_1 f(x,t) = \begin{bmatrix} (V_0' f_{x_0} + P_0)H_0^{-1} & 0 & L_1 H_1^{-1} \\ -V_0' f_{x_0} H_0^{-1} & (V_0+Q_0)h_0^{-1} & \\ 0 & -V_0 h_0^{-1} & (L_1+P_1)H_1^{-1} \end{bmatrix} \tag{0.25}$$

is uniformly column-sum dominant.

<u>Example 0-6</u>. Co-current heat exchanger model, Rosenbrock [26]. Consider a co-current heat exchanger which is described as (n+1) consective elements labelled by r (r=0,1,···,n). Temperature is assumed constant for each liquid in each element. The mass flow rates L and L' are constant. Then, from the heat balance, we obtain:

$$\begin{cases} \dot{\xi}_{2r} = LH_{r-1}^{-1}\,\xi_{2r-2} - LH_r^{-1}\,\xi_{2r} - w_r(c^{-1}H_r^{-1}\,\xi_{2r}, c'^{-1}H_r'^{-1}\,\xi_{2r+1}) \\ \dot{\xi}_{2r+1} = L'H_{r-1}'^{-1}\,\xi_{2r-1} - L'H_r'^{-1}\,\xi_{2r+1} + w_r(c^{-1}H_r^{-1}\,\xi_{2r}, \\ c'^{-1}H_r'^{-1}\,\xi_{2r+1}), \quad r = 0,1,\cdots,n, \end{cases} \tag{0.26}$$

where $\xi_{2r}(t) \triangleq H_r c\theta_r(t)$ and $\xi_{2r+1}(t) = H_r'c'\theta_r'(t)$; H_r and H_r' are positive constant masses of the liquid in the r-th element; c and c' are the specific heats of the two liquids (positive constant); $\theta_r(t)$ and $\theta_r'(t)$ are the temperatures of the r-th element at time t; L and L' are the positive constant mass flow rates; and $w_r(\theta_r(t),\theta_r'(t))$ is the exchange heat rate in the r-th element. We assume that there exists a positive constant $\varepsilon > 0$ such that

$$\frac{\partial w_r(\theta_r,\theta_r')}{\partial \theta_r} \geqq \varepsilon > 0 \text{ and } \frac{\partial w_r(\theta_r,\theta_r')}{\partial \theta_r'} \leqq -\varepsilon < 0$$

$$\text{for all } \theta_r(t),\ \theta_r'(t) \in \mathbb{R},\ \text{for all } r = 0,1,\cdots,n. \tag{0.27}$$

Observe that (0.26) is of the form (0.1). Let $P = \mathrm{diag}(1,1,2^{-1},2^{-1},\cdots,2^{-n},2^{-n})$. Then, typical columns of $-PD_1f(x,t)$ are easily written with only the following non-zero elements:

$$
\begin{array}{lcc}
(2r+1)\text{th row} \hookrightarrow & \dfrac{1}{2^r}\left[\dfrac{L}{H_r}+\dfrac{\partial w_r}{\partial \theta_r}\dfrac{1}{cH_r}\right] & \dfrac{1}{2^r}\dfrac{\partial w_r}{\partial \theta'_r}\dfrac{1}{c'H'_r} \\[2ex]
(2r+2)\text{th row} \hookrightarrow & \dfrac{-1}{2^r}\dfrac{\partial w_r}{\partial \theta_r}\dfrac{1}{cH_r} & \dfrac{1}{2^r}\left[\dfrac{L'}{H'}-\dfrac{\partial w_r}{\partial \theta'_r}\dfrac{1}{c'H'_r}\right] \\[2ex]
(2r+3)\text{th row} \hookrightarrow & \dfrac{-1}{2^{r+1}}\dfrac{L}{H_r} & 0 \\[2ex]
(2r+4)\text{th row} \hookrightarrow & 0 & \dfrac{-1}{2^{r+1}}\dfrac{L'}{H'_r} \\[2ex]
 & \uparrow & \uparrow \\
 & (2r+1)\text{th column} & (2r+2)\text{th column}
\end{array}
\qquad (0.28)
$$

From (0.28) and (0.27), we observe that $-PD_1f(x,t)$ is uniformly column-sum dominant. Since P^{-1} is positive diagonal and $-PD_1f(x,t)$ is uniformly column-sum dominant, the condition (0.10) is satisfied.

A class of O. D. E.'s we discuss in this thesis is of the form (0.1) and contains the class ND and the class NCSD as special cases.

3. Contributions of This Thesis

Lemma 1-2 gives properties of the measure $\mu(\cdot)$. Properties (j), (k) and (ℓ) are new, and these properties play a crucial role in this thesis.

Lemma 1-4 gives equivalent statements to the definitions of row-sum dominant, column-sum dominant and passive matrices in terms of the measure $\mu(\cdot)$.

Theorem 2-2, Corollary 2-3, Corollary 2-4 and Corollary 2-5 unify and generalize previous work on the existence and uniqueness of D. C. solution by Stern [8], Willson Jr. [9], Ohtsuki & Watanabe [10] and Kuh & Hajj [11]. The generalization and unification are two fold: first, the choice of a vector norm is arbitrary and second, the uniformity condition is relaxed.

Theorem 2-6 and Corollary 2-7 determine a guaranteed region of convergence and establish the quadratic convergence for the Newton-Raphson method for infinite precision arithmetic computation.

Lemma 2-8 is a slightly modified version of Hurt's corollaries, [13], which is a kind of Lyapunov stability theorem for difference equations, where the continuity of the Lyapunov function is not required and the Lyapunov function can possibly increase along some solution sequence.

Theorem 2-9 and Corollary 2-10 show the effect of the local round-off error on the radius of the convergence region and on the convergence for the computation.

Lemma 3-1 is a slightly generalized version of Coppel's inequality where it is extended to the piecewise continuous case.

Theorem 3-2 and Corollary 3-3 give an estimate of the upper bound on the exact solution of O. D. E. The estimate is essentially due to Dahlquist [1], but it is extended to the piecewise continuous case. Corollary 3-3 includes previous work under ℓ^2 norms, ℓ^1 norms and weighted ℓ^1 norms, Rosenbrock [14], Sandberg [15], Mitra & So [16].

Theorem 3-4 gives an estimate of the upper bound on the difference of two solutions of O. D. E. starting from different initial states and different inputs. Corollary 3-5 gives an estimate of the upper bound on the difference between the exact solution and the equilibrium point of O. D. E. Both Theorem 3-4 and Corollary 3-5 include as special cases previous work under weighted ℓ^1 norms, Sandberg [15], Mitra & So [16].

Theorem 3-6, Corollary 3-7, Theorem 3-8, and Corollary 3-9 give estimates for lower bounds corresponding to Theorem 3-2, Corollary 3-3, Theorem 3-4 and Corollary 3-5, including special cases under weighted ℓ^1 norms by Sandberg [15].

Theorem 4-1 and Corollary 4-2 give estimates for the bound on the computed sequence by the backward Euler method, which generalize special cases under ℓ^2 norms and weighted ℓ^1 norms, Sandberg & Shichman [17], Sandberg [3].

Theorem 4-3 gives an estimate for the bound on the error between the computed sequence by the backward Euler method and the computed sequence by a modified implementable method. Theorem 4-3 is a generalization of earlier results under ℓ^1 norms and ℓ^2 norms, Sandberg [3], Sandberg & Shichman [17].

Theorem 4-4 gives an explicit estimate for the bound on the computed sequence where we use only one step of the Newton-Raphson method at each time step of the backward Euler method. Similar results under ℓ^2 norms were proved by

Sandberg & Shichman [17], but the estimate in Theorem 4-4 is more explicit and general.

Theorem 4-5 gives an estimate for the bound on the error sequence between the computed sequence by the backward Euler method and the one by the method stated in Theorem 4-4.

Theorem 4-6 gives an estimate for the bound on the so-called accumulated truncation error incurred by the backward Euler method. This is a generalization of a previous work under weighted ℓ^1 norms by Sandberg [3].

In Section 3 of Chapter IV, we make following comments on the implicit equation obtained by the backward Euler method under reasonable assumptions:
(i) The existence and uniqueness of the solution is guaranteed for any (large) step size; (ii) The guaranteed convergence region of the Newton-Raphson method applied to the implicit equation is monotonically enlarged as the step size becomes smaller; (iii) The error estimate between the exact solution and any computed solution is given by (4.65) using a priori known quantities.

4. Notation

$\mathbb{R}$ ($\mathbb{C}$)	field of real (complex) numbers
$\mathbb{R}_+$	set of nonnegative real numbers
$\mathbb{Z}_+$	set of nonnegative integers
$\mathbb{R}^d$ ($\mathbb{C}^d$)	direct product of $\mathbb{R}$'s ($\mathbb{C}$'s), d times

$\mathbb{R}^{d\times d}$ ($\mathbb{C}^{d\times d}$)	set of dxd real (complex) matrices
$\lvert\cdot\rvert$	vector norm on $\mathbb{R}^d$ or $\mathbb{C}^d$
$\lVert\cdot\rVert$	induced matrix norm on $\mathbb{R}^{d\times d}$ or $\mathbb{C}^{d\times d}$
$\mu(\cdot)$	measure of a matrix (definition: Ch.I, Sec.1)
I	identity matrix
$\lambda_i(A)$	i-th eigenvalue of a matrix A
Re z	real part of a complex number z
$\triangleq$	is equal to by definition
$o(\cdot)$	quantity, say x, such that $(x/h) \to 0$ as $h \to 0$
A^*	cojugate transpose of A
A^T	transpose of A
$\langle\cdot,\cdot\rangle$	scalar product on $\mathbb{R}^d$
$\cup$	union
$u(\cdot)$	input
$x(\cdot)$	exact solution of O. D. E.
$\{y_n\}_0^\infty$, $\{\hat{y}_n\}_0^\infty$, $\{\bar{y}_n\}_0^\infty$	computed solution
h	step size
t	time
$Df(x)$	derivative of f at x (Jacobian when $f: \mathbb{R}^d \to \mathbb{R}^d$)
$D_1f(x,t)$	derivative of $x \mapsto f(x,t)$ at x
$D_2f(x,t)$	derivative of $t \mapsto f(x,t)$ at x

$\mathcal{C}^1$	class of continuously differentiable functions
det(A)	determinant of A
x^*	exact D. C. solution
$\{x_n\}_0^\infty$	computed sequence for D. C. solution
$\tilde{x}$	computed D. C. solution
$\Phi(t,t_0)$	state transition matrix
θ_d	zero vector on $\mathcal{R}^d$ or $\mathcal{C}^d$
$\diamond$	Q.E.D.

Equations are sometimes assigned a number which is located in the right margin: (2.3) means eq.(3) of Chapter II. Theorems, Lemmas and Corollaries are numbered consecutively within each chapter: Theorem 2-4 follows Corollary 2-3 which itself follows Lemma 2-2.

CHAPTER I.

PRELIMINARIES

In this chapter we define the measure of a matrix and prove in detail its properties, some of which are new. Also, we explain a class of implicit integration formulae and the Newton-Raphson method.

1. Measure of A Matrix

The measure $\mu(\cdot)$ of a matrix was discussed by Dahlquist [1], and was used to investigate the stability of ordinary differential equations(O. D. E.'s), [1], [2].

Definition. Let $\mathbb{C}^d$ be $\mathbb{C} \times \mathbb{C} \times \cdots \times \mathbb{C}$, d times. Let $|\cdot|$ denote a vector norm on $\mathbb{C}^d$. Let A be a dxd complex matrix, and $\|\cdot\|$ be an induced matrix norm corresponding to $|\cdot|$. The measure $\mu(\cdot): \mathbb{C}^{d\times d} \to \mathbb{R}$ of a matrix is defined by

$$\mu(A) \triangleq \lim_{\theta \downarrow 0+} \frac{\|I + \theta A\| - 1}{\theta},$$

where I is the dxd identity matrix.

Remark. By the definition of $\mu(\cdot)$, $\mu(A)$ is seen to be a one-sided directional derivative of a mapping $\|\cdot\|: \mathbb{C}^{d\times d} \to \mathbb{R}_+$ at the point $I \in \mathbb{C}^{d\times d}$ in the direction of $A \in \mathbb{C}^{d\times d}$.

The following lemma shows that $\mu(\cdot)$ is well-defined.

Lemma 1-1. Dahlquist [1], Coppel [2]. For any dxd complex matrix A, the measure $\mu(A)$ exists.

Proof. Let $k \in (0,1)$.

$$\frac{\|I + k\theta A\| - 1}{k\theta} = \frac{\|k(I + \theta A) + (1-k)\cdot I\| - 1}{k\theta}$$

$$\leq \frac{k\,\|I + \theta A\| + (1-k) - 1}{k\theta} \text{, by triangle inequality.}$$

$$= \frac{\|I + \theta A\| - 1}{\theta} \;.$$

Hence, $\theta \longmapsto \dfrac{\|I + \theta A\| - 1}{\theta}$ is non-decreasing.

$$\frac{\|I + \theta A\| - 1}{\theta} \geq \frac{1 - \theta\|A\| - 1}{\theta} = -\|A\|, \text{ by triangle}$$

inequality and homogeneity.

Since $\dfrac{\|I + \theta A\| - 1}{\theta}$ is bounded from below and decreases as $\theta \downarrow 0+$, the limit $\mu(A)$ exists. $\diamond$

Remark. The measure $\mu(\cdot)$ depends on the choice of the original vector norm $|\cdot|$.

Lemma 1-2. Properties of $\mu(\cdot)$. Let A and B be in $\mathbb{C}^{d\times d}$.

(a) $\mu(I) = 1, \mu(-I) = -1.$

(b) If $A = \theta_{dxd}$ (dxd zero matrix), then $\mu(A) = 0$.

(c) $-\|A\| \leq -\mu(-A) \leq \mu(A) \leq \|A\|$.

(d) $\mu(cA) = c\mu(A)$ for all $c \geq 0$. (positive homogeneity)

(e) $\mu(A + cI) = \mu(A) + c$ for all $c \in \mathbb{R}$.

(f) $\max\{\mu(A) - \mu(-B), -\mu(-A) + \mu(B)\} \leq \mu(A + B)$
$\leq \mu(A) + \mu(B)$. (sub-additivity)

(g) $\mu[\lambda A + (1 - \lambda)B] \leq \lambda\mu(A) + (1 - \lambda)\mu(B)$ for all
$\lambda \in [0,1]$. (convexity)

(h) $|\mu(A) - \mu(B)| \leq \max\{|\mu(A - B)|, |\mu(B - A)|\}$
$\leq \|A - B\|$.

(i) $-\mu(-A) \leq \text{Re}\,\lambda_i(A) \leq \mu(A)$ for all $i = 1, 2, \cdots , d$,
where $\text{Re}\,\lambda_i(A)$ denotes the real part of the eigenvalue
$\lambda_i(A)$ of the matrix A.

(j) $|Ax| \geq \max\{-\mu(-A), -\mu(A)\}\cdot|x|$ for all $x \in \mathbb{C}^d$.

(k) Let $|\cdot| : \mathbb{C}^d \longrightarrow \mathbb{R}_+$ be a vector norm in $\mathbb{C}^d$. Define
$|x|_P \triangleq |Px|$, where P is a nonsingular dxd complex matrix
and call μ_P the measure defined in terms of the corresponding induced norm. Then, $\mu_P(A) = \mu(PAP^{-1})$.

(ℓ) Let A be a nonsingular dxd complex matrix. Then,

$$\frac{1}{\|A^{-1}\|} \geq \max\left\{-\mu(-A), -\mu(A)\right\}.$$

<u>Proof</u>. (a) The results are immediate from the definition of the measure $\mu(\cdot)$.

(b) The result is trivially true by the definition of $\mu(\cdot)$.

(c) Observe that

$$\frac{||I - \theta A|| - 1}{\theta} + \frac{||I + \theta A|| - 1}{\theta} \geq \frac{||2I|| - 2}{\theta} = 0,$$

by triangle inequality, or

$$-\frac{||I - \theta A|| - 1}{\theta} \leq \frac{||I + \theta A|| - 1}{\theta}.$$

So, $\mu(-A) + \mu(A) \geq 0$.

Observe that

$$-||A|| = \frac{1 - \theta||A|| - 1}{\theta} \leq -\frac{||I - \theta A|| - 1}{\theta} \leq \frac{||I + \theta A|| - 1}{\theta}$$

$\leq \dfrac{1 + \theta||A|| - 1}{\theta} = ||A||$, since $\theta > 0$ and by triangle inequality.

(d) If $c = 0$, the result is true by the property (b).

Assume that $c > 0$. Observe that

$\dfrac{||I + c\theta A|| - 1}{\theta} = c \cdot \dfrac{||I + c\theta A|| - 1}{c\theta}$, and that $c\theta \downarrow 0+$ as $\theta \downarrow 0+$ since the constant c is > 0.

(e) Observe that

$$\frac{||I + \theta(A + cI)|| - 1}{\theta} = \frac{(1+\theta c)\left\|I + \frac{\theta}{1 + \theta c}A\right\| - 1}{\theta}$$

$$= \frac{\left\|I + \frac{\theta}{1 + \theta c}A\right\| - 1}{\frac{\theta}{1 + \theta c}} + c.$$

Since $\dfrac{\theta}{1 + \theta c} \downarrow 0+$ as $\theta \downarrow 0+$, the result follows.

(f) Observe that

$$\frac{\|I+\theta(A+B)\|-1}{\theta} = \frac{\|(I+2\theta A)+(I+2\theta B)\|-2}{2\theta}$$

$$\leq \frac{\|I+2\theta A\|-1}{2\theta} + \frac{\|I+2\theta B\|-1}{2\theta}.$$

Hence, $\mu(A+B) \leq \mu(A) + \mu(B)$.

The other inequalities follow from $A = (-B) + (A+B)$ and $B = (-A) + (A+B)$.

(g) The convexity property follows from the positive homogeneity (d) and the sub-additivity (f).

(h) The property (c) implies that

$$\max\{|\mu(A-B)|,\ |\mu(B-A)|\} \leq \|A-B\|.$$

The other inquality is obtained by observing

$$-\mu(B-A) \leq \mu(A) - \mu(B) \leq \mu(A-B) \quad \text{and}$$

$$-\mu(A-B) \leq \mu(B) - \mu(A) \leq \mu(B-A).$$

(i) Let $e \in \mathbb{C}^d$ be a normalized eigenvector of A associated with the eigenvalue λ_i. Observe that

$$\frac{\|I+\theta A\|-1}{\theta} \geq \frac{|e+\theta Ae|-1}{\theta} = \frac{|e+\theta\lambda_i e|-1}{\theta}$$

$$= \frac{|1+\theta\lambda_i|\cdot|e|-1}{\theta} = \frac{|1+\theta\lambda_i|-1}{\theta}, \text{ and that}$$

$$|1+\theta\lambda_i| = 1 + \theta \operatorname{Re}\lambda_i + o(\theta) \quad \text{for sufficiently small } \theta > 0.$$

The other inequality follows from

$$-\frac{\|I-\theta A\|-1}{\theta} \leq -\frac{|e-\theta Ae|-1}{\theta} = -\frac{|1-\theta\lambda_1|-1}{\theta}.$$

(j) Let θ be > 0.

$$|Ax| = \frac{|(x-\theta Ax)-x|}{\theta} = \frac{|(I-\theta A)x-x|}{\theta}$$

$$\geq \frac{|x| - \|I-\theta A\|\cdot|x|}{\theta} = -|x|\frac{\|I-\theta A\|-1}{\theta}.$$

Hence, $|Ax| \geq -\mu(-A)\cdot|x|$ by letting $\theta\downarrow 0+$. Also,
$|Ax| = |(-A)x| \geq -\mu[-(-A)]\cdot|x| = -\mu(A)\cdot|x|$.

(k) Observe that

$$\|I+\theta A\|_P \triangleq \sup_{x\neq\theta_d}\frac{|x+\theta Ax|_P}{|x|_P} = \sup_{x\neq\theta_d}\frac{|P(x+\theta Ax)|}{|Px|}$$

$$= \sup_{x\neq\theta_d}\frac{|Px+\theta(PAP^{-1})Px|}{|Px|} = \|I+\theta PAP^{-1}\|.$$

(ℓ) Claim: $\inf_{|x|=1}|Ax| = \dfrac{1}{\|A^{-1}\|}$.

$$\inf_{|x|=1}|Ax| = \inf_{x\neq\theta_d}\frac{|Ax|}{|x|} = \frac{1}{\sup_{x\neq\theta_d}\frac{|x|}{|Ax|}} = \frac{1}{\sup_{Ax\neq\theta_d}\frac{|A^{-1}(Ax)|}{|Ax|}}$$

$$= \frac{1}{\|A^{-1}\|}.$$

Hence, $\dfrac{1}{\|A^{-1}\|} = \max\{\lambda \mid |Ax| \geq \lambda \text{ and } |x| = 1\}$.

Since $|Ax| \geq \max\{-\mu(-A), -\mu(A)\} \cdot |x|$ by (j),

$$\frac{1}{\|A^{-1}\|} \geq \max\{-\mu(-A), -\mu(A)\}. \quad \diamond$$

<u>Remark</u>. Since the measure $\mu(\cdot)$ is a convex function, it is continuous. As we have shown, the measure $\mu(\cdot)$ is in some ways similar to the norm of a matrix, however $\mu(\cdot)$ is only <u>positively</u> homogeneous and can take on <u>negative</u> values. We can easily verify that a mapping $\theta \mapsto \dfrac{\|I + \theta A\| - 1}{\theta}$ is continuous and monotone increasing except at $\theta = 0$, and that

$$-\|A\| \leq \frac{\|I + \theta A\| - 1}{\theta} \leq \|A\| \quad \text{for all } \theta \in \mathbb{R} \text{ except } 0.$$

Also note that $\lim_{\theta \uparrow 0-} \dfrac{\|I + \theta A\| - 1}{\theta} = -\mu(-A) \leq \mu(A)$

$\triangleq \lim_{\theta \downarrow 0+} \dfrac{\|I + \theta A\| - 1}{\theta}$. We shall obtain tighter bounds for the stability analysis of O. D. E.'s and its numerical integration formulas by the use of the measure $\mu(\cdot)$ rather than by the use of norms. A key tool is the following inequality due to Coppel, [2]:

$$\exp(-\|A\|t) \leq \exp(-\mu(-A)t) \leq \frac{1}{\|(\exp(At))^{-1}\|}$$

$$\leq \|\exp(At)\| \leq \exp(\mu(A)t) \leq \exp(\|A\|t) \quad \text{for all } t \geq 0.$$

Another case is the following: if $h > 0$, then(by Lemma 1-2, (ℓ))

$$\frac{1}{\|(I+hA)^{-1}\|} \geq 1 - h\mu(-A) \geq 1 - h\|A\|.$$

The values of $\mu(A)$ are easy to compute for ℓ^1, ℓ^2 and ℓ^∞ norms.

<u>Lemma 1-3</u>. The values of $\|A\|$ and $\mu(A)$.

Let A be a dxd complex matrix.

(a) If $|x| = |x|_\infty \triangleq \max_{i=1,2,\dots,d} |x_i|$, then

$$\|A\|_\infty = \max_{i=1,2,\dots,d} \sum_{j=1}^{d} |a_{ij}| \text{ and}$$

$$\mu_\infty(A) = \max_{i=1,2,\dots,d} \Big(\operatorname{Re} a_{ii} + \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}| \Big). \quad \text{(row sum)}$$

(b) If $|x| = |x|_1 \triangleq \sum_{i=1}^{d} |x_i|$, then

$$\|A\|_1 = \max_{j=1,2,\dots,d} \sum_{i=1}^{d} |a_{ij}| \text{ and}$$

$$\mu_1(A) = \max_{j=1,2,\dots,d} \Big(\operatorname{Re} a_{jj} + \sum_{\substack{i=1 \\ (i\neq j)}}^{d} |a_{ij}| \Big). \quad \text{(column sum)}$$

(c) If $|x| = |x|_2 \triangleq \Big(\sum_{i=1}^{d} |x_i|^2 \Big)^{1/2}$, then

$$\|A\|_2 = \Big[\max_{i=1,2,\dots,d} \big\{ \lambda_i(A^*A) \big\} \Big]^{1/2} \text{ and}$$

$\mu_2(A) = \max_{i=1,2,\dots,d}\left\{\lambda_i\left(\frac{A + A^*}{2}\right)\right\}$, where A^* is a conjugate transpose of A.

Proof.

(a) $$\|I + \theta A\|_\infty = \max_{i=1,\dots,d} \sum_{j=1}^{d} |\delta_{ij} + \theta a_{ij}|$$

$$= \max_{i=1,\dots,d}\left\{|1 + \theta a_{ii}| + \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |\theta a_{ij}|\right\}$$

$$= \max_{i=1,\dots,d}\left\{1 + \theta \mathrm{Re}a_{ii} + o(\theta) + \theta \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}|\right\}$$

for sufficiently small $\theta > 0$.

Hence, $$\frac{\|I + \theta A\|_\infty - 1}{\theta} = \max_{i=1,\dots,d}\left\{\mathrm{Re}a_{ii} + \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}|\right\} + o(\theta)$$

for sufficiently small $\theta > 0$.

(b) The proof is analogous to that of (a).

(c) $$\|I + \theta A\|_2 = \left[\max_{i=1,\dots,d}\left\{\lambda_i\left((I+\theta A)^*(I+\theta A)\right)\right\}\right]^{1/2}$$

$$= \left[\max_{i=1,\dots,d}\left\{\lambda_i\left(I + \theta(A+A^*) + \theta^2 A^* A\right)\right\}\right]^{1/2}$$

$$= \max_{i=1,\dots,d}\left\{\lambda_i\left(I + \theta(A+A^*) + \theta^2 A^* A\right)\right\}^{1/2}$$

$$= \max_{i=1,\cdots,d} \left\{ \lambda_i \left(I + \theta(A+A^*) + \theta^2 A^* A \right)^{1/2} \right\}$$

$$= \max_{i=1,\cdots,d} \left\{ 1 + \theta\lambda_i \left(\frac{A+A^*}{2}\right) + o(\theta) \right\} \quad \text{for suf-}$$

ficiently small $\theta > 0$. $\diamond$

There are classes of matrices which are called row-sum dominant, column-sum dominant and passive.

Definition. A dxd complex matrix A is said to be strongly (weakly) row-sum dominant iff

$$\mathrm{Re}\, a_{ii} > (\geq) \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}| \quad \text{for all } i = 1, 2, \cdots, d.$$

Definition. A dxd complex matrix A is said to be strongly (weakly) column-sum dominant iff

$$\mathrm{Re}\, a_{jj} > (\geq) \sum_{\substack{i=1 \\ (i\neq j)}}^{d} |a_{ij}| \quad \text{for all } j = 1, 2, \cdots, d.$$

Definition. A dxd real matrix A is said to be strongly (weakly) passive iff

$$\langle x, Ax \rangle > (\geq)\ 0 \quad \text{for all non-zero vector } x \in \mathbb{R}^d.$$

Sometimes, the strongly (or weakly) passive matrix is called positive definite (or positive semidefinite). Note that we do not require that the matrix A is symmetric.

The next lemma shows how those classes of matrices are related to $\mu(\cdot)$ under specific norms.

Lemma 1-4. Let A be a dxd real matrix.

(a) The matrix A is strongly (weakly) row-sum dominant iff $-\mu_\infty(-A) > (\geq)\ 0$.

(b) The matrix A is strongly (weakly) column-sum dominant iff $-\mu_1(-A) > (\geq)\ 0$.

(c) The matrix A is strongly (weakly) passive iff $-\mu_2(-A) > (\geq)\ 0$.

Proof. (a) Observe that

$$a_{ii} > (\geq) \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}| \quad \text{for all } i = 1, 2, \cdots, d.$$

$$\Longleftrightarrow -\Big(-a_{ii} + \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |-a_{ij}| \Big) > (\geq)\ 0 \quad \text{for all } i = 1, 2, \cdots, d.$$

(b) The proof is analogous to that of (a).

(c) Observe that

$$\langle x, Ax \rangle = \langle x, \frac{A+A^T}{2} x \rangle,$$ where A^T is a transpose of A. $\diamond$

Remark. If a dxd real matrix A is strongly column-sum dominant, i.e., there exists a positive constant $\varepsilon > 0$ such that

$$a_{jj} - \sum_{\substack{i=1\\(i\neq j)}}^{d} |a_{ij}| \geq \varepsilon > 0 \quad \text{for all } j = 1, 2, \cdots, d,$$

then by Lemma 1-4, (b) and Lemma 1-2, (j), Sandberg's result, [3] follows:

$$|Ax|_1 \geq \varepsilon |x|_1 \quad \text{for all } x \in \mathbb{R}^d.$$

Similar results are immediately obtained for strongly row-sum and strongly passive matrices.

The inequality in Lemma 1-2, (i) under ℓ^∞ norms can also be proved by the Gerschgorin circle theorem.

Gerschgorin circle theorem, [4] .

Let A be a dxd complex matrix. Then every eigenvalue of A lies in the set

$$\bigcup_{i=1}^{d} \left\{ z \in \mathbb{C} \,\middle|\, |a_{ii} - z| \leq \sum_{\substack{j=1\\(j\neq i)}}^{d} |a_{ij}| \right\} . \diamond$$

For each eigenvalue λ_i of A, there exists $i_0 \in \{1, 2, \cdots, d\}$ such that $|a_{i_0 i_0} - \lambda_i| \leq \sum_{\substack{j=1\\(j\neq i)}}^{d} |a_{i_0 j}|$.

Noting that

$$|\mathrm{Re}\,\lambda_i - \mathrm{Re}\,a_{i_0 i_0}| \leq |a_{i_0 i_0} - \lambda_i| ,$$

we obtain

$$-\mu_\infty(-A) = \min_{i=1,\dots,d} \left\{ \mathrm{Re}\,a_{ii} - \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}| \right\}$$

$$\leq \mathrm{Re}\,a_{i_0 i_0} - \sum_{\substack{j=1 \\ (j\neq i_0)}}^{d} |a_{i_0 j}| \leq \mathrm{Re}\,\lambda_i$$

$$\leq \mathrm{Re}\,a_{i_0 i_0} + \sum_{\substack{j=1 \\ (j\neq i_0)}}^{d} |a_{i_0 j}|$$

$$\leq \max_{i=1,\dots,d} \left\{ \mathrm{Re}\,a_{ii} + \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}| \right\} = \mu_\infty(A), \quad \text{for all}$$

$i = 1, 2, \cdots, d$.

2. Implicit Integration Formulae

One of the main concerns in lumped circuit analysis is the computation of the transient response of a circuit, i.e., to solve the appropriate O. D. E. in an efficient and accurate way. A class of numerical integration formulae is stated in this section.

Consider an O. D. E.:

$$\begin{cases} \dot{x} = f(x,t) + u(t) \\ x(0) = x_0 \end{cases} \tag{1.1}$$

where $x(t),\ u(t) \in \mathbb{R}^d$ for all $t \in \mathbb{R}_+$ and $f: \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$.

It is assumed that the existence and uniqueness of the solution

$x(\cdot)$ of the O. D. E. (1.1) is guaranteed and that it is continuous for all $t \in \mathbb{R}_+$. A sufficient condition is, for example, given in the reference [5].

Let $h > 0$ be a step size. A special class of algorithms for obtaining the numerical solution of O. D. E. (1.1) is:

$$y_{n+1} = \sum_{k=0}^{p} a_k y_{n-k} + \sum_{k=-1}^{p} b_k \dot{y}_{n-k}, \quad \text{with } b_{-1} \neq 0, \tag{1.2}$$

where $\dot{y}_{n-k} \triangleq f(\, y_{n-k}, (n-k)h\,) + u(\,(n-k)h\,)$.

For notational convenience, $x(nh)$, $u(nh)$ and $f(\,x(nh), nh\,)$ will be denoted by x_n, u_n and $f(x_n, n)$ respectively for all $n \in \mathbb{Z}_+$ from now on. The above algorithm (1.2) is called the <u>multi-point formula of closed type</u> or an <u>implicit integration formula</u>. The determination of y_{n+1} is implicit for given $\{ y_{n-p}, y_{n-p+1}, \cdots, y_n \}$, $\{ a_0, a_1, \cdots, a_p \}$, $\{ b_{-1}, b_0, \cdots, b_p \}$ and $\{ \dot{y}_0, \dot{y}_1, \cdots, \dot{y}_p \}$. In particular, when $p = 0$, $a_0 = 1$, $b_{-1} = h$ and $b_0 = 0$, the formula (1.2) is called the <u>backward Euler formula</u>:

$$y_{n+1} = y_n + h\dot{y}_{n+1} = y_n + hf(y_{n+1}, n+1) + hu_{n+1}. \tag{1.3}$$

The corresponding explicit integration formula is the <u>Euler-Cauchy</u> method:

$$y_{n+1} = y_n + h\dot{y}_n = y_n + hf(y_n, n) + hu_n. \tag{1.4}$$

The following example shows that the Euler-Cauchy method is not as good as the backward-Euler method even for a scalar linear O. D. E.

<u>Example 1-1</u>. Consider a scalar O. D. E.:

$$\begin{cases} \dot{\xi} = \lambda \xi \\ \xi(0) = \xi_0 \end{cases} \tag{1.5}$$

where $\xi(t) \in \mathbb{R}$ for all $t \in \mathbb{R}_+$ and $\lambda < 0$.
The exact solution $\xi(t) = \exp(\lambda t) \cdot \xi_0$ of O. D. E. (1.5) converges to 0 as $t \to \infty$. The computed solution $y_n = (1+h\lambda)^n y_0$ by the Euler-Cauchy method (1.4) converges to 0 as $n \to \infty$ if $0 < h < -2/\lambda$, otherwise it does not converge to 0 as $n \to \infty$. So, when $|\lambda|$ is large, the step size $h > 0$ has to be chosen sufficiently small to get over the numerical instability, which requires more computational time. But the computed solution $y_n = (1-h\lambda)^{-n} y_0$ by the backward Euler method (1.3) converges to 0 as $n \to \infty$ for any $h > 0$. Moreover, the accumulated truncation error $|\xi_n - y_n|$ of the backward Euler method has an upper bound:

$$|\xi_n - y_n| \leq (1-h\lambda)^{-n} |\xi_0 - y_0| + \tfrac{1}{2} |\lambda| h |\xi_0| \quad \text{for all } n \geq 1.$$

The error estimate consists of two terms: the first term shows that the effect of the initial error decays exponentially as $n \to \infty$ and the second shows that it is proportional to the step size h for any $h > 0$. In Chapter IV, the backward Euler method is more fully investigated. Using the measure $\mu(\cdot)$, we show that similar desirable properties still hold for an important class of nonlinear O. D. E.'s.

3. Newton-Raphson Method for Solving D. C. Equations

D. C. equations(algebraic equations) are encountered in computing the transient response of a circuit by implicit integration formulae and also in computing the D. C. operating point. The Newton-Raphson method is one of the widely used algorithms for solving D. C. equations. The scheme is stated in this section.

Consider a D. C. equation:

$$f(x) = y \tag{1.6}$$

where $x, y \in \mathbb{R}^d$ and f is a mapping from $\mathbb{R}^d$ into itself. Given $y \in \mathbb{R}^d$ and $f: \mathbb{R}^d \longrightarrow \mathbb{R}^d$, we want to find the D. C. solution $x^* \in \mathbb{R}^d$ such that $f(x^*) = y$ if it exists. The Newton-Raphson method of solving the D. C. equation (1.6) is given by:

$$x_{k+1} = x_k - \left(Df(x_k) \right)^{-1} \left(f(x_k) - y \right), \quad k = 1, 2, \cdots \tag{1.7}$$

with x_0 given; here $Df(x_k)$ denotes the derivative of f (i.e., the Jacobian of f) evaluated at x_k. Note that the Newton-Raphson method is applicable only when $(Df(x_k))$ is nonsingular for all x_k, $k \in \mathbb{Z}_+$. The Newton-Raphson method is essentially a linearization process. At k-th step, the D. C. equation (1.6) is linearized at $x = x_k$:

$$f(x) = y \cong f(x_k) + Df(x_k)\cdot(x - x_k). \tag{1.8}$$

Solving the linearized equation (1.8) for x and letting $x_{k+1} = x$, we obtain the formula (1.7).

CHAPTER II.

D. C. EQUATIONS

In this chapter, using the measure $\mu(\cdot)$ we develop properties of D. C. equations. First, we prove an existence and uniqueness theorem. Second, we determine the guaranteed convergence region and the rate of convergence of the Newton-Raphson method. The effect of the local round-off error is also investigated.

1. Existence and Uniqueness of D. C. Solution

Consider the D. C. equation (1.6), i.e.,

$$f(x) = y \tag{1.6}$$

where $x, y \in \mathbb{R}^d$ and f is a mapping from $\mathbb{R}^d$ into itself. In this section, existence and uniqueness of D. C. solution of eq. (1.6) and continuous dependence of the D. C. solution on a given vector $y \in \mathbb{R}^d$ are discussed. The above requirements of the D. C. solution are met for all $y \in \mathbb{R}^d$ if the mapping $f: \mathbb{R}^d \to \mathbb{R}^d$ is continuous & bijective and if the inverse mapping $f^{-1}: \mathbb{R}^d \to \mathbb{R}^d$ is continuous. (The latter statement follows from the former by The Invariance of Domain Theorem, [4].)

<u>Definition.</u> Let $f: \mathbb{R}^d \to \mathbb{R}^d$ be continuously differentiable ($f \in \mathcal{C}^1$). The mapping $f: \mathbb{R}^d \to \mathbb{R}^d$ is said to be a <u>$\mathcal{C}^1$-diffeomorphism</u> from $\mathbb{R}^d$ onto itself iff f is bijective and f^{-1} is in $\mathcal{C}^1$.

Palais, [6] gave the necessary and sufficient condition for the mapping f: $\mathbb{R}^d \to \mathbb{R}^d$ to be a C^1-diffeomorphism.

Lemma 2-1. (Global Inverse Function Theorem, Palais [6], Holzman & Liu [7], Stern [8], Ortega & Rheinboldt [4], Wu & Desoer [18].)

Let f: $\mathbb{R}^d \to \mathbb{R}^d$ be in C^1. Then, f is a C^1-diffeomorphism

iff (i) $\det(\, Df(x)\,) \neq 0$ for all $x \in \mathbb{R}^d$ (2.1)

and (ii) $\lim_{|x| \to \infty} |f(x)| = +\infty.$ $\diamond$ (2.2)

The condition (ii) of the Global Inverse Function Theorem is often not easy to check in specific cases. Sufficient conditions which are weaker but easier to check are given below.

Definition. A function $m(\cdot)$: $\mathbb{R}_+ \to \mathbb{R}_+$ is said to be in class $\mathcal{M}_0$ iff $m(\alpha) > 0$ for all $\alpha \in \mathbb{R}_+$ and $\int_0^\infty m(\alpha)\, d\alpha = +\infty$.

Theorem 2-2. Let f: $\mathbb{R}^d \to \mathbb{R}^d$ be in C^1. If there exists an $m(\cdot) \in \mathcal{M}_0$ such that either $-\mu(\, Df(x)\,) \geq m(|x|) > 0$ or $-\mu(\, -Df(x)\,) \geq m(|x|) > 0$ for all $x \in \mathbb{R}^d$, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself.

Proof. Use the Global Inverse Function Theorem.

Claim: $\det(\, Df(x)\,) \neq 0$ for all $x \in \mathcal{R}^d$.

Let z be a non-zero vector in $\mathcal{R}^d$, then for all $x \in \mathcal{R}^d$,

$|Df(x)\cdot z| \geq \max\left\{ -\mu(\, -Df(x)\,),\ -\mu(\, Df(x)\,)\right\} \cdot |z|$, by Lemma 1-2, (j),

$$\geq m(\,|x|\,)\cdot |z| > 0 \quad \text{for all } z \neq \theta_d. \tag{2.3}$$

Claim: $\lim_{|x| \to \infty} |f(x)| = +\infty$.

By Taylor's formula,

$$f(x) = f(\theta_d) + \left(\int_0^1 Df(\tau x)\, d\tau \right)\cdot x. \tag{2.4}$$

$$|f(x)| \geq \left| \left(\int_0^1 Df(\tau x)\, d\tau \right)\cdot x \right| - \left| f(\theta_d) \right|$$

$$\geq \max\left\{ -\mu\left(-\int_0^1 Df(\tau x)\, d\tau \right),\ -\mu\left(\int_0^1 Df(\tau x)\, d\tau \right)\right\} \cdot |x|$$

$$- \left| f(\theta_d) \right|, \quad \text{by Lemma 1-2, (j).} \tag{2.5}$$

$$-\mu\left(-\int_0^1 Df(\tau x)\, d\tau \right) \geq -\int_0^1 \mu(\, -Df(\tau x)\,)\cdot d\tau$$

$$= \int_0^1 -\mu(\, -Df(\tau x)\,)\, d\tau \quad \text{by Lemma 1-2,}$$

(d) & (f). (2.6)

Similarly,

$$-\mu\left(\int_0^1 Df(\tau x)\,d\tau\right) \geq -\int_0^1 \mu(\,Df(\tau x)\,)\,d\tau$$

$$= \int_0^1 -\mu(\,Df(\tau x)\,)\,d\tau\,. \qquad (2.7)$$

By assumption, we obtain

$$\max\left\{-\mu\left(-\int_0^1 Df(\tau x)\,d\tau\right),\ -\mu\left(\int_0^1 Df(\tau x)\,d\tau\right)\right\}$$

$$\geq \int_0^1 m(\,|\tau x|\,)\,d\tau \quad \text{for all } x\in\mathbb{R}^d \qquad (2.8)$$

Hence, the inequality (2.5) becomes:

$$|f(x)| \geq \int_0^1 m(\,|\tau x|)\,d\tau\cdot|x| - \left|f(\theta_d)\right|$$

$$= \int_0^{|x|} m(\alpha)\,d\alpha - \left|f(\theta_d)\right| \quad \text{by letting } \alpha=|\tau x|$$

$$= \tau\,|x|\,. \qquad (2.9)$$

So, $|f(x)|\to\infty$ as $|x|\to\infty$. ◇

<u>Remark</u>. Since f is in C^1 and $\mu(\cdot)$ is continuous, the conditions of Theorem 2-2 on Df(x) are mutually exclusive because either $-\mu(\,Df(x)\,) \geq m(\,|x|\,) > 0$ for all $x\in\mathbb{R}^d$ holds, or $-\mu(\,-Df(x)\,) \geq m(\,|x|\,) > 0$ for all $x\in\mathbb{R}^d$ holds.

<u>Definition</u>. A function $m(\cdot)$: $\mathbb{R}_+\to\mathbb{R}_+$ is said to be in <u>class</u>

$\mathcal{M}(\varepsilon)$ iff $m(\alpha) > 0$ for all $\alpha \in \mathbb{R}_+$ and there exists a positive constant $\varepsilon > 0$ such that $\int_0^\alpha m(\xi)\, d\xi \geq \varepsilon\alpha$ for all $\alpha \in \mathbb{R}_+$. (2.10)

Since the class $\mathcal{M}(\varepsilon)$ is a subset of the class $\mathcal{M}_0$, the next corollary follows.

Corollary 2-3. Let $f: \mathbb{R}^d \to \mathbb{R}^d$ be in C^1. If there exists an $m(\cdot) \in \mathcal{M}(\varepsilon)$ such that either $-\mu(\,Df(x)\,) \geq m(|x|) > 0$ or $-\mu(\,-Df(x)\,) \geq m(|x|) > 0$ for all $x \in \mathbb{R}^d$, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself. ◇

Corollary 2-4. Let $f: \mathbb{R}^d \to \mathbb{R}^d$ be in C^1. If there exists a positive constant $m > 0$ such that either $-\mu(\,Df(x)\,) \geq m > 0$ or $-\mu(\,-Df(x)\,) \geq m > 0$ for all $x \in \mathbb{R}^d$, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself.

Proof. The constant function $m \in \mathcal{M}(\varepsilon) \subset \mathcal{M}_0$. ◇

Examples of $m(\cdot)$ are given below.

Example 2-1. Consider a function $m(\cdot): \mathbb{R}_+ \to \mathbb{R}_+$ defined by

$$m(\alpha) \triangleq \varepsilon_0(\alpha + \alpha_0)^{-p} \tag{2.11}$$

where $\varepsilon_0 > 0$, $\alpha_0 > 1$ and $p \leq 1$.

First observe that $m(\cdot)$ defined above is in class $\mathcal{M}_0$.

(a) If $p \leq -1$, then $m(\cdot) \in \mathcal{M}(\varepsilon)$.

(b) If $-1 < p \leq 1$, then $m(\cdot) \notin \mathcal{M}(\varepsilon)$, but $m(\cdot) \in \mathcal{M}_0$.
In particular, if $p = 0$, then $m(\alpha) = \varepsilon_0 =$ constant.

By choosing specific norms, ℓ^1, ℓ^2 and ℓ^∞, for Corollary 2-4, we can derive more special cases. Before giving the next corollary, uniformly row-sum, uniformly column-sum and uniformly positive definite(or negative definite) matrices are defined. Let $A(x)$ denote a dxd real matrix with a parameter $x \in \mathbb{R}^d$.

Definition. The matrix $A(x)$ is said to be uniformly row-sum dominant iff there exists a positive constant $m > 0$ such that

$$a_{ii}(x) - \sum_{\substack{j=1 \\ (j\neq i)}}^{d} |a_{ij}(x)| \geq m > 0 \quad \text{for all } i = 1, 2, \cdots, d,$$

for all $x \in \mathbb{R}^d$.

Definition. The matrix $A(x)$ is said to be uniformly column-sum dominant iff there exists a positive constant $m > 0$ such that

$$a_{jj}(x) - \sum_{\substack{i=1 \\ (i\neq j)}}^{d} |a_{ij}(x)| \geq m > 0 \quad \text{for all } j = 1, 2, \cdots, d,$$

for all $x \in \mathbb{R}^d$.

Definition. The matrix $A(x)$ is said to be uniformly positive definite(or uniformly passive) iff there exists a positive constant $m > 0$ such that

$$\langle y, A(x)y \rangle \geq m|y|_2^2 \quad \text{for all } y \in \mathbb{R}^d, \text{ for all } x \in \mathbb{R}^d.$$

The matrix $A(x)$ is said to be uniformly negative definite iff $-A(x)$ is uniformly positive definite.

Corollary 2-5. Let $f: \mathbb{R}^d \to \mathbb{R}^d$ be in C^1.
(a) If either $Df(x)$ or $-Df(x)$ is uniformly column-sum dominant, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself.
(b) If either $Df(x)$ is either uniformly positive definite or uniformly negative definite, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself.
(c) If either $Df(x)$ or $-Df(x)$ is uniformly row-sum dominant, then f is a C^1-diffeomorphism from $\mathbb{R}^d$ onto itself.
Proof. Use Lemma 1-4 and Corollary 2-4. $\diamond$

Remark. Stern, [8] and A. N. Wilson Jr., [9] showed essentially that a continuously differentiable function $f: \mathbb{R}^d \to \mathbb{R}^d$ is a C^1-diffeomorphism if $Df(x)$ is uniformly row-sum dominant. Corollary 2-5 (b) was proved by Stern, [8], Ohtsuki & Watanabe, [10] and Kuh & Hajj, [11].

2. Newton-Raphson Method

The Newton-Raphson method is an attractive method of computing D. C. solutions because of its quadratic convergence under certain reasonable conditions. That is, if the initial point is sufficiently close to the exact D. C. solution, the (k+1)-th error is at least proportional to the square of the k-th error, [4], [12]. In this section, the guaranteed convergence region of the Newton-Raphson method is determined and the quadratic convergence is established again using the measure $\mu(\cdot)$. The effect of the local round-off error on the region of convergence and on the convergence is also investigated. For this problem we use a key result due to Hurt, [13].

Consider the D. C. equation (1.6):

$$f(x) = y \tag{1.6}$$

where $x, y \in \mathbb{R}^d$ and $f : \mathbb{R}^d \to \mathbb{R}^d$.

Throughout this section, we assume that

(Ai) $f : \mathbb{R}^d \to \mathbb{R}^d$ is in C^1,

(Aii) there exists a positive constant $m > 0$ such that either

$-\mu(\ Df(x)\ ) \doteq m > 0$ or $-\mu(\ -Df(x)\ ) \doteq m > 0$ for all $x \in \mathbb{R}^d$.

We note that the existence and uniqueness of the D. C. solution $x^* \in \mathbb{R}^d$ of eq. (1.6) is guaranteed and that $Df(x)$ is nonsingular for all $x \in \mathbb{R}^d$ by Corollary 2-4.

The Newton-Raphson method of solving the D. C. equation (1.6) with an infinite-precision machine is defined by the iteration rule

$$x_{k+1} = x_k - (\ Df(x_k)\ )^{-1} (\ f(x_k) - y\ ), \quad k = 1, 2, \cdots \tag{1.7}$$

with x_0 given.

<u>Definition.</u> Let $\{x_k\}_0^{\infty}$ be a sequence in $\mathbb{R}^d$ which converges to x^*. The sequence $\{x_k\}_0^{\infty}$ is said to converge to x^* <u>at least quadratically</u> iff there exist an integer $k_0 \geq 0$, and a constant c such that $|x_{k+1} - x^*| \leq c|x_k - x^*|^2$ for all $k \geq k_0$. (2.12)

<u>Theorem 2-6.</u> Consider the D. C. equation (1.6) with assumptions (Ai) and (Aii). Assume that there exists a continuous monotone increasing function $k^*(\cdot): \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $r > 0$

$$\|Df(u) - Df(v)\| \leq k^*(r)|u - v|, \quad \text{for all } u,v \in B(x^*,r). \tag{2.13}$$

Define r^* to be the unique solution of $r = 2m/k^*(r)$, $r > 0$. Under these conditions, if $x_0 \in B(x^*;r^*)$ then the corresponding sequence $\{x_k\}_{k=0}^{\infty}$ defined by eq. (1.7) remains in $B(x^*;r^*)$ and converges to the unique solution x^* at least quadratically.

<u>Proof.</u> Let an error vector e_k be defined by

$$e_k \triangleq x^* - x_k \quad \text{for all } k \in \mathbb{Z}_+. \tag{2.14}$$

Then, from eq. (1.7) and the definition of e_k, we obtain:

$$e_{k+1} \triangleq x^* - x_{k+1}$$

$$= x^* - \left\{ x_k - (Df(x_k))^{-1} (f(x_k) - y) \right\}$$

$$= (Df(x_k))^{-1} \left\{ Df(x_k)\cdot(x^* - x_k) + f(x_k) - f(x^*) \right\}$$

$$= (Df(x_k))^{-1} \left\{ Df(x_k)\cdot(x^* - x_k) + \int_0^1 Df(x^* + \tau(x_k - x^*))\cdot d\tau (x_k - x^*) \right\}.$$

Thus,

$$e_{k+1} = (Df(x^* - e_k))^{-1} \left\{ \int_0^1 (Df(x^* - e_k) - Df(x^* - \tau e_k))\cdot d\tau \cdot e_k \right\}. \tag{2.15}$$

Let $V(e) \triangleq |e|$, and $\Delta V(e_k) \triangleq V(e_{k+1}) - V(e_k)$ for all $k \in \mathbb{Z}_+$.

From eq. (2.15), we obtain

$$\Delta V(e) \leq \left| (Df(x^* - e))^{-1} \left\{ \int_0^1 (Df(x^* - e) - Df(x^* - \tau e))\cdot d\tau \cdot e \right\} \right| - |e|$$

$$\leq \left\| (Df(x^* - e))^{-1} \right\| \cdot \left| \int_0^1 (Df(x^* - e) - Df(x^* - \tau e))\cdot d\tau \cdot e \right| - |e| \tag{2.16}$$

The assumption (Aii) and Lemma 1-2, (ℓ) imply that

$$\left\| \left(Df(x^* - e) \right)^{-1} \right\| \leq 1/m \quad \text{for all } e \in \mathbb{R}^d. \tag{2.17}$$

Furthermore, if for any given $r > 0$ x^* and x^*-e are in $B(x^*; r)$, we obtain

$$\left| \int_0^1 \left(Df(x^* - e) - Df(x^* - \tau e) \right) d\tau \cdot e \right|$$

$$\leq \int_0^1 \left\| Df(x^* - e) - Df(x^* - \tau e) \right\| d\tau \cdot |e|$$

$$\leq \int_0^1 k^*(r) \left| (\tau - 1)e \right| d\tau \cdot |e|$$

$$= \frac{k^*(r)}{2} |e|^2. \tag{2.18}$$

Hence, for all $r > 0$ $\Delta V(e) \leq \frac{1}{m} \cdot \frac{k^*(r)}{2} |e|^2 - |e|$ for all $e \in B(\theta_d; r)$. (2.19)

In particular, $\Delta V(e) \leq \frac{1}{m} \cdot \frac{k^*(r^*)}{2} |e|^2 - |e|$

$$= \frac{k^*(r^*)}{2m} |e|^2 - |e|$$

$$= |e| \left(|e|/r^* - 1 \right) < 0 \quad \text{for all}$$

$0 \leq |e| < r^*$. (2.20)

Consider any sequence $\left\{e_k\right\}_0^\infty$ defined by eq. (2.15) with some initial condition $e_0 \in \mathbb{R}^d$ subject to $|e_0| \leq \gamma < r^*$ for some $\gamma > 0$. From eq. (2.20), we obtain

$$\left|e_{k+1}\right| \leq \frac{\left|e_k\right|^2}{r^*} \quad \text{for all } 0 \leq \left|e_k\right| \leq \gamma < r^*. \tag{2.21}$$

By induction, $0 \leq |e_k| \leq \gamma < r^*$ for all $k \in \mathbb{Z}_+$.

From eq. (2.21), we get

$$\left|e_k\right| \leq \left[\frac{\gamma}{r^*}\right]^{2k} r^* \quad \text{for all } k \in \mathbb{Z}_+. \tag{2.22}$$

So, the sequence $\left\{e_k\right\}_0^\infty$ converges to θ_d as $k \to \infty$, since $\gamma < r^*$. In terms of the iterates, eq. (2.21) is rewritten as

$$\left| x_{k+1} - x^*\right| \leq (1/r^*)\cdot\left|x_k - x^*\right|^2 \quad \text{for all } k \in \mathbb{Z}_+. \quad \diamond \tag{2.23}$$

<u>Remark.</u> Since $r^* = \max_{r>0} \min\left\{r, 2m/k^*(r)\right\}$, the open (2.24)

ball $B(x^*;r^*)$ is the best possible convergence region obtainable from eq. (2.13). If either m becomes large or if $f(\cdot)$ becomes smoother, i.e., $k^*(r)$ is decreased for each fixed $r > 0$, r^* becomes large by eq. (2.24); the region of convergence is enlarged. If $k^*(\cdot)$ is a constant function, r^* becomes $2m/k^*$, and the effect of m and k^* on the convergence region is obvious.

Since the D. C. solution $x^* \in \mathbb{R}^d$ is unknown a priori, conditions of Theorem 2-6, i.e., eq. (2.13) and $x_0 \in B(x^*;r^*)$, are impossible to check. Those conditions can be replaced by other stronger conditions which do not include the unknown x^*. The next corollary is stated in terms of a priori known quantities.

Corollary 2-7. Consider the D. C. equation (1.6) with assumptions (Ai) and (Aii). Assume that given $x_0 \in \mathbb{R}^d$, there exists a continuous monotone increasing function $k_0(\cdot): \mathbb{R}_+ \to \mathbb{R}_+$ such that for all $r > 0$

$$\| Df(u) - Df(v) \| \leq k_0(r) |u - v| \quad \text{for all } u,v \in B(x_0;r). \tag{2.25}$$

Define r^* to be the unique solution of

$$r = \frac{2m}{k_0\left(r + \frac{|f(x_0) - y|}{m} \right)}, \quad r > 0. \tag{2.26}$$

Under these conditions, if $|f(x_0) - y| \leq mr^*$, then the corresponding sequence $\{x_k\}_0^\infty$ defined by eq. (1.7) remains in $B(x^*;r^*)$ and converges to the unique solution x^* at least quadratically.

Proof. Claim: $|f(x) - y| \geq m|x - x^*|$ for all $x \in \mathbb{R}^d$.

$$\tag{2.27}$$

$|f(x) - y| = |f(x) - f(x^*)|$

$$= \left| \int_0^1 Df(\, x^* + \tau (x - x^*)\,)\, d\tau \cdot (x - x^*) \right| \qquad \text{by Taylor's}$$

formula.

$$\geq m \left| x - x^* \right| \qquad \text{by (2.6), (2.7) and Lemma 1-2, (j).}$$

Claim: the condition (2.25) implies the condition (2.13).

Let $r_0 \in \mathbb{R}_+$ be such that $r_0 > \left| x^* - x_0 \right|$, and define

$$r \triangleq r_0 - \left| x^* - x_0 \right| > 0. \tag{2.28}$$

Since $\left| x - x^* \right| < r$ implies that

$$\left| x - x_0 \right| \leq \left| x - x^* \right| + \left| x^* - x_0 \right| < r + \left| x^* - x_0 \right| = r_0,$$

we obtain the relation: $B(x^*;\, r) \subset B(x_0;\, r_0)$. (2.29)

From the condition (2.25), for all $r_0 > \left| x^* - x_0 \right|$,

$$\| Df(u) - Df(v) \| \leq k_0(r_0) |u - v| \quad \text{for all } u, v \in B(x_0;\, r_0) \tag{2.30}$$

Hence, for all $r > 0$,

$$\| Df(u) - Df(v) \| \leq k_0(r_0) |u - v| \quad \text{for all } u, v \in B(x^*;\, r)$$

$$\subset B(x_0;\, r_0). \tag{2.31}$$

Since $k_0(\cdot)$ is monotone increasing, we obtain for all $r > 0$,

$$\| Df(u) - Df(v) \| \leq k_0(\, r + \left| x^* - x_0 \right| \,) |u - v|$$

$$\leq k_0(r + |f(x_0) - y|/m)|u-v| \quad \text{for all } u,v \in B(x^*; r)$$

by (2.28) and (2.27). (2.32)

Let $k^*(r) \triangleq k_0(r + |f(x_0) - y|/m)$. Then, eq. (2.32) becomes the condition (2.13), since $k^*(\cdot)$ is continuous and monotone increasing. Also, note that by eq. (2.27),

$$\left\{ x \in \mathbb{R}^d \,\middle|\, |f(x) - y| \leq mr^* \right\} \subset B(x^*; r^*). \tag{2.33}$$

Then, Theorem 2-6 is applied to complete the proof. ◇

The Newton-Raphson method of solving the D. C. equation (1.6) with finite-precision machine gives:

$$x_{k+1} = x_k - (Df(x_k))^{-1}(f(x_k) - y) + \mathcal{E}(x_k), \quad k = 1, 2, \cdots \tag{2.34}$$

with x_0 given, where $\mathcal{E}(x_k)$ denotes the local round-off error incurred at (k+1)-th step. We assume that there exists an $\mathcal{E}_\infty > 0$ such that

$$|\mathcal{E}(x_k)| \leq \mathcal{E}_\infty \quad \text{for all } x_k \text{ generated by eq. (2.34).} \tag{2.35}$$

In order to discuss the effect of the local round-off error, we use a modified version of Hurt's corollaries, [13] .

Consider a difference equation:

$$\begin{cases} x_{k+1} = f(x_k) \\ x_0 \text{: given,} \end{cases} \tag{2.36}$$

where $x_k \in \mathbb{R}^d$ for all $k \in \mathbb{Z}_+$, and $f: \mathbb{R}^d \to \mathbb{R}^d$ is continuous.

Consequently, for all $x_0 \in \mathbb{R}^d$, the solution $\{x(k;x_0)\}_0^\infty$ of (2.36) is uniquely defined and for each fixed $k \in \mathbb{Z}_+$, the mapping $x_0 \mapsto x(k;x_0)$ is continuous.

Lemma 2-8. (Modified version of Hurt's corollaries, [13].)
Let V and W map $\mathbb{R}^d$ into $\mathbb{R}$, and let W be continuous. For some $\gamma > 0$, let $G \triangleq \{x \in \mathbb{R}^d \mid V(x) \leq \gamma\}$.

Assume further that

(i) $V(x) \geq 0$ for all $x \in G$;

(ii) G is compact;

(iii) there exists a constant $w \geq 0$ such that

$$\underset{(2.36)}{\Delta V(x)} \triangleq V(\, f(x)\,) - V(x) \leq -W(x) \leq w \quad \text{for all } x \in G; \tag{2.37}$$

(iv) Let $N \triangleq \{x \in G \mid W(x) \leq 0\}$ and $b \triangleq \sup_{x \in N} V(x) < \infty$;

(v) Let $A \triangleq \{x \in \mathbb{R}^d \mid V(x) \leq b + w\}$; $b + w < \gamma$.

Let $\delta = \inf_{x \in G-A} W(x)$.

Under these conditions,

(a) $N \subset A \subset G$, N is closed and $\delta \geq 0$.

(b) For all $x_0 \in G$, $x(k;x_0) \in G$ for all $k \in \mathbb{Z}_+$, i.e., G is an invariant set of eq. (2.36).

(c) For all $x_0 \in G$, $x(k;x_0) \to A$ as $k \to \infty$ and A is an invariant set of eq. (2.36). If, in addition $\delta > 0$, then there is a $k'(x_0)$ such that $x(k;x_0) \in A$ for all $k > k'(x_0)$.

(d) For all $x_0 \in G$, the positive limit set (set of all the limit points) $M(x_0)$ of the sequence $\{x(k;x_0)\}_0^\infty$ is a subset of A and $M(x_0)$ is an invariant set of eq. (2.36).

Proof. (a) If $x \in N$, then $V(x) \leq b \leq b + w < \gamma$, by assumptions (iv), (iii) and (v). Hence $N \subset A \subset G$.

Since $N = W^{-1}((-\infty, 0]) \cap G$ and W is continuous, N is closed as the intersection of two closed sets.

Since $W(x) > 0$ for all $x \in G-N$ and $G-A \subset G-N$,

$$\delta \triangleq \inf_{x \in G-A} W(x) \geq 0.$$

(b) Since $x_0 \in G$, it is enough to show that for all $i \in \mathbb{Z}_+$ $x(i;x_0) \in G$ implies $x(i+1;x_0) \in G$.

Case i) $x(i;x_0) \in G-N$.

By assumptions (iii) and (iv),

$$V(\, x(i+1;x_0)\,) \leq V(\, x(i;x_0)\,) - W(\, x(i;x_0)\,)$$

$$< V(\, x(i;x_0)\,) \leq \gamma.$$

So, $x(i+1;x_0) \in G$.

Case ii) $x(i;x_0) \in N$.

By assumptions (iii), (iv) and (v),

$$V(\, x(i+1;x_0)\,) \leq V(\, x(i;x_0)\,) - W(\, x(i;x_0)\,)$$

$$\leq b - W(\ x(i;x_0)\ ) \leq b + w < \gamma .$$

So, $x(i+1;x_0) \in A \subset G$.

Hence, G is an invariant set of eq. (2.36).

(c) Claim: A is an invariant set of eq. (2.36).

It is enough to show that $x \in A$ implies $f(x) \in A$. Similar to the proof of the invariance of G, if $x \in A-N$, then $f(x) \in A$ and if $x \in N$, then $f(x) \in A$.

Claim: for all $x_0 \in G$, $x(k;x_0) \rightarrow A$ as $k \rightarrow \infty$.

Let $d(\cdot,\cdot)$ be a distant function defined by

$$d(x,A) \triangleq \inf_{a \in A} |x-a| . \tag{2.38}$$

Proof is done by contradiction. Suppose not, i.e.,

$$\sim \left[\forall\ x_0 \in G \quad \forall\ \varepsilon > 0 \quad \exists N \quad \forall\ k \geq N \quad d(\ x(k;x_0),A\ ) \leq \varepsilon \right] . \tag{2.39}$$

That is,

$$\exists\ x_0' \in G \quad \exists \varepsilon' > 0 \quad \forall\ N \quad \exists k \geq N \quad d(\ x(k;x_0'),A\ ) > \varepsilon' . \tag{2.40}$$

Let J be the infinite set of integers defined as

$$\left\{ k \in \mathbb{Z}_+ \;\middle|\; d(\ x(k;x_0'),A\ ) > \varepsilon' \right\} .$$

Note that $\left\{ x(k;x_0') \right\}_{k \in J}$ is a subsequence of $x(\cdot;x_0')$ and that the sequence $x(\cdot;x_0')$ stays outside the set A, because A is invariant. Thus $\forall\ k \in J$, $x(k;x_0') \in G-B(A;\ \varepsilon') \subset G-A \subset G-N$.

Hence by assumptions (iii) and (iv), the subsequence $k \mapsto V(\ x(k;x_0')\ )$ is strictly monotone decreasing. Since $V \geqq 0$ on G, this subsequence converges. Hence, $\triangle V(\ x(k;x_0')\ ) \to 0$ as $k \to \infty$, $k \in J$, and so, $W(\ x(k;x_0')\ )$ tends to 0 as $k \to \infty$, $k \in J$, since $W(\ x(k;x_0')\ ) > 0$ for all $k \in J$.

Now,

$$\inf_{z \in G-B(A;\varepsilon)} W(z) = \min_{z \in G-B(A;\varepsilon)} W(z), \quad \text{since } G-B(A;\varepsilon) \text{ is}$$

compact,

$$\triangleq \varepsilon_w > 0 \ , \quad \text{since } W(z) > 0 \ \text{ for all}$$

$$z \in G-B(A;\varepsilon) \subset G-N. \tag{2.41}$$

So, $W(\ x(k;x_0')\ ) \geqq \varepsilon_w > 0$ for all $k \in J$. This is a contradiction.

Claim: if $\delta > 0$, then $\forall\, x_0 \in G\ \exists k'(x_0)$ such that $x(k;x_0) \in A$ $\forall\, k > k'(x_0)$.

Since A is an invariant set, it is enough to show that $\forall\, x_0 \in G$ $\exists k'(x_0)$ such that $x(\ k'(x_0);x_0\ ) \in A$.

Use contradiction. Suppose not, i.e.,

$$\exists x_0' \in G \ \forall\, k \in \mathbb{Z}_+ \quad x(k;x_0') \in G-A. \tag{2.42}$$

Then,

$$V(\ x(k;x_0')\ ) \leqq V(\ x(k-1;x_0')\ ) - W(\ x(k-1;x_0')\ )$$

$$\leq V(\ x(k-1;x_0')\ ) - \delta, \quad \text{by (iii) and (v).}$$

Thus,

$$V(\ x(k;x_0')\ ) \leq V(x_0') - k\delta. \tag{2.43}$$

So, $V(\ x(k;x_0')\ ) \to -\infty$ as $k \to \infty$. But, $\forall\ k \in \mathbb{Z}_+$ $V(\ x(k;x_0')\ ) > b + w \geq 0$ since $x(k;x_0') \in G-A$, $\forall\ k \in \mathbb{Z}_+$. This is a contradiction.

(d) Claim: $M(x_0)$ is an invariant set of eq. (2.36).

$M(x_0)$ is compact, since $M(x_0) \subset \bar{G} = G$ and $M(x_0)$ is closed. Let $p \in M(x_0)$. Then, there exists a convergent subsequence $\{x(k_n;\ x_0)\}_{n=0}^{\infty}$ such that $x(k_n;\ x_0) \to p$ as $n \to \infty$.

Define:

$$y_n(k) \triangleq x(k+k_n;\ x_0) \triangleq x(k+k_n) \quad \forall\ k \in \mathbb{Z}_+, \quad \forall\ n \in \mathbb{Z}_+. \tag{2.44}$$

Then, $y_n(\cdot)$ is the solution of eq. (2.36) with the initial condition $y_n(0) = x(k_n;\ x_0)$. Also, $y_n(0) \to p$ as $n \to \infty$.

Since the solution of eq. (2.36) is continuous with respect to x_0, $y_n(\cdot) \to x(\cdot;p)$ in the sense of pointwise convergence of sequences: $\forall\ k \in \mathbb{Z}_+$ $y_n(k) \to x(k;p)$ as $n \to \infty$.

Since $\forall\ n \in \mathbb{Z}_+$ $\forall\ k \in \mathbb{Z}_+$ $y_n(k) \in \mathbb{R}^d$ is on the sequence $x(\cdot;x_0)$, for fixed $k \in \mathbb{Z}_+$, $\{y_n(k)\}_{n=1}^{\infty}$ is a subsequence of

$x(\cdot;x_0)$, such that $y_n(k) \to x(k;p)$ as $n \to \infty$. So, $x(k;p) \in M(x_0)$ $\forall k \in \mathbb{Z}_+$, i.e., $M(x_0)$ is invariant under eq. (2.36).

Claim: $M(x_0) \subset A$.

By the definition of $M(x_0)$, $\forall\, p \in M(x_0)$ $\exists$ a subsequence $S(p)$ of $x(\cdot;x_0)$ which tends to p. We have shown that the sequence $x(\cdot;x_0)$ tends to A. Since $S(p)$ is a subsequence of $x(\cdot;x_0)$, $S(p)$ also tends to A. Hence $p \in A$. $\diamond$

<u>Remark 1</u>. V is said to be a Lyapunov function. In the above Lemma, V takes on nonnegative values on G and is bounded from below on G. The continuity of V is not required, and V can possibly increase on N along the solution sequence.

<u>Remark 2</u>. We note that Lemma 2-8 can be used to prove Theorem 2-6. By letting $V(e) \triangleq |e|$, $W(e) \triangleq -(|e|/r^* - 1)|e|$, and $0 < \gamma < r^*$, we obtain $A = N = \{\theta_d\}$ and $G = B(\theta_d;r^*)$, using eq. (2.15).

Now, we state the theorem concerning the effect of the local round-off error.

<u>Theorem 2-9</u>. Consider the D. C. equation (1.6) with assumptions (Ai) and (Aii). Assume that f satisfies the condition (2.13) of Theorem 2-6. Assume further that the local round-off error $\varepsilon(x_k)$ is bounded as in (2.35) and that

$$\varepsilon_\infty < r^*/5. \qquad (2.45)$$

Under these conditions, if $x_0 \in \bar{B}(x^*; r^* - 2\mathcal{E}_\infty)$, then the corresponding sequence $\{x_k\}_{k=0}^{\infty}$ defined by eq. (2.34) remains in $\bar{B}(x^*; r^* - 2\mathcal{E}_\infty)$ and enters the region $\bar{B}(x^*; 3\mathcal{E}_\infty)$ after a finite number of steps and remains in it forever after.

<u>Proof.</u> From eq. (2.34), we can derive a difference equation analogous to (2.15):

$$e_{k+1} = (\, Df(x^* - e_k)\,)^{-1}\left\{ \int_0^1 (\, Df(x^* - e_k) - Df(x^* - \tau e_k)\,)\, d\tau \cdot e_k \right\}$$

$$+ \mathcal{E}(x^* - e_k), \quad k = 1, 2, \cdots \tag{2.46}$$

Let $V(e) = |e|$. As we obtained eq. (2.19), we get: for all $r > 0$

$$\Delta V(e) \leq |e|(\, k^*(r)/2m \cdot |e| - 1\,) + \left|\mathcal{E}(x^* - e_k)\right|$$

$$\leq |e|(\, k^*(r)/2m \cdot |e| - 1\,) + \mathcal{E}_\infty, \text{ for all } e \in B(\theta_d; r). \tag{2.47}$$

Corresponding to eq. (2.20), we obtain:

$$\Delta V(e) \leq |e|(\,|e|/r^* - 1\,) + \mathcal{E}_\infty, \quad \text{for all } 0 \leq |e| < r^* \tag{2.48}$$

In order to apply Lemma 2-8, let $-W(e)$ be the right-hand side of eq. (2.48):

$$W(e) \triangleq -|e|(\,|e|/r^* - 1\,) - \mathcal{E}_\infty. \tag{2.49}$$

Observe that W is continuous and $w = \mathcal{E}_\infty$. Choose $\gamma = r^* - 2\mathcal{E}_\infty < r^*$. Hence we obtain

$$G = \left\{ e \in \mathbb{R}^d \,\middle|\, |e| \leq r^* - 2\mathcal{E}_\infty \right\} = \bar{B}(\theta_d;\ r^* - 2\mathcal{E}_\infty). \tag{2.50}$$

Check all the conditions of Lemma 2-8.

(i) $V(e) = |e| \geq 0$ for all $e \in G \subset \mathbb{R}^d$.

(ii) $G = \bar{B}(\theta_d;\ r^* - 2\mathcal{E}_\infty)$ is compact.

(iii) Let $w = \mathcal{E}_\infty$. Then,

$$\triangle V(e)_{(2.46)} \leq -W(e) \leq \mathcal{E}_\infty \quad \text{for all } e \in \bar{B}(\theta_d;\ r^* - 2\mathcal{E}_\infty). \tag{2.51}$$

(iv) $$N \triangleq \left\{ e \in G \,\middle|\, W(e) \leq 0 \right\} = \left\{ e \in \mathbb{R}^d \,\middle|\, |e| \leq b \right\} = \bar{B}(\theta_d;\ b) \tag{2.52}$$

where b is the smallest zero of

$$W(e) = -|e|^2/r^* + |e| - \mathcal{E}_\infty = 0 \tag{2.53}$$

Therefore,

$$b = \frac{-1 + \sqrt{1 - 4\mathcal{E}_\infty/r^*}}{-2/r^*} = \frac{1 - \sqrt{1 - 4\mathcal{E}_\infty/r^*}}{2/r^*}$$

$$= \mathcal{E}_\infty + \mathcal{E}_\infty^2/r^* + \cdots, \quad \text{since } 4\mathcal{E}_\infty/r^* < 1. \tag{2.54}$$

Since $W(e)\Big|_{|e| = \mathcal{E}_\infty} = -\mathcal{E}_\infty^2/r^* < 0$ and $W(e)\Big|_{|e| = 2\mathcal{E}_\infty} =$

$$\mathcal{E}_\infty(-4\mathcal{E}_\infty/r^* + 1) > 0, \quad \mathcal{E}_\infty < b < 2\mathcal{E}_\infty < \infty. \tag{2.55}$$

(v) $$A \triangleq \left\{ e \in \mathbb{R}^d \,\middle|\, V(e) \leq b + \mathcal{E}_\infty \right\} = \bar{B}(\theta_d;\ b + \mathcal{E}_\infty). \tag{2.56}$$

$$b + \varepsilon_\infty < 3\,\varepsilon_\infty < r^* - 2\,\varepsilon_\infty \quad \text{by (2.55) and (2.45).} \tag{2.57}$$

From eq. (2.57), we obtain

$$A \subset \bar{B}(\theta_d;\ 3\varepsilon_\infty) \subset G. \tag{2.58}$$

Note that

$$\delta \triangleq \inf_{e \in G-A} W(e) = \inf_{b+\varepsilon_\infty < |e| \leq r^*-2\varepsilon_\infty} W(e)$$

$$= \min\left\{ W(e)\Big|_{|e|=b+\varepsilon_\infty},\ W(e)\Big|_{|e|=r^*-2\varepsilon_\infty} \right\} > 0. \tag{2.59}$$

Hence, all the conditions of Lemma 2-8 are satisfied and, consequently, the conclusion of the theorem follows. ◇

Remark. Theorem 2-9 shows that if the local round-off error ε_∞ is sufficiently small, then the radius of the convergence region is $2\varepsilon_\infty$ smaller than that of the infinite precision arithmetic case, and instead of quadratic convergence to the unique solution x^*, we obtain the convergence to a ball centered on x^* with a radius $3\varepsilon_\infty$ in a finite number of steps.

Corresponding to Corollary 2-7, the following corollary which is stated using only a priori known quantities is obtained from Theorem 2-9 in a similar manner.

Corollary 2-10. Assume that f satisfies all the conditions of Corollary 2-7. Assume that the local round-off error $\varepsilon(x_k)$ is bounded as in (2.35) and that $\varepsilon_\infty < r^*/5$.

Under these conditions, if $|f(x_0) - y| \leq m(r^* - 2\varepsilon_\infty)$, then the corresponding sequence $\{x_k\}_0^\infty$ defined by eq. (2.34) remains in $\bar{B}(x^*; r^* - 2\varepsilon_\infty)$ and enters the region $\bar{B}(x^*; 3\varepsilon_\infty)$ after a finite number of steps and remains in it forever after.

Proof. Using the same techniques for proving Corollary 2-7, we can show that Corollary 2-10 is the special case of Theorem 2-9. $\diamond$

It is worthwile to note that error estimate is obtained by eq. (2.27). Let $\tilde{x} \in \mathbb{R}^d$ be a computed point. Then, the estimate of the error $x^* - \tilde{x}$ is given by:

$$|x^* - x| \leq |f(x) - y|/m. \qquad (2.60)$$

CHAPTER III.

ORDINARY DIFFERENTIAL EQUATIONS

In this chapter, the upper and lower bounds of the solution of O. D. E.'s are estimated using the measure $\mu(\cdot)$.

1. Estimates for Upper Bounds on Solutions

We consider nonlinear time-varying O. D. E.'s of the form:

$$\begin{cases} \dot{x} = f(x,t) + u(t) \\ x(0) = x_0 \end{cases} \tag{1.1}$$

where $x(t)$, $u(t) \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$, and $f: \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$. We assume A1: $f(\theta_d, t) = \theta_d$ for all $t \in \mathbb{R}_+$; A2: $x \mapsto f(x,t)$ is in C^1 for all $t \in \mathbb{R}_+$; and A3: the input $u(\cdot)$ and for each fixed $x \in \mathbb{R}^d$ $t \mapsto f(x,t)$ are piecewise continuous on $\mathbb{R}_+$. We say that a function from $\mathbb{R}_+$ into $\mathbb{R}^d$ is piecewise continuous iff on every compact interval $J = [t_0, t_1] \subset \mathbb{R}_+$ (i) the function is continuous on J except for at most a finite number of points; (ii) if $t' \in (t_0, t_1)$ is a point of discontinuity, then the right- and left-hand limits of the function exist and are finite; and (iii) at $t = t_0$ the right-hand limit exists and at $t = t_1$ the left-hand limit exists, [5], [23].

We utilize Coppel's theorem for estimating those bounds. Coppel's theorem gives the upper and lower bounds for linear time-varying O. D. E.'s, where the measure $\mu(\cdot)$ was originally used for the stability analysis of O. D. E.'s, Dahlquist [1], Coppel [2].

<u>Lemma 3-1</u>. Slightly generalized version of Coppel's inequality [2].

Let $A(\cdot): \mathbb{R}_+ \to \mathbb{R}^{d\times d}$ be piecewise continuous. Let $\Phi(t,t_0)$ be the state transition matrix associated with $A(\cdot)$, i.e., by definition:

$$\begin{cases} \dfrac{\partial}{\partial t}\Phi(t,t_0) = A(t)\Phi(t,t_0) \\ \Phi(t_0,t_0) = I \end{cases} \tag{3.1}$$

for all $t \geq t_0 \geq 0$.

Then,

$$\exp\left[-\int_{t_0}^{t}\mu(\,-A(\tau)\,)\,d\tau\right] \leq 1/\left\|\left[\Phi(t,t_0)\right]^{-1}\right\|$$

$$\leq \left\|\Phi(t,t_0)\right\| \leq \exp\left[\int_{t_0}^{t}\mu(\,A(\tau)\,)\,d\tau\right] \tag{3.2}$$

for all $t \geq t_0 \geq 0$.

<u>Proof</u>. Consider a linear time-varying O. D. E.:

$$\begin{cases} \dot{x} = A(t)\cdot x \\ x(t_0) = x_0 \end{cases} \tag{3.3}$$

where $t \geq t_0$ and $t_0 \in \mathbb{R}_+$.

Since $A(\cdot)$ is piecewise continuous, letting D be an atmost denumerable subset of $\mathbb{R}_+$ where for all $t' \in D$ there exists some pair (i,j), $i,j \in \{1,2, \cdots ,d\}$ such that $a_{ij}(\cdot)$ is discontinuous at t'. The solution $x(\cdot)$ of (3.3) is by definition a <u>continuous</u> function: $\mathbb{R}_+ \to \mathbb{R}^d$ such that (3.3) holds in $\mathbb{R}_+ - D$. The inequalities (3.2) will follow if we show that for $t \geq t_0$ and for all $x_0 \neq \theta$

$$\exp\left[-\int_{t_0}^{t} \mu(-A(\tau))\,d\tau\right] |x_0| \leq |x(t)| \leq$$

$$\exp\left[\int_{t_0}^{t} \mu(A(\tau))\,d\tau\right] |x_0|. \tag{3.4}$$

This is easily seen by taking the infimum and supremum over $x_0 \neq \theta$. We first observe that $\mu(A(\cdot))$ is piecewise continuous, since $\mu(\cdot)$ is continuous and $A(\cdot)$ is piecewise continuous.

Claim: the right-hand derivative $|x(t)|^{\bullet}_{+}$ of the norm $|x(\cdot)|$ of any solution of (3.3) exists for all $t \in \mathbb{R}_+$, and

$$|x(t)|^{\bullet}_{+} = \lim_{h \downarrow 0+} \frac{|x(t) + h\dot{x}(t+0)| - |x(t)|}{h} \tag{3.5}$$

Observe that from (3.3) the right-hand derivative $\dot{x}(t+0)$ of $x(\cdot)$ at t exists for all $t \in \mathbb{R}_+$.

Let $0 < \theta < 1$. Then we have

$$|x(t) + \theta h\dot{x}(t+0)| = |\theta\cdot(x(t) + h\dot{x}(t+0)) + (1-\theta)\cdot x(t)|$$

$$\leq \theta|x(t) + h\dot{x}(t+0)| + (1-\theta)|x(t)| \tag{3.6}$$

or

$$\frac{|x(t) + \theta h\dot{x}(t+0)| - |x(t)|}{\theta h} \leq \frac{|x(t) + h\dot{x}(t+0)| - |x(t)|}{h} \tag{3.7}$$

Since $h \longmapsto \dfrac{|x(t) + h\dot{x}(t+0)| - |x(t)|}{h}$ is nondecreasing and it is bounded from below by $-|\dot{x}(t+0)|$, the limit in (3.5) is finite. We now establish equality (3.5). For sufficiently small $h > 0$,

$$\left| \, |x(t)|^{\cdot}_{+} - \frac{|x(t) + h\dot{x}(t+0)| - |x(t)|}{h} \right|$$

$$= \left| \frac{|x(t+h)| - |x(t)|}{h} + o(h)/h - \frac{|x(t) + h\dot{x}(t+0)| - |x(t)|}{h} \right|$$

$$= \left| \frac{|x(t+h)| - |x(t) + h\dot{x}(t+0)| + o(h)}{h} \right|$$

$$\leq \frac{1}{h} \left| x(t+h) - x(t) - h\dot{x}(t+0) + o(h) \right|$$

$$= \frac{1}{h} \left| \int_{t}^{t+h} A(t')\, x(t')\, dt' - h\dot{x}(t+0) + o(h) \right| \tag{3.8}$$

Since for sufficiently small $h > 0$ $A(\cdot)$ is continuous in $(t, t+h]$, the integral is $A(t+0)x(t)h + o(h)$. Therefore, the left-hand side of (3.8) is equal to $o(h)/h$. Hence, (3.5) follows.

Since, $|x(t) + h\dot{x}(t+0)| - |x(t)|$

$$\leq \|I + hA(t+0)\| \cdot |x(t)| - |x(t)|, \tag{3.9}$$

$$|x(t)|^{\bullet}_{+} \leq \lim_{h \downarrow 0+} \frac{\| I + hA(t+0) \| - 1}{h} |x(t)|$$

$$= \mu(\ A(t+0)\ )|x(t)| \quad \text{for all } t \in \mathbb{R}_{+}. \tag{3.10}$$

Let $w(t) \triangleq \exp\left[- \int_{t_0}^{t} \mu(\ A(\tau)\ )\, d\tau \right] \cdot |x(t)|$ for all

$$t \in \mathbb{R}_{+}. \tag{3.11}$$

Since $\mu(\ A(\cdot)\ )$ is piece-wise continuous, the set D is a set of measure zero and

$$w(t) = \exp\left[- \int_{t_0}^{t} \mu(\ A(\tau+0)\, d\tau \right] \cdot |x(t)| \quad \text{for all}$$

$$t \in \mathbb{R}_{+}. \tag{3.12}$$

By eq.(3.10) and eq.(3.12),

$$\dot{w}_{+}(t) = \exp\left[- \int_{t_0}^{t} \mu(\ A(\tau+0)\ )\, d\tau \right] \cdot \Big\{ -\mu(\ A(t+0)\ )|x(t)|$$

$$+ |x(t)|^{\bullet}_{+} \Big\} \leq 0 \quad \text{for all } t \in \mathbb{R}_{+}. \tag{3.13}$$

Hence w(t) is monotone decreasing, [2] and then

$$\exp\left[- \int_{t_0}^{t} \mu(\ A(\tau)\ )\, d\tau \right] |x(t)| = w(t) \leq w(t_0)$$

$$= |x_0|, \tag{3.14}$$

or

$$|x(t)| \leq \exp\left[\int_{t_0}^{t} \mu(A(\tau))\, d\tau \right] |x_0| \quad \text{for all } t \geq t_0. \tag{3.15}$$

The proof of the other part of the inequality (3.4) is analogous to the previous one and uses left-hand derivatives. We also obtain that the left-hand derivative $|x(t)|^{\bullet}_{-}$ of $|x(t)|$ exists for all $t \in \mathbb{R}_+$, and

$$|x(t)|^{\bullet}_{-} = \lim_{h \downarrow 0+} \frac{|x(t)| - |x(t) - h\dot{x}(t-0)|}{h}. \tag{3.16}$$

We also obtain:

$$|x(t)|^{\bullet}_{-} \geq -\mu(A(t-0))|x(t)| \quad \text{for all } t \in \mathbb{R}_+. \tag{3.17}$$

Let $w(t) \triangleq \exp\left[\int_{t_0}^{t} \mu(-A(\tau))\, d\tau \right] \cdot |x(t)|$ for all $t \in \mathbb{R}_+$. Then, it is easily verified that $\dot{w}_-(t) \geq 0$ for all $t \in \mathbb{R}_+$. Hence, we obtain:

$$|x(t)| \geq \exp\left[-\int_{t_0}^{t} \mu(-A(\tau))\, d\tau \right] \cdot |x_0| \quad \text{for all}$$

$$t \geq t_0. \quad \diamond \tag{3.18}$$

<u>Comment</u>. The following calculation gives insight to the meaning of $\mu(\cdot)$ and its relation to the solution of O. D. E.'s. This was suggested by Prof. W. Kahan. For simplicity, let $A(\cdot)$ be continuous.

Define $y(t) \triangleq \exp(\sigma t)x(t)$ for all $t \geq t_0$ (3.19)

where $\sigma > 0$.

Then, $\dot{y} = (\sigma I + A(t))y$ with $y(t_0) = x_0$. (3.20)

Claim: $|y(t)|^{\bullet}_{+} \leq |\dot{y}(t)|$ for all $t \in \mathbb{R}_+$, (3.21)

where $|y(t)|^{\bullet}_{+}$ is the right-hand derivative of $|y(\cdot)|$.

Observe that for all $dt > 0$,

$$\frac{|y(t+dt)| - |y(t)|}{dt} = \frac{|y(t) + \dot{y}(t)dt + o(dt)| - |y(t)|}{dt}$$

$$\leq |\dot{y}(t)| + o(dt)/dt. \tag{3.22}$$

From (3.20) and (3.21), we obtain:

$$|y(t)|^{\bullet}_{+} \leq \|\sigma I + A(t)\| \cdot |y(t)| \quad \text{with } |y(t_0)| = \tag{3.23}$$

$|x_0|$ for all $t \geq t_0$.

In terms of $|x(t)|$, eq.(3.23) becomes:

$$|x(t)|^{\bullet}_{+} \leq \left[\|\sigma I + A(t)\| - \sigma \right] \cdot |x(t)|, \quad \text{with} \tag{3.24}$$

$|x(t_0)| = |x_0|$ for all $t \geq t_0$.

Let $\theta = 1/\sigma$ and let $\sigma \to +\infty$, and then

$$|x(t)|^{\bullet}_{+} \leq \lim_{\theta \downarrow 0+} \frac{\|I + \theta A(t)\| - 1}{\theta} |x(t)| = \mu(A(t)) \cdot |x(t)|.$$

Hence, $|x(t)| \leq \exp\left[\int_{t_0}^{t} \mu(A(\tau))\, d\tau \right] \cdot |x_0|$. (3.25)

Thus, $\|\Phi(t,t_0)\| = \sup_{x_0 \neq \theta} \dfrac{|x(t;x_0)|}{|x_0|}$

$$\leq \exp\left[\int_{t_0}^{t} \mu(\, A(\tau)\,)\, d\tau \right]. \tag{3.26}$$

Similarly, by using the left-hand derivative $|y(t)|^{\bullet}_{-}$, we obtain:

$$\exp\left[-\int_{t_0}^{t} \mu(\, -A(\tau)\,)\, d\tau \right] \leq \inf_{x_0 \neq \theta_d} \frac{|x(t;x_0)|}{|x_0|}$$

$$= 1/ \left\| \left[\Phi(t,t_0)\right]^{-1} \right\| . \tag{3.27}$$

Recall that we defined the class $\mathcal{M}(\varepsilon)$ of functions by (2.10). That is, a function $m(\cdot): \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ is said to be in $\mathcal{M}(\varepsilon)$ iff $m(\alpha) > 0$ for all $\alpha \in \mathbb{R}_+$ and there exists a positive constant $\varepsilon > 0$ such that

$$\int_0^{\alpha} m(\xi)\, d\xi \geq \varepsilon\alpha \quad \text{for all } \alpha \in \mathbb{R}_+. \tag{2.10}$$

Theorem 3-2. Dahlquist, [1].

Consider the O. D. E. (1.1) with assumptions A1, A2 and A3. Assume that there exists an $m(\cdot) \in \mathcal{M}(\varepsilon)$ such that

$$-\mu[D_1 f(x,t)] \geq m(|x|) > 0 \quad \text{for all } x \in \mathbb{R}^d, \tag{3.28}$$

for all $t \in \mathbb{R}_+$.

Under these conditions, the solution $x(\cdot;x_0)$ of eq.(1.1) with an initial condition $x_0 \in \mathbb{R}^d$ satisfies:

$$|x(t)| \le \exp(-\varepsilon t)\cdot |x_0| + \int_0^t \exp\left[-\varepsilon(t-\tau)\right]\cdot|u(\tau)|\, d\tau \quad (3.29)$$

for all $t \in \mathbb{R}_+$.

Proof. Since the solution $x(\cdot;x_0)$ of eq.(1.1) exists and is unique, $x(\cdot;x_0)$ is equal to the solution of the following linear time varying differential equation:

$$\begin{cases} \dot{x} = A(t)x + u(t) \\ x(0) = x_0 \end{cases} \quad (3.30)$$

$$\text{where } A(t) = \int_0^1 D_1 f(\tau x, t)\, d\tau \quad \text{for all } t \in \mathbb{R}_+. \quad (3.31)$$

Here, we used the Taylor formula:

$$f(x,t) = f(\theta,t) + \int_0^1 D_1 f(\tau x,t)\, d\tau \cdot x$$

$$= \int_0^1 D_1 f(\tau x,t)\, d\tau \cdot x \quad \text{for all } t \in \mathbb{R}_+. \quad (3.32)$$

We note that $A(\cdot)$ is piecewise continuous.

Claim: $\mu(\, A(t)\,) \le -\varepsilon$ for all $t \in \mathbb{R}_+$.

$$\mu(\, A(t)\,) = \mu\left[\int_0^1 D_1 f(\tau x,t)\, d\tau\right]$$

$$\le \int_0^1 \mu\left[D_1 f(\tau x,t)\right] d\tau, \quad \text{by Lemma 1-2, (d) \& (f)}$$

$$\leq \int_0^1 -m(|\tau x|)\, d\tau$$

$$= \int_0^{|x|} -\frac{m(\alpha)}{|x|} d\alpha \quad \text{by letting } \alpha = |\tau x| = \tau|x|$$

$$\leq -\varepsilon < 0 \text{ , since } m(\cdot) \in \mathcal{M}(\varepsilon). \tag{3.33}$$

By Lemma 3-1, we obtain:

$$\|\Phi(t,t_0)\| \leq \exp\left[-\varepsilon(t-t_0)\right] \text{ , for all } t \geq t_0, \tag{3.34}$$

$t, t_0 \in \mathbb{R}_+$.

Thus the inequality (3.29) follows. ◇

Remark. The inequality (3.29) shows that if the input $u(\cdot)$ is bounded on $[0,\infty)$ and if $u(t) \to \theta_d$ as $t \to \infty$, then starting from any initial condition $x_0 \in \mathbb{R}^d$, $x(t;x_0) \to \theta_d$ as $t \to \infty$.

Since a constant function m is in $\mathcal{M}(m)$, the following corollary follows immediately.

Corollary 3-3. Consider the O. D. E. (1.1) satisfying A1, A2 and A3. Assume that there exists a positive constant $m > 0$ such that

$$-\mu\left[D_1 f(x,t)\right] \geq m > 0 \quad \text{for all } x \in \mathbb{R}^d, \quad \text{for all} \tag{3.35}$$

$t \in \mathbb{R}_+$.

Then, the solution $x(\cdot;x_0)$ of eq.(1.1) satisfies:

$$|x(t)| \leq \exp(-mt)\cdot|x_0| + \int_0^t \exp[-m(t-\tau)]\cdot|u(\tau)|d\tau,$$

for all $t \in \mathbb{R}_+$. $\diamond$ (3.36)

Relation to previous work. The special case under ℓ^2 norm for Corollary 3-3 is classical. The ℓ^1 norm case was studied by Rosenbrock [14], and the modification was done by Sandberg [15] and Mitra & So [16], where $|x| = |Dx|_1$, with positive diagonal dxd matrix $D > 0$.

Theorem 3-4. Consider the O. D. E. (1.1). Assume all the conditions of Corollary 3-3 are satisfied. Let $x_a(\cdot)$ & $x_b(\cdot)$ be solutions of eq.(1.1) with initial conditions $x_a(0)$ & $x_b(0)$, due to inputs $u_a(\cdot)$ and $u_b(\cdot)$, respectively.
Under these conditions, the difference $x_a(\cdot) - x_b(\cdot)$ of the two solutions satisfies:

$$|x_a(t) - x_b(t)| \leq \exp(-mt)\cdot|x_a(0) - x_b(0)|$$

$$+ \int_0^t \exp[-m(t-\tau)]\cdot|u_a(\tau) - u_b(\tau)|d\tau$$

for all $t \in \mathbb{R}_+$. (3.37)

Proof. Note that:

$$\dot{x}_a = f(x_a,t) + u_a(t) \quad \text{for all } t \in \mathbb{R}_+. \qquad (3.38)$$

and that

$$\dot{x}_b = f(x_b, t) + u_b(t) \quad \text{for all } t \in \mathbb{R}_+. \tag{3.39}$$

By subtracting eq.(3.39) from eq.(3.38), we obtain

$$\frac{d}{dt}\left[x_a(t) - x_b(t)\right] = f(x_a, t) - f(x_b, t) + \left[u_a(t) - u_b(t)\right]$$

$$= \int_0^1 D_1 f\left[x_b + \tau(x_a - x_b), t\right] d\tau \cdot (x_a - x_b) + \left[u_a(t) - u_b(t)\right]$$

for all $t \in \mathbb{R}_+$. (3.40)

Observe that:

$$\mu\left[\int_0^1 D_1 f\left[x_b + \tau(x_a - x_b), t\right] d\tau\right]$$

$\leq \int_0^1 \mu\left[D_1 f\left[x_b + \tau(x_a - x_b), t\right]\right] d\tau$ by Lemma 1-2, (d) &

(f)

$\leq -m < 0$ for all $t \in \mathbb{R}_+$. (3.41)

Similarly to the proof of Theorem 3-2, we obtain the inequality (3.37). $\diamond$

Remark. As before, the inequality (3.37) shows that if the difference $u_a(\cdot)-u_b(\cdot)$ of the two inputs is bounded on $[0,\infty)$ and converges to θ_d as $t \to \infty$, then starting from any two initial conditions $x_a(0)$ & $x_b(0)$, the difference $x_a(\cdot) - x_b(\cdot)$ of the two solutions converges to θ_d as $t \to \infty$. This guarantees a unique steady state solution for broad classes of electric circuits.

Corollary 3-5. Consider the O. D. E. (1.1). Assume all the conditions of Theorem 3-4 are satisfied. Let $x(\cdot;x_0)$ be the solution of eq.(1.1) with the initial condition $x_0 \in \mathbb{R}^d$ due to a constant input $u_\infty \in \mathbb{R}^d$. Let $x_\infty \in \mathbb{R}^d$ be the D. C. solution of

$$\theta_d = f(x) + u_\infty \tag{3.42}$$

Under these conditions, the difference $x(t) - x_\infty$ satisfies:

$$|x(t) - x_\infty| \leq \exp(-mt)\cdot|x_0 - x_\infty| \quad \text{for all } t \in \mathbb{R}_+. \tag{3.43}$$

Proof. In view of Corollary 2-4, the D. C. solution of eq. (3.42) exists and is unique. Then, the inequality (3.43) is the immediate consequence of Theorem 3-4. $\diamond$

Relation to previous work. The special cases under the weighted ℓ^1 norm, i.e., $|x| \triangleq |Dx|_1$ & $D > 0$ is diagonal, for Theorem 3-4 and Corollary 3-5 were proved by Sandberg [15] and Mitra & So [16].

2. Estimates for Lower Bounds on Solutions

Using the other half of Coppel's inequality (3.2), we can state theorems corresponding to those of Section 1, giving estimates for lower bounds on solutions.

<u>Theorem 3-6</u>. Consider the O. D. E. (1.1) with assumptions A1, A2 and A3. Assume that there exists an $m(\cdot) \in \mathcal{M}(\varepsilon)$ such that

$-\mu[-D_1 f(x,t)] \geq -m(\,|x|\,)$ for all $x \in \mathbb{R}^d$, for all

$t \in \mathbb{R}_+$, (3.44)

and that $u(t) \equiv \theta_d$ for all $t \in \mathbb{R}_+$.

Under these conditions, the solution $x(\cdot;x_0)$ of eq.(1.1) with an initial condition $x_0 \in \mathbb{R}^d$ satisfies:

$$|x(t)| \geq \exp(-\varepsilon t)\,|x_0| \quad \text{for all } t \in \mathbb{R}_+. \qquad (3.45)$$

<u>Proof</u>. The proof is analogous to that of Theorem 3-2. Let $A(\cdot)$ be defined as in (3.16).

Observe that:

$$-\mu[-A(t)] = -\mu\left[-\int_0^1 D_1 f(\tau x,t)\, d\tau\right]$$

$$\geq -\int_0^1 \mu\left[-D_1 f(\tau x,t)\right] d\tau \quad \text{by Lemma 1-2, (d) \&}$$

(f)

$$\geq -\int_0^{|x|} \frac{m(\sigma)}{|x|}\, d\sigma \quad \text{by letting } \sigma = |\tau x|$$

$$\geq -\varepsilon. \qquad (3.46)$$

Then, Lemma 3-1 is applied to obtain:

$$\exp\left[-\varepsilon(t-t_0)\right] \leq 1/\left\|\left[\Phi(t,t_0)\right]^{-1}\right\|, \text{ for all } t \geq t_0,$$

$$t, t_0 \in \mathbb{R}_+. \tag{3.47}$$

Hence, the inequality (3.45) follows. ◇

Corollary 3-7. Consider the O. D. E. (1.1) with assumptions A1, A2 and A3. Assume that there exists a positive constant $m > 0$ such that:

$$-\mu\left[-D_1 f(x,t)\right] \geq -m \quad \text{for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+,$$

and that $u(t) \equiv \theta_d$ for all $t \in \mathbb{R}_+$.

Under these conditions, the solution $x(\cdot;x_0)$ of eq.(1.1) with an initial condition $x_0 \in \mathbb{R}^d$ satisfies:

$$|x(t)| \geq \exp(-mt)|x_0|. \quad ◇ \tag{3.48}$$

Theorem 3-8. Consider the O. D. E. (1.1) with assumptions A1, A2 and A3. Assume that there exists a positive constant $m > 0$ such that

$$-\mu\left[-D_1 f(x,t)\right] \geq -m \quad \text{for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+.$$

Let $x_a(\cdot)$ & $x_b(\cdot)$ be solutions of eq.(1.1) with initial conditions $x_a(0)$ & $x_b(0)$, due to the same input $u(\cdot) \equiv u_a(\cdot) \equiv u_b(\cdot)$, respectively.

Under these conditions, the difference $x_a(\cdot) - x_b(\cdot)$ of the two solutions satisfies:

$$|x_a(t) - x_b(t)| \doteq \exp(-mt)\cdot|x_a(0) - x_b(0)| \quad \text{for all}$$

$$t \in \mathbb{R}_+. \quad \diamond \tag{3.49}$$

Corollary 3-9. Consider eq.(1.1). Assume that all the conditions of Theorem 3-8 are satisfied. Let $x(\cdot;x_0)$ be the solution of eq.(1.1) with the initial condition $x_0 \in \mathbb{R}^d$ due to a constant input $u_\infty \in \mathbb{R}^d$. Let $x_\infty \in \mathbb{R}^d$ be the D. C. solution of eq.(3.42).

Under these conditions, the difference $x(\cdot) - x_\infty$ satisfies:

$$|x(t) - x_\infty| \doteq \exp(-mt)|x_0 - x_\infty| \quad \text{for all } t \in \mathbb{R}_+. \tag{3.50}$$

Relation to previous work. If we take ℓ^1 norm, Theorem 3-8 and Corollary 3-9 are led to the results by Sandberg [15].

Remark. In this chapter, the estimate of lower and upper bounds is stated only for exponentially stable case: there exist positive constants m_{max} and m_{min} such that

$$-m_{max} \leq -\mu[-D_1f(x,t)] \leq \mu[D_1f(x,t)] \leq -m_{min} \quad \text{for all}$$

$$x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+. \tag{3.51}$$

Using the same technique, it is easy to show the similar estimates for exponentially unstable cases: there exist positive constants m_{max} and m_{min} such that

$$m_{min} \leq -\mu[-D_1f(x,t)] \leq \mu[D_1f(x,t)] \leq m_{max} \quad \text{for all}$$

$$x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+. \tag{3.52}$$

CHAPTER IV.

COMPUTATION OF SOLUTIONS OF ORDINARY DIFFERENTIAL EQUATIONS

In this chapter, estimates for bounds on computed solutions of O. D. E. with infinite precision arithmetic and on accumulated truncation errors are given using the measure $\mu(\cdot)$. Also, we extend and relate the earlier results on D. C. equation (Ch. II) to the implicit equation required by the backward Euler method.

Section 1 gives estimates for bounds on computed solutions and on errors, obtained from several computational schemes. Theorem 4-1 and Corollary 4-2 give estimates for the bound on the computed sequence by the backward Euler method. The estimates consist of two terms: the first term shows that the effect of the initial value decays exponentially and the second is bounded if the input $u(\cdot)$ is bounded. Since the backward Euler method is implicit, it requires in principle an infinite number of arithmetical operations and function evaluations at each time step. In implementing the backward Euler method at each time step, we modify it by truncating the iteration when the computed value is within some ε of the exact value. Theorem 4-3 gives an estimate for the bound on the error between the computed sequence by the backward Euler method and the computed sequence by the modified implementable method. The estimate is the sum of two terms: the first term shows that the effect of the initial error decays exponentially, and the second is proportional to the chosen δ incurred by truncating the iterative method at each time step. We consider next the algorithm where at each time step of the backward Euler method we use only one step of the Newton-Raphson method. Theorem 4-4 gives an estimate for the bound on the computed sequence thus obtained. Theorem 4-5 gives an estimate for the bound on the error sequence between the computed sequence by the backward Euler method and the one thus obtained. These estimates obtained are similar to those obtained in the previous theorems of this chapter.

In Section 2, the estimate for the bound on the so-called accumulated truncation error incurred by the backward Euler method is given by Theorem 4-6. Again the estimate is of a similar form, consisting of two terms: the first term shows that the effect of initial errors decays exponentially and the second is proportional to the step size.

In Section 3, we extend and relate the results of Chapter II to the implicit equation obtained by the backward Euler method. The effect of the step size on the existence and uniqueness of the D. C. solution as well as on the region of convergence for Newton-Raphson method with infinite and finite precision arithmetic is stressed.

1. Properties of The Computed Solution of O. D. E. (1.1) When It Is Computed by The Backward Euler Method (1.3) And Some of Its Simplified Versions

Throughout we assume an infinite precision arithmetic for all computations. Consider the O. D. E. (1.1):

$$\begin{cases} \dot{x} = f(x,t) + u(t) \\ x(0) = x_0 \end{cases} \qquad (1.1)$$

where $x(t), u(t) \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$ and $f: \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}^d$. We assume A1: $f(\theta_d, t) = \theta_d$, for all $t \in \mathbb{R}_+$; A2: $x \mapsto f(x,t)$ is in C^1 for all $t \in \mathbb{R}_+$; and A3: the input $u(\cdot)$ and for each fixed $x \in \mathbb{R}^d$, $t \mapsto f(x,t)$ are piecewise continuous on $\mathbb{R}_+$. Recall the backward Euler formula (1.3) and let $\{y_n\}_0^\infty$ denote the computed solution of eq.(1.1) by the backward Euler formula (1.3).

<u>Theorem 4-1.</u> If there exists an $m(\cdot) \in \mathcal{M}(\cdot)$ such that

$$-\mu[D_1 f(x,t)] \geq m(|x|) > 0 \quad \text{for all } x \in \mathbb{R}^d, \text{ for all}$$

$t \in \mathbb{R}_+$, then the computed solution $\{y_n\}_0^\infty$ of eq.(1.1) by the formula (1.3) satisfies:

$$|y_n| \leq (1+\varepsilon h)^{-n} |y_0| + \sum_{k=0}^{n-1} (1+\varepsilon h)^{-(k+1)} \cdot h \cdot |u_{n-k}| \quad \text{for all}$$

$$n \geq 1. \qquad (4.1)$$

<u>Proof.</u> From (1.3), we obtain:

$$y_{n+1} - hf(y_{n+1}, n+1) = y_n + hu_{n+1}. \qquad (4.2)$$

By Taylor's formula, we have:

$$\text{LHS of (4.2)} = y_{n+1} - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau \cdot y_{n+1}$$

$$= \left[I_{dxd} - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau \right] \cdot y_{n+1}. \qquad (4.3)$$

As in (3.33), observe that from the assumption, we obtain:

$$\mu\left[\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \leq -\varepsilon < 0. \qquad (4.4)$$

$$-\mu\left[-I_{dxd} + h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] = -\left\{-1 + \mu\left[h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right]\right\}, \quad \text{by Lemma 1-2, (e)}$$

$$= 1 - h\,\mu\left[\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right], \quad \text{by Lemma 1-2, (d)}$$

$$\geq 1 + h\varepsilon > 1. \qquad (4.5)$$

Using Lemma 1-2, (j), (4.2), (4.3) and (4.5), we obtain:

$$|y_n| + h|u_{n+1}| \geq |y_n + hu_{n+1}|$$

$$= \left|\left[I_{dxd} - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \cdot y_{n+1}\right|$$

$$\geq (1+h\varepsilon)|y_{n+1}|. \qquad (4.6)$$

$$|y_{n+1}| \leq (1+\varepsilon h)^{-1}|y_n| + (1+\varepsilon h)^{-1} \cdot h \cdot |u_{n+1}|. \qquad (4.7)$$

Hence, we obtain:

$$|y_n| \leq (1+\varepsilon h)^{-n}|y_0| + \sum_{k=0}^{n-1}(1+\varepsilon h)^{-(k+1)} \cdot h \cdot |u_{n-k}| . \quad \diamond \tag{4.8}$$

<u>Remark.</u> From Theorem 3-2, under the assumptions of the above theorem, the exact solution $x(\cdot;x_0)$ of (1.1) is also bounded-input bounded-output(B. I. B. O.) stable.

<u>Corollary 4-2.</u> Suppose $f(\cdot,\cdot)$ satisfies conditions A1, A2 and A3. If there exists a positive constant $m > 0$ such that

$$-\mu[D_1 f(x,t)] \geq m > 0 \quad \text{for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+,$$

then the computed solution $\{y_n\}_0^\infty$ of eq.(1.1) by eq.(1.3) satisfies:

$$|y_n| \leq (1+mh)^{-n}|y_0| + \sum_{k=0}^{n-1}(1+mh)^{-(k+1)} \cdot h \cdot |u_{n-k}| \quad \text{for all}$$

$$n \geq 1. \quad \diamond \tag{4.9}$$

<u>Relation to previous work.</u> Special cases of Corollary 4-2 were proved by Sandberg & Shichman under ℓ^2 norms, [17] ; and by Sandberg under weighted ℓ^1 norms, [3].

In order to solve the implicit equation (4.2) we use an iterative method, say the Newton-Raphson method. In practice we have to truncate the iterative method at each step of the backward Euler method. For example, at each step of the backward Euler method, instead of solving eq.(4.2) exactly for

y_{n+1} and thus obtaining the sequence $\{y_n\}_0^\infty$, we truncate the procedure; this will give us a sequence $\{\tilde{y}_n\}_0^\infty$. More precisely, at the (n+1)-th step we should solve (see eq.(4.2)) the equation:

$$y^*_{n+1} - hf(y^*_{n+1}, n+1) = \tilde{y}_n + hu_{n+1}, \text{ for all } n \geq 0 \tag{4.10}$$

for y^*_{n+1}. Note that y^*_{n+1} is the (exact) solution of (4.10). We solve (4.10) by iteration and we stop the iteration when we obtain an iterate, say $\tilde{y}_{n+1}$, such that for some $\varepsilon > 0$

$$|\tilde{y}_{n+1} - y^*_{n+1}| \leq \varepsilon, \text{ for all } n \geq 0. \tag{4.11}$$

Note that we have three sequences in mind: $\{y_n\}_0^\infty$, $\{\tilde{y}_n\}_0^\infty$, and $\{y^*_n\}_0^\infty$, where $y^*_0 \triangleq \tilde{y}_0$ and $\tilde{y}_0$ is the initial condition for our simplified calculation. The next theorem gives an estimate for a bound on $\tilde{y}_n - y_n$.

<u>Theorem 4-3</u>. Assume that all the conditions of Corollary 4-2 are satisfied. Let $\{y_n\}_0^\infty$ and $\{\tilde{y}_n\}_0^\infty$ be defined by eq.(4.2) and eq.(4.10) & (4.11). Then the difference between $\tilde{y}_n$ and y_n satisfies:

$$|\tilde{y}_n - y_n| \leq (1+mh)^{-n}|\tilde{y}_0 - y_0| + \varepsilon \sum_{k=0}^{n-1}(1+mh)^{-k}, \text{ for all}$$

$$n \geq 1. \tag{4.12}$$

<u>Proof</u>. From eq.(4.2) and (4.10), we obtain:

$$y_{n+1} - y^*_{n+1} - h\left[f(y_{n+1}, n+1) - f(y^*_{n+1}, n+1)\right] = y_n - \tilde{y}_n. \tag{4.13}$$

By Taylor's formula applied to $f(y_{n+1}, n+1) - f(y^*_{n+1}, n+1)$, eq. (4.13) becomes:

$$\left\{ I - h \int_0^1 D_1 f\left[(1-\tau)y^*_{n+1} + \tau y_{n+1}, n+1\right] d\tau \right\} \cdot (y_{n+1} - y^*_{n+1})$$

$$= y_n - \tilde{y}_n. \tag{4.14}$$

Similar to the proof of Theorem 4-1, we have, as in (4.5),

$$-\mu\left[-I + h \int_0^1 D_1 f\left[(1-\tau)y^*_{n+1} + \tau y_{n+1}, n+1\right] d\tau\right]$$

$$\geq 1 + hm > 1. \tag{4.15}$$

Then, by using Lemma 1-2, (j), eq. (4.14) and eq. (4.15) we get:

$$(1+hm)\left|y_{n+1} - y^*_{n+1}\right| \leq \left|y_n - \tilde{y}_n\right|. \tag{4.16}$$

From eq. (4.11) and eq. (4.16), we conclude that:

$$\begin{aligned} \left|\tilde{y}_{n+1} - y_{n+1}\right| &\leq \left|\tilde{y}_{n+1} - y^*_{n+1}\right| + \left|y^*_{n+1} - y_{n+1}\right| \\ &\leq \varepsilon + (1+mh)^{-1}\left|\tilde{y}_n - y_n\right|. \end{aligned} \tag{4.17}$$

Hence, we obtain:

$$\left|\tilde{y}_n - y_n\right| \leq (1+mh)^{-n}\left|\tilde{y}_0 - y_0\right| + \varepsilon \sum_{k=0}^{n-1} (1+mh)^{-k}, \text{ for all } n \geq 1.$$

$\diamond$

<u>Relation to previous work.</u> Theorem 4-3 is a generalization of earlier results: Sandberg in [3] proved the same result under

weighted ℓ^1 norms and Sandberg & Shichman in [17] proved the similar result under ℓ^2 norms. In the literature [17], the estimate of bounds includes a Lipschitz constant, but in Theorem 4-3 this constant is eliminated. In fact, it can be verified that Theorem 4-3 gives a tighter estimate in view of the fact $-\mu(-A) \leq \mu(A) \leq \|A\|$.

Next, we consider a simplified computational algorithm where, at each step of the backward Euler method, we use only one step of the Newton-Raphson method. The iteration is then given by:

$$\begin{cases} \bar{y}_{n+1} = \bar{y}_n - [I - hD_1 f(\bar{y}_n, n+1)]^{-1} [-hf(\bar{y}_n, n+1) - hu_{n+1}] \\ \bar{y}_0 \text{: given,} \end{cases}$$

for all $n \geq 0$. (4.18)

The next theorem shows that under natural assumptions the sequence $\{\bar{y}_n\}_0^\infty$ computed by the formula (4.18) has an estimate consisting of two terms as in (4.20) below: the first term shows that the effect of the initial condition is constant or decays exponentially as $n \to \infty$ and the second shows that it is bounded if the series $\sum_{k=0}^{\infty} |u_k|$ is convergent.

Theorem 4-4. Assume that all conditions of Corollary 4-2 are satisfied. Assume further that there exists a constant $\delta \in [0,m]$ such that

$$\|D_1 f(x,t) - D_1 f(\alpha x,t)\| \leq \delta \leq m, \text{ for all } x \in \mathbb{R}^d,$$

for all $t \in \mathbb{R}_+$, for all $\alpha \in [0,1]$. (4.19)

Under these conditions, the computed solution $\{\bar{y}_n\}_0^\infty$ by formula (4.18) satisfies:

$$|\bar{y}_n| \leq \left[\frac{1+\varepsilon h}{1+mh}\right]^n |\bar{y}_0| + \frac{h}{1+mh} \sum_{k=0}^{n-1} \left[\frac{1+\varepsilon h}{1+mh}\right]^k |u_{n-k}|, \; n \geq 1. \quad (4.20)$$

Proof. By applying Taylor's formula to $f(\bar{y}_n, n+1)$, from eq. (4.18) we get:

$$\bar{y}_{n+1} = \bar{y}_n - \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \left[-h\int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau \cdot \bar{y}_n\right.$$

$$\left.-hu_{n+1}\right], \text{ or} \quad (4.21)$$

$$y_{n+1} = \left\{I + h\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\right\} \cdot y_n$$

$$+ \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot hu_{n+1}. \quad (4.22)$$

Thus,

$$|\bar{y}_{n+1}| \leq \left\|I + h\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\right\| \cdot$$

$$|\bar{y}_n| + \left\|\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1}\right\| \cdot h\, |u_{n+1}|. \quad (4.23)$$

We have:

$$\left\|\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1}\right\| \leq 1/(1+mh), \text{ for all } y_n \in \mathbb{R}^d, \text{ for}$$

all $n \geq 0$, because (4.24)

$$\frac{1}{\left\| \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} \right\|} \geq 1 + mh > 1,$$ by the assumption

and Lemma 1-2, (ℓ), (e) & (d).

We now claim that

$$\left\| I + h\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right\|$$

$$\leq \frac{1+\varepsilon h}{1+mh} \leq 1. \tag{4.25}$$

$$\left\| I + h\left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right\|$$

$$= \left\| \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \left[I - hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right] \right\|$$

$$\leq \left\| \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \right\| \cdot \left\{ 1 + h \int_0^1 \left\| D_1 f(\tau \bar{y}_n, n+1) - D_1 f(\bar{y}_n, n+1) \right\| d\tau \right\}$$

$$\leq 1/(1+mh) \cdot \left\{ 1 + h \int_0^1 \varepsilon \, d\tau \right\}$$

$= (1+\varepsilon h)/(1+mh)$, by (4.24) and the assumption (4.19).

Thus, from (4.23), (4.24) and (4.25), we have:

$$|y_{n+1}| \leq \frac{1+\varepsilon h}{1+mh} |y_n| + \frac{h}{1+mh} |u_{n+1}| . \tag{4.26}$$

Hence, the result (4.20) follows. $\diamond$

Remark. Roughly speaking, the assumption (4.19) requires that for each fixed t the function $f(\cdot,t)$ is not too nonlinear. The above theorem shows that if the function $f(\cdot,t)$ is not too nonlinear and if there exists a constant $m > 0$ such that

$$-\mu[D_1 f(x,t)] \geq m > 0 \quad \text{for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+,$$

then the above seemingly crude algorithm still gives a computed sequence which is bounded by two terms as in (4.20).

Relation to previous work. Sandberg & Shichman in [17] proposed the above algorithm and proved the similar results under ℓ^2 norms. In the above theorem the flexibility of the measure $\mu(\cdot)$ led us to more general and explicit estimate on bounds.

Using the same technique we are going to obtain an estimate on the bound of $\bar{y}_n - y_n$ where $\{\bar{y}_n\}_0^\infty$ and $\{y_n\}_0^\infty$ are computed sequences by the above algorithm and the exact backward Euler method, respectively.

Theorem 4-5. Let $\{\bar{y}_n\}_0^\infty$ and $\{y_n\}_0^\infty$ satisfy respectively the simplified algorithm (4.18) and the backward Euler method (4.2). Assume that all the conditions of Corollary 4-2 are satisfied. Assume further that there exists a constant $\varepsilon \in [0,m)$ such that

$$\|D_1 f(x,t) - D_1 f(\alpha x,t)\| \leq \varepsilon < m \quad \text{for all } x \in \mathbb{R}^d,$$

$$\text{for all } t \in \mathbb{R}_+, \text{ for all } \alpha \in [0,1] \; ; \qquad (4.21)$$

that there exists a constant $\ell > 0$ such that

$$\|D_1 f(x,t)\| \leq \ell, \text{ for all } x \in \mathcal{R}^d, \text{ for all } t \in \mathcal{R}_+ \tag{4.28}$$

and that $u(\cdot)$ is bounded on $[0,\infty)$.

Under these conditions, the difference between $\bar{y}_n$ and y_n satisfies:

$$\left|\bar{y}_n - y_n\right| \leq \left[\frac{1+\varepsilon h}{1+mh}\right]^n \left|\bar{y}_0 - y_0\right| + \frac{2\ell + \varepsilon(1+mh)}{\varepsilon} \frac{(1+\varepsilon h)^n - 1}{(1+mh)^{n+1}} \left|y_0\right|$$

$$+ \frac{2\ell+\varepsilon}{m(m-\varepsilon)} \|u(\cdot)\|_\infty , \tag{4.29}$$

where $\|u(\cdot)\|_\infty \triangleq \sup_{t \in [0,\infty)} |u(t)|$.

Proof. From (4.2) and (4.3), we have:

$$\left[I - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \cdot y_{n+1} = y_n + hu_{n+1}. \tag{4.30}$$

Equation (4.21) is rewritten as:

$$\left[I - hD_1 f(\bar{y}_n, n+1)\right] \cdot \bar{y}_{n+1} = \left[I - hD_1 f(\bar{y}_n, n+1)\right] \cdot \bar{y}_n$$

$$+ h\int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \cdot \bar{y}_n + hu_{n+1}. \tag{4.31}$$

Subtracting (4.30) from (4.31), we obtain:

$$\left[I - hD_1 f(\bar{y}_n, n+1)\right] \bar{y}_{n+1} - \left[I - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \cdot y_{n+1}$$

$$= \left[I - hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right] \cdot \bar{y}_n - y_n. \quad (4.32)$$

Note that $\left[I - hD_1 f(\bar{y}_n, n+1)\right]$ is nonsingular, since

$$\left| [I - hD_1 f(\bar{y}_n, n+1)]\, z \right| \geq (1+mh)|z| > 0, \text{ for all } z \neq \theta_d.$$

Equation (4.32) becomes:

$$\bar{y}_{n+1} - \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \left[I - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \cdot$$

$$y_{n+1}$$

$$= \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \left[I - hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau \bar{y}_n,\right.$$

$$\left. n+1)\, d\tau \right] \cdot \bar{y}_n - \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot y_n. \quad (4.33)$$

$$\text{LHS of } (4.33) = (\bar{y}_{n+1} - y_{n+1}) + \left\{ I - \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1}\right.$$

$$\left. \left[I - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\right] \right\} \cdot y_{n+1}. \quad (4.34)$$

$$\text{RHS of } (4.33) = \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \left[I - hD_1 f(\bar{y}_n, n+1)\right.$$

$$\left. + h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right] \cdot (\bar{y}_n - y_n)$$

$$+ \left[I - hD_1 f(\bar{y}_n, n+1)\right]^{-1} \cdot \left[I - hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau y_n, n+1)\right.$$

$$d\tau\Big]\cdot y_n - \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\cdot y_n. \tag{4.35}$$

Using (4.34) and (4.35), eq.(4.33) becomes:

$$\bar{y}_{n+1} - y_{n+1} = -\Big\{I - \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\Big[I - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\Big]\Big\}\cdot y_{n+1} + \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\Big[I - hD_1 f(\bar{y}_n, n+1) + h\int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\Big]\cdot(\bar{y}_n - y_n) + \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\cdot \Big[-hD_1 f(\bar{y}_n, n+1) + h\int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\Big]\cdot y_n. \tag{4.36}$$

Hence we have:

$$\left|\bar{y}_{n+1} - y_{n+1}\right| \leq \left\|I - \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\Big[I - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau\Big]\right\|\cdot\left|y_{n+1}\right| + \left\|I + \Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1} h\int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\right\|\cdot\left|\bar{y}_n - y_n\right| + \left\|\Big[I - hD_1 f(\bar{y}_n, n+1)\Big]^{-1}\right\|\cdot\left\|-hD_1 f(\bar{y}_n, n+1) + h\int_0^1 D_1 f(\tau\bar{y}_n, n+1)\, d\tau\right\|\cdot\left|y_n\right|. \tag{4.37}$$

In the proof of Theorem 4-4, we proved that

$$\left\|\Big[I - hD_1 f(y_n, n+1)\Big]^{-1}\right\| \leq 1/(1+mh), \tag{4.24}$$

and that

$$\left\| I + \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right\| \leq$$

$$(1+\varepsilon h)/(1+mh) < 1. \tag{4.25}$$

Claim: $\left\| I - \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} \left[I - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\right.\right.$

$$\left.\left. d\tau \right] \right\| \leq 2\ell h/(1+mh). \tag{4.38}$$

$$\left\| I - \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} \left[I - h \int_0^1 D_1 f(\tau y_{n+1}, n+1)\, d\tau \right] \right\|$$

$$= \left\| \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} \left[I - hD_1 f(\bar{y}_n, n+1) - I + h \int_0^1 \right.\right.$$

$$\left.\left. D_1 f(\tau y_{n+1}, n+1)\, d\tau \right] \right\|$$

$$\leq \left\| \left[I - hD_1 f(\bar{y}_n, n+1) \right]^{-1} \right\| h \left\{ \left\| D_1 f(\bar{y}_n, n+1) \right\| + \int_0^1 \left\| D_1 f(\right.\right.$$

$\left.\left. \tau y_{n+1}, n+1) \right\| d\tau \right\} \leq h \cdot 2\ell/(1+mh)$, by (4.24) and the assumption.

Claim: $\left\| -hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right\| \leq \varepsilon h.$ (4.39)

$$\left\| -hD_1 f(\bar{y}_n, n+1) + h \int_0^1 D_1 f(\tau \bar{y}_n, n+1)\, d\tau \right\|$$

$$= h \int_0^1 \left\| -D_1 f(\bar{y}_n, n+1) + D_1 f(\tau \bar{y}_n, n+1) \right\| d\tau$$

$\leq h\varepsilon$, by the assumption (4.19).

From the part of the proof of Corollary 4-2, we can obtain:

$$|y_{n+1}| \leq \frac{1}{1+mh}|y_n| + \frac{h}{1+mh}|u_{n+1}|, \tag{4.40}$$

and, as a result, we get:

$$|y_n| \leq (1+mh)^{-n}|y_0| + \sum_{k=0}^{n-1}(1+mh)^{-(k+1)}h|u_{n-k}|. \tag{4.41}$$

Thus, using (4.24), (4.25), (4.38), (4.39) and (4.40), eq.(4.37) becomes:

$$\begin{aligned}|\bar{y}_{n+1} - y_{n+1}| &\leq 2\ell h/(1+mh)\cdot\left[(1/1+mh)|y_n| + (h/1+mh)|u_{n+1}|\right]\\ &+ \left[(1+\varepsilon h)/(1+mh)\right]|\bar{y}_n - y_n| + \varepsilon h/1+mh\cdot|y_n|\\ &= \left[(1+\varepsilon h)/(1+mh)\right]|\bar{y}_n - y_n| + \left[2\ell h/(1+mh)^2 + \varepsilon h/(1+mh)\right]\cdot|y_n|\\ &+ 2\ell h^2/(1+mh)^2\cdot|u_{n+1}|.\end{aligned} \tag{4.42}$$

Let $\rho_1 \triangleq (1+\varepsilon h)(1+mh)^{-1} < 1$ and $\rho_2 \triangleq 2\ell h(1+mh)^{-2} + \varepsilon h(1+mh)^{-1}$.

Then, from (4.42), we obtain:

$$|\bar{y}_n - y_n| \leq \rho_1^n|\bar{y}_0 - y_0| + \rho_2\sum_{k=0}^{n-1}\rho_1^k|y_{n-k-1}| + 2\ell h^2(1+mh)^{-2}\cdot$$

$$\sum_{k=0}^{n-1}\rho_1^k|u_{n-k}|. \tag{4.43}$$

Claim: $\sum_{k=0}^{n-1} \rho_1^k |y_{n-k-1}| \leq \left[(1+\varepsilon h)^n - 1\right]\left[(1+mh)^{n-1}\varepsilon h\right]^{-1} |y_0|$

$$+(1+mh)\left[m(m-\varepsilon)h\right]^{-1} \|u(\cdot)\|_{\infty} . \qquad (4.44)$$

Using (4.41), we have:

$$|y_{n-k-1}| \leq (1+mh)^{-(n-k-1)} |y_0| + \sum_{j=0}^{n-k-2} (1+mh)^{-(j+1)} h |u_{n-k-1-j}|$$

$$\leq (1+mh)^{-n+k+1} |y_0| + h \|u(\cdot)\|_{\infty} \sum_{j=0}^{n-k-2} (1+mh)^{-(j+1)}$$

$$\leq (1+mh)^{-n+k+1} |y_0| + h \|u(\cdot)\|_{\infty} \sum_{j=0}^{\infty} (1+mh)^{-(j+1)}$$

$$= (1+mh)^{-n+k+1} |y_0| + h \|u(\cdot)\|_{\infty} /mh. \qquad (4.45)$$

Thus,

$$\sum_{k=0}^{n-1} \rho_1^k |y_{n-k-1}| \leq \sum_{k=0}^{n-1} (1+\varepsilon h)^k (1+mh)^{-n+1} |y_0| + m^{-1} \|u(\cdot)\|_{\infty} \cdot$$

$$\sum_{k=0}^{n-1} \left[(1+\varepsilon h)(1+mh)^{-1}\right]^k$$

$$\leq (1+mh)^{-n+1} |y_0| \sum_{k=0}^{n-1} (1+\varepsilon h)^k + m^{-1} \|u(\cdot)\|_{\infty} \sum_{k=0}^{\infty} \left[(1+\varepsilon h)(1+mh)^{-1}\right]^k$$

$$= (1+mh)^{-n+1} |y_0| \left[(1+\varepsilon h)^n - 1\right]\left[\varepsilon h\right]^{-1} + m^{-1} \|u(\cdot)\|_{\infty} (1+mh)(m-\varepsilon)^{-1} \cdot$$

$$h^{-1}. \qquad (4.46)$$

Note that $\sum_{k=0}^{n-1} \rho_1^k |u_{n-k}| \leq (1+mh)(m-\varepsilon)^{-1} h^{-1} \|u(\cdot)\|_{\infty}$. (4.47)

Using (4.44) and (4.47), we obtain from (4.43):

$$|\bar{y}_n - y_n| \leq \rho_1^n |\bar{y}_0 - y_0| + \rho_2 \Big\{ \big[(1+\varepsilon h)^n - 1\big] (1+mh)^{-n+1} \varepsilon^{-1} h^{-1} \cdot |y_0| + (1+mh)m^{-1}(m-\varepsilon)^{-1}h^{-1} \|u(\cdot)\|_\infty \Big\} + 2\ell h^2(1+mh)^{-2} \cdot (1+mh)(m-\varepsilon)^{-1}h^{-1} \|u(\cdot)\|_\infty$$

$$= \big[(1+\varepsilon h)(1+mh)^{-1}\big]^n |\bar{y}_0 - y_0| + \big[2\ell h + (1+mh)\varepsilon h\big] (1+mh)^{-2} \cdot \big[(1+\varepsilon h)^n - 1\big] (1+mh)^{-n+1} \varepsilon^{-1} h^{-1} |y_0| + \Big\{ \big[2\ell h + (1+mh)\varepsilon h\big](1+mh)^{-2} \cdot (1+mh)m^{-1}(m-\varepsilon)^{-1}h^{-1} + 2\ell h^2(1+mh)^{-2} \cdot (1+mh)(m-\varepsilon)^{-1}h^{-1} \Big\} \|u(\cdot)\|_\infty$$

$$= \big[(1+\varepsilon h)(1+mh)^{-1}\big]^n |\bar{y}_0 - y_0| + \big[(1+\varepsilon h)^n - 1\big] (1+mh)^{-n-1} \cdot \big[2\ell + \varepsilon(1+mh)\big] \varepsilon^{-1} |y_0| + \big[2\ell + (1+mh)\varepsilon + 2\ell hm\big] (1+mh)^{-1} \cdot m^{-1}(m-\varepsilon)^{-1} \|u(\cdot)\|_\infty$$

$$= \big[(1+\varepsilon h)(1+mh)^{-1}\big]^n |\bar{y}_0 - y_0| + \big[2\ell + \varepsilon(1+mh)\big] \varepsilon^{-1} \cdot \big[(1+\varepsilon h)^n - 1\big] (1+mh)^{-n-1} |y_0| + (2\ell + \varepsilon)m^{-1}(m-\varepsilon)^{-1} \|u(\cdot)\|_\infty . \quad \diamond$$

2. Comparison of The Exact Solution of O. D. E. (1.1) with The Computed Solution of The Backward Euler Method (1.3)

Throughout this section we assume infinite precision arithmetic for all computaions.

Consider the solution $x(\cdot; x_0)$ of O. D. E. (1.1). Let $\{y_n\}_0^\infty$ be the computed solution of (1.1) by the backward Euler formula (1.3). The error vector $x_n - y_n$ is said to be the accumulated truncation error. In this section, under reasonable assumptions, we give an estimate of the accumulated truncation error. We show that the error does not build up indefinitely, and that the effect of an initial error decays exponentially.

<u>Theorem 4-6</u>. Assume that all the conditions of Corollary 4-2 are satisfied. If, in addition, for any fixed $x \in \mathbb{R}^d$, $D_2 f(x,\cdot)$ is piecewise continuous and $\dot{u}(\cdot)$ is piecewise continuous, if both $u(\cdot)$ and $\dot{u}(\cdot)$ are bounded on $\mathbb{R}_+$, and if there exist positive constant α and β such that

$\|D_1 f(x,t)\| \leq \alpha$ and $|D_2 f(x,t)| \leq \beta$, for all $x \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$, then there exists a $\rho > 0$ independent of h such that

$$|x_n - y_n| \leq (1+mh)^{-n}|x_0 - y_0| + \rho h, \text{ for all } n \geq 0. \qquad (4.48)$$

<u>Proof</u>. From Corollary 3-3, the solution $x(\cdot;x_0)$ of (1.1) satisfies the inequality (3.36):

$$|x(t)| \leq \exp(-mt)|x_0| + \int_0^t \exp[-m(t-\tau)] \cdot |u(\tau)| d\tau. \qquad (3.36)$$

Since $u(\cdot)$ is bounded on $\mathbb{R}_+$, $x(\cdot)$ is also bounded on $\mathbb{R}_+$, i.e.,

$$\|x(\cdot)\|_\infty \triangleq \sup_{t \in \mathbb{R}_+} |x(t)| < \infty.$$

Claim: $\ddot{x}(\cdot)$ is bounded on $\mathbb{R}_+$, i.e., $\|\ddot{x}(\cdot)\|_\infty < \infty$.

$$\dot{x}(t) = f(x,t) + u(t). \qquad (1.1)$$

Differentiate both sides of (1.1) with respect to t:

$$\ddot{x}(t) = D_1 f(x,t)\cdot \dot{x}(t) + D_2 f(x,t) + \dot{u}(t)$$

$$= D_1 f(x,t)\cdot \left[f(x,t) + u(t)\right] + D_2 f(x,t) + \dot{u}(t). \qquad (4.49)$$

So,

$$|\ddot{x}(t)| \leq \|D_1 f(x,t)\| \cdot \left\{ \int_0^1 \|D_1 f(\tau x,t)\| d\tau \cdot |x(t)| + |u(t)| \right\}$$

$$+ |D_2 f(x,t)| + |\dot{u}(t)|$$

$$\leq \alpha\left\{\alpha\|x(\cdot)\|_\infty + \|u(\cdot)\|_\infty\right\} + \beta + \|\dot{u}(\cdot)\| < \infty,$$

for all $t \in \mathbb{R}_+$. (4.50)

Thus, $\ddot{x}(\cdot)$ is bounded on $\mathbb{R}_+$. Define the <u>local truncation error</u> $\{\xi_n\}_0^\infty$ by:

$$\xi_n \triangleq x_{n+1} - x_n - h\dot{x}_{n+1}, \; n \geq 0. \qquad (4.51)$$

Claim the local truncation error ξ_n has an upper bound, more precisely, there exists a positive constant independent of h such that

$$|\xi_n| \leq \tfrac{1}{2}h^2 \rho_1, \text{ for all } n \geq 0. \qquad (4.52)$$

By applying Taylor's formula to each component of x_n, we obtain:

$$x_n = x_{n+1} - h\dot{x}_{n+1} + \tfrac{1}{2}h^2 U_n, \qquad (4.53)$$

where j-th component $[U_n]_j$ of U_n is equal to the j-th component $\ddot{x}_j$ of $\ddot{x}$ evaluated at some point of $[nh,(n+1)h]$. By the definition of ξ_n, (4.52) and (4.53), we have:

$$\xi_n = -\tfrac{1}{2}h^2 U_n, \text{ for all } n \geq 0. \tag{4.54}$$

Since by (3.36) $\ddot{x}$ is bounded on $\mathbb{R}_+$, U_n is also bounded. Thus,

$|U_n| \leq \rho_1$, for some $\rho_1 > 0$, for all $n \geq 0$.

Hence, by (4.54),

$$|\xi_n| \leq \tfrac{1}{2}h^2 \rho_1, \text{ for all } n \geq 0. \tag{4.52}$$

Next, we derive a difference inequality with respect to $|y_n - x_n|$.

$$y_{n+1} - hf(y_{n+1}, n+1) = y_n + hu_{n+1}. \tag{4.2}$$

From (4.51),

$$x_{n+1} - hf(x_{n+1}, n+1) = x_n + hu_{n+1} + \xi_n. \tag{4.55}$$

Subtracting (4.55) from (4.2), we get:

$$y_{n+1} - x_{n+1} - h\int_0^1 D_1 f\,[(1-\tau)x_{n+1} + \tau y_{n+1}, n+1]\,d\tau \cdot (y_{n+1} - x_{n+1})$$

$$= y_n - x_n - \xi_n, \tag{4.56}$$

or

$$\left[I - h\int_0^1 D_1 f\,[(1-\tau)x_{n+1} + \tau y_{n+1}, n+1]\,d\tau\right] \cdot (y_{n+1} - x_{n+1})$$

$$= (y_n - x_n) - \xi_n. \tag{4.57}$$

Analogous to the part of the proof in Theorem 4-1, as in (4.6), we get:

$$|y_n - x_n| + |\xi_n| \geq |(y_n - x_n) - \xi_n| \geq (1+mh)|y_{n+1} - x_{n+1}| \tag{4.58}$$

or

$$\begin{aligned} |y_{n+1} - x_{n+1}| &\leq (1+mh)^{-1}|y_n - x_n| + (1+mh)^{-1}|\xi_n| \\ &\leq (1+mh)^{-1}|y_n - x_n| + \tfrac{1}{2}h^2(1+mh)^{-1}\rho_1. \end{aligned} \tag{4.59}$$

Solving the recursive inequality (4.59),

$$\begin{aligned} |y_n - x_n| &\leq (1+mh)^{-n}|y_0 - x_0| + \tfrac{1}{2}h^2\rho_1 \sum_{k=1}^{n} (1+mh)^{-k} \\ &\leq (1+mh)^{-n}|y_0 - x_0| + \tfrac{1}{2}h^2\rho_1 \sum_{k=1}^{\infty} (1+mh)^{-k} \\ &= (1+mh)^{-n}|y_0 - x_0| + \tfrac{1}{2}h^2\rho_1/mh. \end{aligned} \tag{4.60}$$

By letting $\rho = \rho_1/2m$, we obtain the result. ◇

Remark. As in (4.48) the estimate of the accumulated truncation error shows that the effect of the initial error decays exponentially as $(1+mh)^{-n}$ and that the effect of the local truncation error does not build up indefinitely; in fact proportional to h.

Relation to previous work. The special case of Theorem 4-6 under weighted ℓ^1 norms was proved by Sandberg [3].

3. Extensions and Relation to Results from Earlier Chapters

In Chapter II, we discussed properties of D. C.

equations. In this section we extend and relate those results to the implicit equations obtained by the backward Euler method:

$$y_{n+1} - hf(y_{n+1}, n+1) = y_n + hu_{n+1}. \qquad (4.2)$$

We assume that all conditions of Corollary 4-2 are satisfied, i.e.,

$f(\theta_d,t) = \theta_d$ for all $t \in \mathbb{R}_+$; $x \mapsto f(x,t)$ is in C^1 for all $t \in \mathbb{R}_+$; and there exists a positive constant $m > 0$ such that $-\mu[D_1 f(x,t)] \geq m > 0$ for all $x \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$.

Observe that Lemma 1-2, (e), (d) and the above assumption imply

$$-\mu\left[-(I - hD_1 f(y_{n+1}, n+1))\right] \geq 1 + mh > 1 \quad \text{for all}$$

$$n \geq 0. \qquad (4.61)$$

3.1 Using Corollary 2-4, it follows that for each integer $n \geq 0$, for any fixed $h > 0$ and for any u_{n+1} & y_n, the solution y^*_{n+1} of (4.2) exists and is unique. Furthermore, y_{n+1} is a continuously differentiable function of the previous value y_n, the step size h and the input value u_{n+1}.

3.2 Let $\{y^i_{n+1}\}_{i=0}^{\infty}$ be a computed sequence of (4.2) by the Newton-Raphson method with infinite-precision arithmetic. From Theorem 2-6, we conclude that if the mapping $f(\cdot,n+1): \mathbb{R}^d \to \mathbb{R}^d$ satisfies the condition(2.13), then by defining r^*_h to be the unique solution of

$$r = 2(1+mh)/hk^*(r), \ r > 0, \tag{4.62}$$

the computed solution $\left\{ y^i_{n+1} \right\}_{i=0}^{\infty}$ starting from inside the ball $B(y^*_{n+1}; r^*_h)$ remains in this ball and converges to the unique solution y^*_{n+1} at least quadratically. Since,

$$r^*_h = \max_{r>0} \min \left\{ r, \ (2m + 2/h)/k^*(r) \right\},$$

the convergence region is enlarged if either m becomes large, or if h becomes small, or if $f(\cdot)$ becomes less nonlinear, i.e., $k^*(r)$ is decreased for each fixed $r > 0$. For any fixed m and for any fixed $k^*(\cdot)$, $h \mapsto r^*_h$ is strictly decreasing; $r^*_h \downarrow r^*$ as $h \to +\infty$, where

$$r^* \triangleq \max_{r>0} \min \left\{ r, \ 2m/k^*(r) \right\} \text{ as in } (2.24).$$

Furthermore, $r^*_h \to \infty$ as $h \downarrow 0+$.

These conclusions can easily be made obvious by considering the original implicit equation (4.2). If h is sufficiently small, then (4.2) is close to a linear equation. If h is sufficiently large, then (4.2) is approximated by:

$$-f(y_{n+1}, n+1) = u_{n+1}. \tag{4.63}$$

Then, using Theorem 2-6 directly, the convergence region is $B(y^*_{n+1}; r^*)$, where r^* is defined as in (2.24).

3.3 Using Corollary 2-7, it follows that if the mapping $f(\cdot, n+1): \mathbb{R}^d \to \mathbb{R}^d$ satisfies the condition (2.25), then by defining r^*_h to be the unique solution of

$$r = \frac{2(1+mh)}{hk_0\left[r + \left|y_{n+1}^0 - hf(y_{n+1}^0, n+1) - y_n - hu_{n+1}\right|/(1+mh)\right]},$$

$$r > 0, \tag{4.64}$$

and assuming $\left|y_{n+1}^0 - hf(y_{n+1}^0, n+1) - y_n - hu_{n+1}\right| \leq (1+mh)r_h^*$, then the corresponding Newton-Raphson sequence $\left\{ y_{n+1}^i \right\}_{i=0}^{\infty}$ remains in $B(y_{n+1}^*; r_h^*)$ and converges to the unique solution y_{n+1}^* at least quadratically.

3.4 If we take into account the local round-off error on the Newton-Raphson method, then as in Theorem 2-9, for sufficiently small local round-off error, the radius of the convergence region is $2\mathcal{E}_\infty$ smaller than that of the infinite precision arithmetic case, and instead of quadratic convergence to the unique solution y_{n+1}^*, we obtain convergence to within a ball centered on y_{n+1}^* with a radius $3\mathcal{E}_\infty$ in a finite number of steps.

3.5 Let $\tilde{y}_{n+1} \in \mathbb{R}^d$ be an intermediate result in the course of solving (4.2) by any iterative algorithm. Let y_{n+1}^* be the exact solution. The error, namely $y_{n+1} - \tilde{y}_{n+1}$, is bounded by

$$\left|y_{n+1}^* - \tilde{y}_{n+1}\right| \leq \left|\tilde{y}_{n+1} - hf(\tilde{y}_{n+1}, n+1) - y_n - hu_{n+1}\right|/(1+mh),$$

$$\text{for all } n \geq 0. \tag{4.65}$$

3.6 In Section 1 and Section 2, we assumed that the infinite precision arithmetic for integrating the O. D. E. (1.1). Concerning local round-off errors note that the effect of local round-off errors is equivalent to some additional input. So, if the local round-off errors are bounded on $\mathbb{Z}_+$, then under conditions of Theorem 4-1, Corollary 4-2, Theorem 4-3 or Theorem 4-6, the <u>accumulated</u> round-off error is bounded on $\mathbb{Z}_+$.

APPENDIX

<u>Lemma A-1</u>. Let $A \in \mathbb{R}^{d\times d}$ and $B(x,t) \in \mathbb{R}^{d\times d}$ for all $x \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$. If A is symmetric positive definite and B(x,t) is uniformly positive definite in $\mathbb{R}^d \times \mathbb{R}_+$(not necessarily symmetric), more precisely there exists a positive constant $\varepsilon_B > 0$ such that

$$\langle y, B(x,t)y\rangle \geq \varepsilon_B |y|^2 \text{ for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+, \quad \text{(A-1)}$$

for all $y \in \mathbb{R}^d$, then there exists a nonsingular constant matrix P such that $PAB(x,t)P^{-1}$ is uniformly positive definite: there exists a positive constant $\varepsilon_{AB} > 0$ such that

$$\langle y, PAB(x,t)P^{-1}y\rangle \geq \varepsilon_{AB} |y|^2 \text{ for all } x \in \mathbb{R}^d, \text{ for all } t \in \mathbb{R}_+,$$

for all $y \in \mathbb{R}^d$. (A-2)

<u>Proof</u>. Since A is symmetric positive definite, $A^{\frac{1}{2}}$ is uniquely defined, real, symmetric and positive definite. Furthermore $A^{-\frac{1}{2}} \cdot A^{\frac{1}{2}} = I$. So, we pick $P = A^{-\frac{1}{2}}$. Then, we obtain:

$$\langle y, PAB(x,t)P^{-1}y\rangle = \langle y, A^{\frac{1}{2}}B(x,t)A^{\frac{1}{2}}y\rangle = \langle A^{\frac{1}{2}}y, B(x,t)A^{\frac{1}{2}}y\rangle$$

$$\geq \varepsilon_B |A^{\frac{1}{2}}y|^2. \quad \text{(A-3)}$$

Note that $|A^{\frac{1}{2}}y| \geq \left[\|A^{-\frac{1}{2}}\|\right]^{-1} \cdot |y|$, for all $y \in \mathbb{R}^d$, (A-4)

and that $\|A^{-\frac{1}{2}}\| > 0$. (A-5)

By letting $\varepsilon_{AB} \triangleq \varepsilon_B \cdot \left[\|A^{-\frac{1}{2}}\|\right]^{-2} > 0$, we obtain the (A-6) inequality (A-2). ◇

<u>Corollary A-2.</u> Assume that all the conditions of Lemma A-1 are satisfied. Then all the real parts of the eigenvalues of $AB(x,t)$ is greater than or equal to $\varepsilon_{AB} > 0$ for all $x \in \mathbb{R}^d$, for all $t \in \mathbb{R}_+$.

<u>Proof.</u> Observe that the eigenvalues of any matrix are invariant under similarity transformations. For any $i = 1,2,\cdots,d$,

$$\mathcal{R}e\,\lambda_i(AB(x,t)) = \mathcal{R}e\,\lambda_i(PAB(x,t)P^{-1})$$

$$= -\mu_2(-PAB(x,t)P^{-1}) \quad \text{by Lemma 1-2, (i)}$$

$$= \min_j \lambda_j(\text{symmetric part of } PAB(x,t)P^{-1}) \quad \text{by Lemma 1-3, (c)}$$

$$= \inf_{y \neq 0} \frac{\langle y, PAB(x,t)P^{-1}y\rangle}{|y|^2} \geq \varepsilon_{AB} > 0 \quad \text{by Lemma A-1.} \quad ◇ \qquad \text{(A-7)}$$

<u>Remark.</u> All the conclusions in Lemma A-1 and Corollary A-2 hold true for $B(x,t)A$ where the order of the product is reversed.

<u>Relation to previous work.</u> Similar results for the product of two constant matrices are found in Oster & Desoer [25] and Chua & Alexander [22].

REFERENCES

1. G. Dahlquist, "Stability and error bounds in the numerical integration of ordinary differential equations," Trans. of The Royal Inst. of Tech., Stockholm, Sweden, No. 130, 1959.

2. W. A. Coppel, Stability and Asymptotic Behavior of Differential Equations. Boston: D. C. Heath & Co., 1965.

3. I. W. Sandberg, "Theorems on the computation of the transient response of nonlinear networks containing transistors and diodes," BSTJ, vol. 49, No. 8, Oct. 1970, pp. 1739-1776.

4. J. M. Ortega and W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables. N. Y.: Academic Press, 1970.

5. C. A. Desoer, Notes for A Second Course on Linear Systems. N. Y.: Van Nostrand Reinhold Co., 1970.

6. R. S. Palais, "Natural operations on differential forms," Trans. Amer. Math. Soc., vol. 92, No. 1, 1959, pp. 125-141.

7. C. A. Holzman and R. W. Liu, "On the dynamical equations of nonlinear networks with n-coupled elements," 1965 Proc. Third Ann. Allerton Conf. Circuit and Syst. Theory(Univ. of Illinois), pp. 536-545, Oct. 1965.

8. T. E. Stern, Theory of Nonlinear Networks and Systems. Reading, Mass.: Addison-Wesley, 1965.

9. A. N. Wilson Jr., "On the solutions of equations for nonlinear resistive networks," BSTJ, vol. 47, No. 8, Oct. 1968, pp. 1755-1773.

10. T. Ohtsuki and H. Watanabe, "State-variable analysis of RLC networks containing nonlinear coupling elements," IEEE Trans. on Circuit Theory, vol. CT-16, No. 1, Feb. 1969, pp. 26-38.

11. E. S. Kuh and I. N. Hajj, "Nonlinear circuit theory: resistive networks," Proc. IEEE, vol. 59, No. 3, March 1971, pp. 340-355.

12. E. Polak, Computational Methods in Optimization: A Unified Approach. N. Y.: Academic Press, 1971.

13. J. Hurt, "Some stability theorems for ordinary difference equations," SIAM J. Numer. Anal., vol. 4, No. 4, 1967, pp. 582-596.

14. H. H. Rosenbrock, "A Lyapunov function for some naturally-occurring linear homogeneous time-dependent equations," Automatica, vol. 1, No. 1, 1963, pp. 97-109.

15. I. W. Sandberg, "Some theorems on the dynamic response of nonlinear transistor networks," BSTJ, vol. 48, No. 1, Jan. 1969, pp. 35-54.

16. D. Mitra and H. C. So' "Linear inequalities and P matrices, with applications to stability theory," Fifth Asilomar Conf. on Circuits and Systems, Nov. 1971.

17. I. W. Sandberg and H. Shichman, "Numerical integration of systems of stiff nonlinear differential equations," BSTJ, vol. 47, No. 4, April 1968, pp. 511-527.

18. F. F. Wu and C. A. Desoer, "Global Inverse Function Theorem," IEEE Trans. on Circuit Theory, vol. CT-19, No.2, March 1972.

19. E. Isaacson and H. B. Keller, Analysis of Numerical Methods. N. Y.: John Wiley, 1966.

20. C. W. Gear, Numerical Initial Value Problems in Ordinary Differential Equations. Englewood Cliffs, N. J.: Prentice-Hall, 1971.

21. C. A. Desoer and M. J. Shensa, "Networks with very small and very large parasitics: Natural frequencies and stability," Proc. IEEE, vol. 58, No. 12, Dec. 1970, pp. 1933-1938.

22. L. O. Chua and G. R. Alexander,"The effects of parasitic reactances on nonlinear networks," IEEE Trans. on Circuit Theory, vol. CT-18, No. 5, Sept. 1971, pp. 520-532.

23. R. G. Bartle, The Elements of Real Analysis. N. Y.: John Wiley, 1964, pp. 311.

24. O. I. Elgerd, Electric Energy Systems Theory: An Introduction. N. Y.: McGraw-Hill, 1971, pp. 82-86.

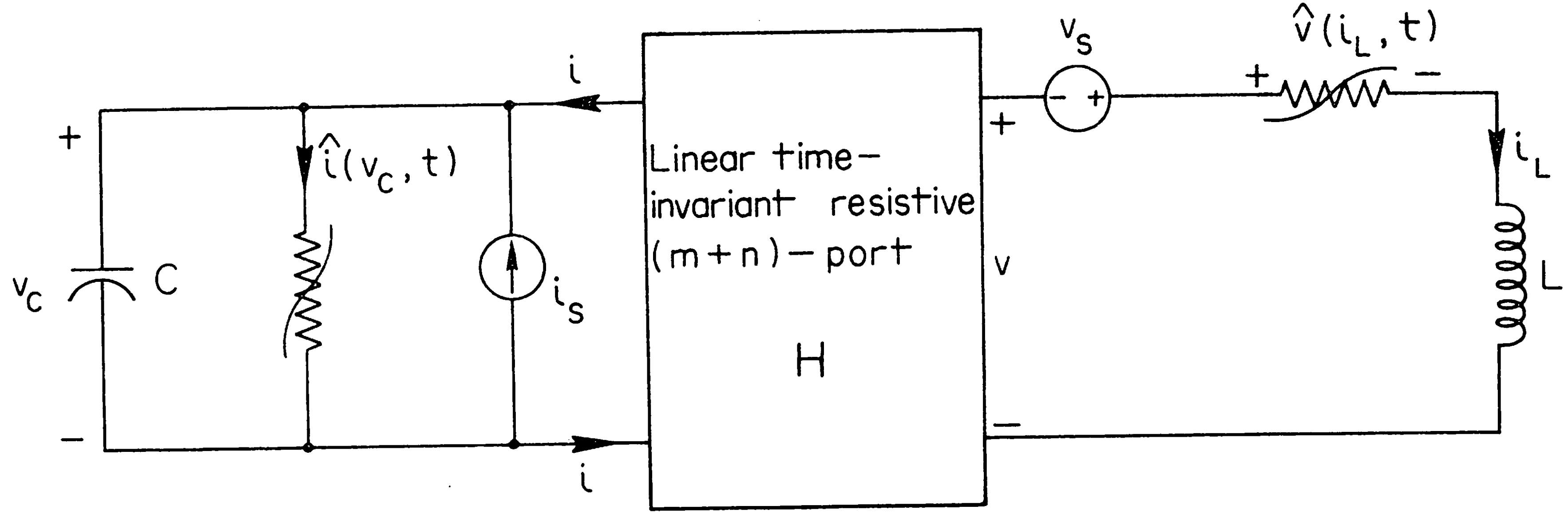
25. G. F. Oster and C. A. Desoer, "Tellegen's theorem and thermodynamic inequalities," J. theor. Biol., vol. 32, 1971, pp. 219-241.

26. H. H. Rosenbrock, "A Lyapunov function with applications to some nonlinear physical systems," Automatica, vol. 1, no. 2/3 1963, pp. 31-53.

27. M. A. Schultz, Control of Nuclear Reactors and Power Plants. N. Y.: McGraw-Hill, 1955, pp.24.

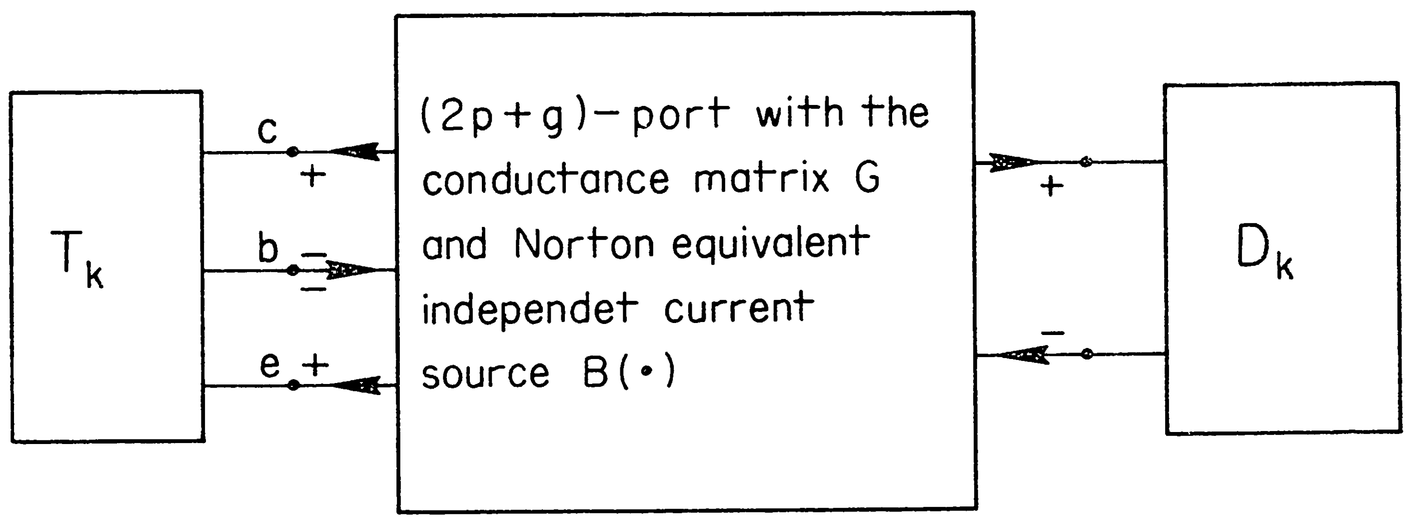
28. L. A. Gould, Chemical Process Control: Theory and Applications. Reading, Mass.: Addison-Wesley, 1969, pp. 192-193.

29. C. A. Desoer and H. Haneda, "The measure of a matrix as a tool to analyze computer algorithms for circuit analysis," to be presented at 1972 IEEE International Symposium on Circuit Theory, April 1972, U. S. A.



i
+
$\hat{i}(v_C, t)$
v_C
C
i_S
i
Linear time-
invariant resistive
(m+n)-port
H
v_S
$\hat{v}(i_L, t)$
+
−
+
i_L
v
L
−
−

Figure 1



T_k
c
+
b
−
−
e
+
(2p+g)−port with the conductance matrix G and Norton equivalent independet current source B(•)
+
−
D_k

Figure 2

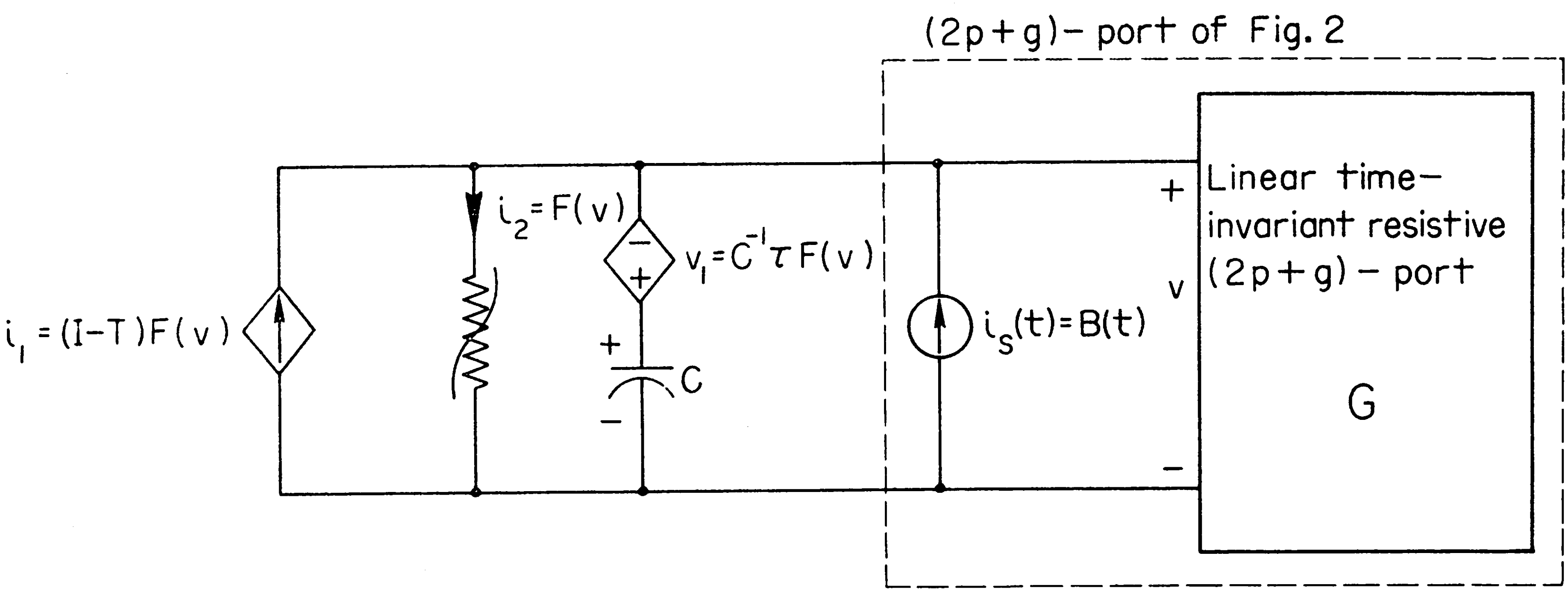
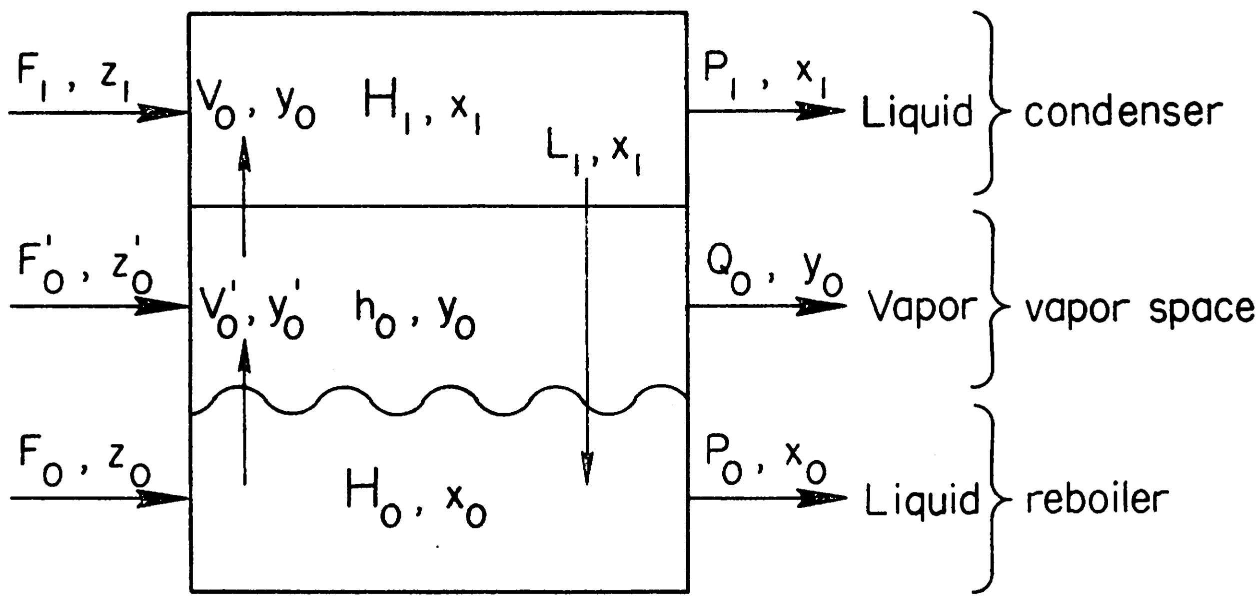

(2p + g) - port of Fig. 2
$i_1 = (I-T)F(v)$
$i_2 = F(v)$
$v_1 = C^{-1}\tau F(v)$
+
−
C
$i_s(t) = B(t)$
+
v
−
Linear time-
invariant resistive
(2p + g) - port
G

Figure 3



F_1 , z_1
V_0 , y_0
H_1 , x_1
L_1 , x_1
P_1 , x_1
Liquid
condenser
F'_0 , z'_0
V'_0 , y'_0
h_0 , y_0
Q_0 , y_0
Vapor
vapor space
F_0 , z_0
H_0 , x_0
P_0 , x_0
Liquid
reboiler

Figure 4

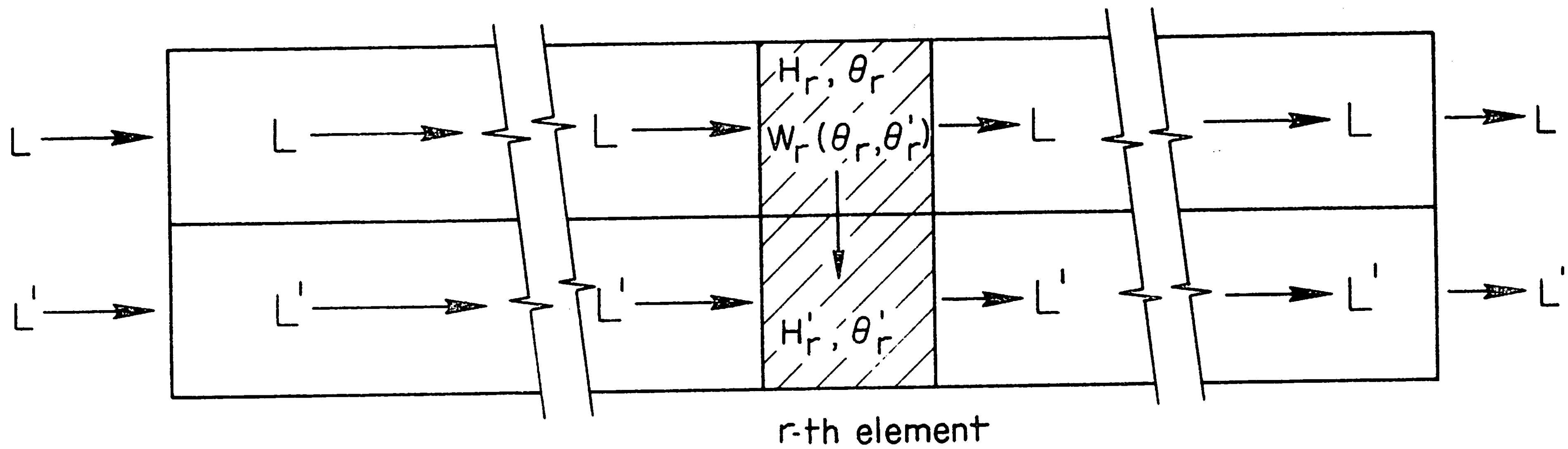

L
L'
L
L'
L
L'
H_r, θ_r
$W_r(\theta_r, \theta_r')$
H_r', θ_r'
L
L'
L
L'
L
L'
r-th element

Figure 5