The Measure of a Matrix as a Tool to Analyze Computer Algorithms for Circuit Analysis

CHARLES A. DESOER, FELLOW, IEEE, AND HIROMASA HANEDA

***Abstract*—The measure of a matrix is used to, first, bound solutions of ordinary differential equations, bound the computer solution by the backward Euler method, and bound the accumulated truncation error; second, to give conditions for the existence and uniqueness of a dc operating point; third, to determine a convergence region for the Newton-Raphson technique and establish its convergence properties. The unifying idea of the paper is the use of the measure of a matrix.**

I. Introduction

THE COMPUTATION of the transient response and of the operating point of nonlinear networks is an important subject in computer-aided circuit analysis. This paper unifies the analysis of computational methods in terms of the measure $\mu(\cdot)$ of a matrix, which was discussed by Dahlquist [1] and Coppel [2]. First, a bound on solutions of some classes of ordinary differential equations, a bound on the solution computed by the backward Euler method, and a bound on the accumulated truncation error are estimated; the estimate given for these bounds is sharper than that obtained by using norms. Second, an existence and uniqueness theorem for dc operating point is given. This theorem includes as special cases well-known results [3]–[6]. Third, the measure $\mu(\cdot)$ of a matrix is used to determine a convergence region of the Newton–Raphson method and quadratic convergence is established. The effect of local roundoff error is also discussed. It should be stressed that not only does our approach unify known results, but also sharpens them.

II. Measure of a Matrix

The measure $\mu(A)$ of a matrix A is a mapping from $\boldsymbol{R}^{d\times d}$ into $\boldsymbol{R}$, [1], [2], which is in some ways analogous to a norm but which leads to sharper estimates of convergence than norms, principally because $\mu(A)$ may take negative values.

Definition

Let A be a $d\times d$ real matrix and $\|\cdot\|$ be an induced norm for matrices. The measure $\mu(A)$ of the matrix A is defined by

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

We can think of $\mu(A)$ as being the one-sided directional derivative of the mapping $\|\cdot\|$: $\boldsymbol{R}^{d\times d}\rightarrow\boldsymbol{R}_+$ at the point I (the $d\times d$ identity matrix), in the direction of A.

The norm of a vector, say x, is denoted by $|x|$, the induced norm of a matrix, say A, by $\|A\|$, and the derivative of a mapping from $\boldsymbol{R}^d$ into itself, say f, at a point $x\in\boldsymbol{R}^d$, by $Df(x)\in\boldsymbol{R}^{d\times d}$. Let $B(a;\, r)$ denote an open ball with a center $a\in\boldsymbol{R}^d$ and radius $r\geq 0$. The closure of the open ball is denoted by $\bar{B}(a;\, r)$. The zero vector in $\boldsymbol{R}^d$ and the zero matrix in $\boldsymbol{R}^{d\times d}$ are denoted by θ_d and $\theta_{d\times d}$, respectively.

The measure $\mu(\cdot)$ has the following properties:

1) For any given $d\times d$ real matrix A, the limit in (1) exists.

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

3) If $A=\theta_{d\times d}$, then $\mu(A)=0$.

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

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

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

7) $\max\{\mu(A)-\mu(-B),\, -\mu(-A)+\mu(B)\}\leq\mu(A+B)\leq\mu(A)+\mu(B)$ (subadditivity).

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

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

10) $|Ax|\geq\max\{-\mu(-A),\ -\mu(A)\}\cdot|x|$ for all $x\in\boldsymbol{R}^d$.

11) Let a vector norm $|\cdot|$ in $\boldsymbol{R}^d$ be given. Define $|x|_P\triangleq|Px|$, where P is a nonsingular $d\times d$ real matrix and call $\mu_P(\cdot)$ the measure defined in terms of the corresponding induced norm. Then $\mu_P(A)=\mu(PAP^{-1})$.

12) If A is a nonsingular $d\times d$ real matrix, then

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

Manuscript received January 3, 1972; revised April 12, 1972. This work was supported by the National Science Foundation under Grant GK-10656X1 and by the Joint Services Electronics Program under Contract F44620-71-C-0087.

C. A. Desoer is with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, Calif. 94720.

H. Haneda was with the Department of Electrical Engineering and Computer Sciences and the Electronics Research Laboratory, University of California, Berkeley, Calif. 94720. He is now with the Department of Electrical Engineering, Kobe University, Kyogo, Japan.

Proof: All properties, except 6), 10), 11), and 12) are in the literature [1], [2]. In fact they are easy to verify.

6) Observe that

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

for sufficiently small $\theta>0$.

10) Observe that for $\theta>0$

$$|Ax|=\frac{|x-(I-\theta A)x|}{\theta}\geq-\frac{\|I-\theta A\|-1}{\theta}|x|.$$

Similarly observe that for $\theta>0$

$$|Ax|\geq-\frac{\|I+\theta A\|-1}{\theta}|x|.$$

11) Observe that

$$\|I+\theta A\|_P=\|I+\theta PAP^{-1}\|.$$

12) Observe that

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

Then the result follows from 10).

Note that the measure $\mu(\cdot)$ is a continuous function from $\boldsymbol{R}^{d\times d}$ into $\boldsymbol{R}$ since it is convex, and that $\mu(\cdot)$ can take on negative values.

Examples [1], [2]

Considering the l^1, l^2, and l^∞ norms in $\boldsymbol{R}^d$, we obtain

$$\mu_1(A)=\max_j\left(a_{jj}+\sum_{\substack{i=1\\(i\neq j)}}^{d}|a_{ij}|\right)$$

$$\mu_2(A)=\max_i\left\{\lambda_i\left(\frac{A+A^T}{2}\right)\right\}$$

and

$$\mu_\infty(A)=\max_i\left(a_{ii}+\sum_{\substack{i=1\\(j\neq i)}}^{d}|a_{ij}|\right).$$

We can easily verify that a $d\times d$ real matrix $A(\omega)$ with a parameter ω in a parametric space Ω is uniformly column-sum dominant (or uniformly positive definite, or uniformly row-sum dominant) iff there exists a positive constant $m>0$ such that $-\mu_1[-A(\omega)]\geq m>0$ (or $-\mu_2[-A(\omega)]\geq m>0$, or $-\mu_\infty[-A(\omega)]\geq m>0$) for all $\omega\in\Omega$, respectively. Sandberg [7] stated the inequality 9) for uniformly column-sum dominant matrices.

III. A Differential Equation

Consider a differential equation

$$\begin{aligned}\dot{x}&=f(x,t)+u(t)\\ x(0)&=x_0\end{aligned}\tag{2}$$

where $x(t)$, $u(t)\in\boldsymbol{R}^d$ for all $t\in\boldsymbol{R}_+$; f: $\boldsymbol{R}^d\times\boldsymbol{R}_+\to\boldsymbol{R}^d$; $u(\cdot)$ and $t\mapsto f(x,t)$ is piecewise continuous (see Appendix). To compute the solution of (2), consider the backward Euler integration formula

$$\begin{aligned}&y_{n+1}=y_n+h\dot{y}_{n+1}, \qquad \text{for all } n\in\boldsymbol{Z}_+\\ &y_0\text{: given}\end{aligned}\tag{3}$$

where $h>0$ is the step size.

Theorem A

Assume that $f(\theta_d,t)=\theta_d$ for all $t\in\boldsymbol{R}_+$ and that $x\mapsto f(x,t)$ is in C^1 for all $t\in\boldsymbol{R}_+$. Under these conditions if there exist some vector norm in $\boldsymbol{R}^d$ and a constant $m>0$ such that $-\mu[D_1f(x,t)]\geq m>0$ for all $x\in\boldsymbol{R}^d$, for all $t\in\boldsymbol{R}_+$ (where $D_1f(x,t)$ is the derivative of the mapping $x\mapsto f(x,t)$), then

1)

$$|x(t)|\leq\exp(-mt)|x_0|+\int_0^t\exp[-m(t-\tau)]\cdot|u(\tau)|\,d\tau \qquad ([1],[2]) \tag{4}$$

2)

$$|y_n|\leq(1+mh)^{-n}|y_0|+\sum_{k=0}^{n-1}(1+mh)^{-(k+1)}\cdot h\cdot|u_{n-k}| \tag{5}$$

3) if, in addition, for any fixed $x\in\boldsymbol{R}^d$, $D_2f(x,\cdot)$ (where $D_2f(x,\cdot)$ is the derivative of the mapping $t\mapsto f(x,t)$ is piecewise continuous and $\dot{u}(\cdot)$ is piecewise continuous, both $u(\cdot)$ and $\dot{u}(\cdot)$ are bounded on $\boldsymbol{R}_+$, and there exist constants k_1 and l_1 such that $\|D_1f(x,t)\|\leq k_1$ and $|D_2f(x,t)|\leq l_1$ for all $x\in\boldsymbol{R}^d$, $\forall t\in\boldsymbol{R}_+$, then there exists $\rho>0$ independent of h such that

$$|x_n-y_n|\leq(1+mh)^{-n}|x_0-y_0|+\rho h, \qquad n\geq 1. \tag{6}$$

Theorem A shows that the exact and computed solutions are both exponentially stable and that the accumulated truncation error has an upper bound. In the estimate of this bound the effect of the initial error decays exponentially and the effect of the local truncation error does not build up indefinitely. Using the same technique, it is easy to show that the difference between two solutions of (2) due to different bounded inputs and different initial conditions converges to 0 as $t\to\infty$ if the difference between the two inputs converges to 0. The flexibility resulting from the freedom of choice for the vector norm in $\boldsymbol{R}^d$ should be stressed. In fact, many authors have shown the same conclusion for special cases: l^1, l^2, and weighted l^1 norms, i.e., $|x|_D = |Dx|_1$, $D \triangleq \text{diag}\,(d_1, d_2, \cdots, d_d) > 0$ (Rosenbrock [8], Sandberg and Shichman [9], Sandberg [7], [10], Mitra and So [11]).

Proof: 1) Observe that the solution $x(\cdot)$ of (2) is equal to the solution of the following linear time varying differential equation:

$$\dot{x} = A(t)x + u(t)$$
$$x(0) = x_0 \tag{7}$$

where

$$A(t) \triangleq \int_0^1 D_1 f(\tau x(t), t)\,d\tau. \tag{8}$$

By assumption and properties of $\mu(\cdot)$, we obtain

$$\mu[A(t)] \leq -m < 0, \qquad \forall t \in R_+.$$

Then Coppel's inequality (see Lemma A in the Appendix) is applied to conclude the inequality (4).

2) From (2) and (3), we obtain

$$y_{n+1} - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\,d\tau \cdot y_{n+1} = y_n + hu_{n+1} \tag{9}$$

where we used Taylor's formula for $f(y_{n+1}, n+1)$. As before, observe that

$$\mu\left[\int_0^1 D_1 f(\tau y_{n+1}, n+1)\,d\tau\right] \leq -m < 0,$$
$$\text{for all } y_{n+1} \in \boldsymbol{R}^d, \quad \text{for all } n \geq 0. \tag{10}$$

So

$$-\mu\left[-I + h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\,d\tau\right] = 1 - h\mu\left[\int_0^1 D_1 f(\tau y_{n+1}, n+1)\,d\tau\right]$$
$$\geq 1 + mh > 1, \qquad n \geq 0. \tag{11}$$

By using properties of $\mu(\cdot)$, (9), and (11), we obtain

$$|y_n| + h|u_{n+1}| \geq |y_n + hu_{n+1}| = \left|\left[I - h\int_0^1 D_1 f(\tau y_{n+1}, n+1)\,d\tau\right] y_{n+1}\right|$$
$$\geq (1+mh)\,|y_{n+1}|. \tag{12}$$

Since $1+mh>1>0$, the inequality (12) becomes

$$|y_{n+1}| \leq (1+mh)^{-1}|y_n| + (1+mh)^{-1}|u_{n+1}|, \qquad n \geq 0. \tag{13}$$

Thus the bound (5) of the computed solution follows.

3) Define the local truncation error $\{\xi_n\}_0^\infty$ by

$$\xi_n \triangleq x_{n+1} - x_n - h\dot{x}_{n+1}, \qquad n \geq 0. \tag{14}$$

First, note that $x(\cdot)$ and $u(\cdot)$ are bounded on $[0, \infty)$. Differentiate both sides of (2) with respect to t:

$$\ddot{x}(t) = D_1 f(x, t) \cdot [f(x, t) + u(t)] + D_2 f(x, t) + \dot{u}(t). \tag{15}$$

So

$$|\ddot{x}(t)| \leq \|D_1 f(x, t)\| \cdot \left\{\int_0^1 \|D_1 f(\tau x, t)\|\,d\tau \cdot |x(t)| + |u(t)|\right\} + |D_2 f(x, t)| + |\dot{u}(t)| < \infty,$$
$$\text{for all } t \in R_+. \tag{16}$$

By the definition of ξ_n in (14) and Taylor's formula applied to each component

$$\xi_n = -\tfrac{1}{2}h^2 U_n \tag{17}$$

where the jth component $[U_n]_j$ of U_n is equal to the jth component $\ddot{x}_j$ of $\ddot{x}$ evaluated at some point of $[nh, (n+1)h]$. Since by (15), $\ddot{x}$ is bounded on $[0, \infty)$, U_n is bounded; thus $|U_n| \leq \rho_1$ for some $\rho_1 > 0$ for all $n \geq 0$. Hence

$$|\xi_n| \leq \tfrac{1}{2}h^2\rho_1, \qquad \text{for all } n \geq 0. \tag{18}$$

Now, from (2), (3), and (14),

$$(y_{n+1} - x_{n+1}) - h\int_0^1 D_1 f[(1-\tau)x_{n+1} + \tau y_{n+1}, n+1]\,d\tau \cdot (y_{n+1} - x_{n+1}) = y_n - x_n - \xi_n \tag{19}$$

where we again used Taylor's formula for $f(y_{n+1}, n+1) - f(x_{n+1}, n+1)$. Analogously for the proof of part 2), we obtain

$$|y_{n+1} - x_{n+1}| \leq (1+mh)^{-1}|y_n - x_n| + (1+mh)^{-1}|\xi_n|. \tag{20}$$

Hence

$$|y_n - x_n| \leq (1+mh)^{-n}|y_0 - x_0| + \frac{1}{2}h^2\rho_1 \sum_{k=1}^{n}(1+mh)^{-k}$$
$$\leq (1+mh)^{-n}|y_0 - x_0| + \frac{1}{2}h^2\rho_1 \sum_{k=1}^{\infty}(1+mh)^{-k}$$
$$= (1+mh)^{-n}|y_0 - x_0| + \frac{1}{2}h^2\rho_1 \frac{1}{mh}\cdot \tag{21}$$

Let $\rho \triangleq \rho_1/(2m)$, then the inequality (6) follows.

IV. Existence and Uniqueness of DC Operating Point

Consider a dc equation

$$f(x) = y \tag{22}$$

where f: $\boldsymbol{R}^d \rightarrow \boldsymbol{R}^d$, $x \in \boldsymbol{R}^d$, and $y \in \boldsymbol{R}^d$ is any given point.

An equation such as (22) has to be solved when calculating dc operating points and at each step of any implicit algorithm for solution of ordinary differential equations such as the backward Euler method.

Theorem B

Let f: $\boldsymbol{R}^d \to \boldsymbol{R}^d$ be in C^1. Assume that there exists a function m: $\boldsymbol{R}_+ \to \boldsymbol{R}_+$ with $m(\alpha) > 0$ for all $\alpha \in \boldsymbol{R}_+$ and $\int_0^\infty m(\alpha) d\alpha = \infty$ such that either $-\mu[Df(x)] \geq m(|x|) > 0$ or $-\mu[-Df(x)] \geq m(|x|) > 0$ for all $x \in \boldsymbol{R}^d$. Then, f is a C^1-diffeomorphism from $\boldsymbol{R}^d$ onto itself.

Proof: We use Palais' global inverse function theorem in the proof [12], [18]. For any nonzero vector $z \in \boldsymbol{R}^d$

$$\begin{aligned} | Df(x) \cdot z | &\geq \max \{ -\mu[-Df(x)], -\mu[Df(x)] \} \cdot | z | \\ &\geq m(|x|) \cdot |z| > 0. \end{aligned} \tag{23}$$

Thus det $[Df(x)] \neq 0$, $\forall x \in \boldsymbol{R}^d$. Next, we want to show $|f(x)| \to \infty$ as $|x| \to \infty$.

By Taylor's formula and properties of $\mu(\cdot)$

$$\begin{aligned} |f(x)| &\geq \left| \left[\int_0^1 Df(\tau x) d\tau \right] \cdot x \right| - \left| f(\theta_d) \right| \\ &\geq \max \left\{ -\mu \left[- \int_0^1 Df(\tau x) d\tau \right], \right. \\ &\qquad \left. -\mu \left[\int_0^1 Df(\tau x) d\tau \right] \right\} \cdot \left| x \right| - \left| f(\theta_d) \right|. \end{aligned} \tag{24}$$

By assumption, we observe

$$\begin{aligned} \max \left\{ -\mu \left[- \int_0^1 Df(\tau x) d\tau \right], -\mu \left[\int_0^1 Df(\tau x) d\tau \right] \right\} \\ \geq \int_0^1 m(|\tau x|) d\tau = \int_0^{|x|} \frac{m(\alpha) d\alpha}{|x|}, \\ \text{by letting } \alpha = |\tau x| = \tau |x|. \end{aligned} \tag{25}$$

From (24) and (25)

$$|f(x)| \geq \int_0^{|x|} m(\alpha) d\alpha - |f(\theta_d)| \to \infty, \qquad \text{as } |x| \to \infty. \tag{26}$$

Hence the result follows from Lemma B.

Theorem B is a generalization of previous results in two directions. First, any vector norm in $\boldsymbol{R}^d$ can be used for testing the condition of the theorem. Second, if, for example, with the l^2 norm in $\boldsymbol{R}^d$, $Df(x)$ is neither uniformly positive definite nor uniformly negative definite, there may be cases where an $m(\cdot)$ is found to satisfy the condition of the theorem, e.g., $m(|x|) = (\alpha + \beta |x|)^{-1}$, $\alpha > 0$ and $\beta > 0$. Also, Theorem B can be used to show that the implicit equation obtained from the backward Euler method has a unique solution. Several authors have obtained special cases of Theorem B; most considered $m(\cdot)$ constant: Stern [3], l^∞; Willson [4], l^∞; Ohtsuki and Watanabe [5], l^2; Kuh and Hajj [6], l^2, and Sandberg used a variable $m(\cdot)$ in a different approach [7].

V. Newton–Raphson Method

Consider the difference equation

$$x_{k+1} = f(x_k) \tag{27}$$

where $x_k \in \boldsymbol{R}^d$ for all $k \in \boldsymbol{Z}_+$ and f: $\boldsymbol{R}^d \to \boldsymbol{R}^d$ is continuous. Consequently, for all $x_0 \in \boldsymbol{R}^d$, the solution $\{x(k; x_0)\}_0^\infty$ of (27) is uniquely defined and for each fixed $k \in \boldsymbol{Z}_+$, $x_0 \mapsto x(k, x_0)$ is continuous.

Convergence of the Newton–Raphson Method

Problem: Let f: $\boldsymbol{R}^d \to \boldsymbol{R}^d$ and $f \in C^1$. Given some $y \in \boldsymbol{R}^d$ find x^* such that $f(x^*) = y$. The method uses the following iteration:

$$\begin{aligned} & x_{k+1} = x_k - [Df(x_k)]^{-1}[f(x_k) - y], \qquad k = 0, 1, 2, \cdots. \\ & x_0 \text{: given.} \end{aligned} \tag{28}$$

Theorem C

Suppose that

1) f: $\boldsymbol{R}^d \to \boldsymbol{R}^d$ and $f \in C^1$;

2) for some $m > 0$ either $-\mu[Df(x)] \geq m > 0$ or $-\mu[-Df(x)] \geq m > 0$, $\forall x \in \boldsymbol{R}^d$;

3) there exists a continuous monotone increasing function $k^*(\cdot)$: $\boldsymbol{R}_+ \to \boldsymbol{R}_+$ such that $\forall r > 0$

$$\| Df(u) - Df(v) \| \leq k^*(r) \, | u - v |, \qquad \forall u, v \in B(x^*; r).$$

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 $\{x_k\}_0^\infty$ defined by (28) lies in $B(x^*; r^*)$ and converges to the unique solution x^* at least quadratically.

By the definition of r^*, the convergence region is enlarged if either m becomes larger or if $f(\cdot)$ becomes smoother, i.e., $k^*(r)$ is decreased for each fixed $r > 0$. If $k^*(\cdot)$ is a constant function, r^* becomes $2m/k^*$, and the effect of m and k^* on the convergence region is obvious. The key point in the above analysis is that we can find a uniform upper bound on $\| [Df(x)]^{-1} \|$ using properties of $\mu(\cdot)$ under these assumptions.

Proof: b) implies that f is a C^1 diffeomorphism of $\boldsymbol{R}^d$ onto itself; hence for any $y \in \boldsymbol{R}^d$, $f(x) = y$ has one and only one solution. Let

$$e_k \triangleq x^* - x_k, \qquad \forall k \in \boldsymbol{Z}_+.$$

So

$$\begin{aligned} e_{k+1} &= x^* - x_{k+1} \\ &= x^* - x_k + [Df(x_k)]^{-1}[f(x_k) - y] \\ &= [Df(x_k)]^{-1}[Df(x_k)(x^* - x_k) + f(x_k) - f(x^*)] \\ &= [Df(x_k)]^{-1} \left[Df(x_k)(x^* - x_k) - \int_0^1 Df(x^* - \tau e_k) d\tau \cdot e_k \right] \end{aligned}$$

where we used Taylor's formula. In terms of e_k, we obtain

$$e_{k+1} = [Df(x^* - e_k)]^{-1} \cdot \left\{ \int_0^1 [Df(x^* - e_k) - Df(x^* - \tau e_k)] d\tau \cdot e_k \right\}. \quad (29)$$

To study the convergence of the procedure, let $V(e) = |e|$. From (29) with $\Delta V(e_k) \triangleq V(e_{k+1}) - V(e_k)$, $\forall k \in \boldsymbol{Z}_+$, we obtain

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

As a consequence of 2) and the properties of $\mu(\cdot)$, the first factor is smaller than or equal to $1/m$. Furthermore, using 3) we observe that $\forall r > 0$,

$$\begin{aligned} &\left| \int_0^1 [Df(x^* - e) - Df(x^* - \tau e)] d\tau \cdot e \right| \\ &\leq \int_0^1 \|Df(x^* - e) - Df(x^* - \tau e)\| d\tau \cdot |e| \\ &\leq \frac{k^*(r) |e|^2}{2}, \qquad \text{where } x^* \text{ and } x^* - e \in B(x^*; r). \end{aligned}$$

Hence for $r \geq |e|$

$$\Delta V(e) \leq |e| \left(\frac{k^*(r)}{2m} |e| - 1 \right). \quad (30)$$

Observe that for all $0 \leq |e| < r^*$,

$$\Delta V(e) \leq |e| (|e| (r^*)^{-1} - 1) < 0. \quad (31)$$

Consider any sequence $\{e_k\}_0^\infty$ defined by (29) and starting from some e_0 subject to $|e_0| \leq \gamma < r^*$.
By (31),

$$\begin{aligned} V(e_{k+1}) = |e_{k+1}| &= |e_k| + \Delta V(e_k) \\ &\leq |e_k|^2 (r^*)^{-1}. \end{aligned}$$

By induction

$$|e_k| \leq \left(\frac{\gamma}{r^*} \right)^{2k} r^*, \qquad k = 0, 1, 2, \cdots. \quad (32)$$

So the sequence converges to 0 since $\gamma < r^*$. In terms of the iterates,

$$|x_{k+1} - x^*| \leq |x_k - x^*|^2 (r^*)^{-1}. \quad (33)$$

Hence for all x_0 such that $|x_0 - x^*| \leq \gamma < r^*$

$$\limsup_{k \to \infty} \frac{|x_{k+1} - x^*|}{|x_k - x^*|^2} \leq (r^*)^{-1}.$$

Remark: Since $x^* \in \boldsymbol{R}^d$ is unknown *a priori*, the condition 3) of Theorem C and the condition $x_0 \in B(x^*; r^*)$ are impossible to check *a priori*. These conditions can be replaced by weaker conditions which do not include the unknown x^*.

Corollary A

Suppose that f: $\boldsymbol{R}^d \to \boldsymbol{R}^d$ satisfies conditions 1) and 2) of Theorem C. Given $x_0 \in \boldsymbol{R}^d$, assume further that

1') there exists a continuous monotone increasing function $k_0(\cdot)$: $\boldsymbol{R}_+ \to \boldsymbol{R}_+$ such that $\forall r > 0$

$$\|Df(u) - Df(v)\| \leq k_0(r) |u - v|, \qquad \forall u, v \in B(x_0; r).$$

Define r_0^* to be the unique solution of

$$r = \frac{2m}{k_0\left(r + \frac{|f(x_0) - y|}{m} \right)}, \qquad r > 0. \quad (34)$$

Under these conditions, if $|f(x_0) - y| \leq m_0^*$, then the corresponding sequence $\{x_k\}_0^\infty$ defined by (28) remains in $B(x^*; r_0^*)$ and converges to the unique solution x^* at least quadratically.

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

Observe that

$$|f(x) - y| = \left| \int_0^1 Df(x^* + \tau(x - x^*)) d\tau \cdot (x - x^*) \right|.$$

Using properties of $\mu(\cdot)$, 1) and 2), we obtain

$$|f(x) - y| \geq m|x - x^*|.$$

Claim 1')→1).

Considering only $r_0 > |x^* - x_0|$, let $r \triangleq r_0 - |x^* - x_0|$. Hence $B(x^*; r) \subset B(x_0; r_0)$.

The condition 1') implies that $\forall r_0 > |x^* - x_0|$

$$\|Df(u) - Df(v)\| \leq k_0(r_0) |u - v|, \qquad \forall u, v \in B(x_0, r_0).$$

So $\forall r > 0$

$$\begin{aligned} &\|Df(u) - Df(v)\| \\ &\leq k_0(r + |x^* - x_0|) |u - v|, \qquad \forall u, v \in B(x^*; r) \subset B(x_0, r_0) \\ &\leq k_0\left(r + \frac{|f(x_0) - y|}{m} \right) \cdot (u - v|, \quad \forall u, v \in B(x^*; r) \subset B(x_0, r_0). \end{aligned}$$

Let

$$k^*(r) \triangleq k_0\left(r + \frac{|f(x_0) - y|}{m} \right).$$

Noting that the function $k^*(\cdot)$ is continuous and monotone increasing we obtain the condition 3) of Theorem C. Also

$$\{x \in \boldsymbol{R}^d \mid |f(x) - y| \leq m r_0^*\} \subset B(x^*; r_0^*).$$

Q.E.D.

In order to discuss the effect of roundoff errors, we use a slightly modified version of Hurt's lemma [15].

Lemma B (Hurt, [15])

Let V and W map $\boldsymbol{R}^d \rightarrow \boldsymbol{R}$ and let W be continuous. For some fixed $\gamma > 0$, let $G \triangleq \{x \in \boldsymbol{R}^d | \; V(x) \leq \gamma\}$. Assume further that

1) $V \geq 0$ on G;
2) G is compact;
3) there exists a constant $\omega \geq 0$ such that

$$\Delta V(x) \triangleq V[f(x)] - V(x) \leq W(x) \leq \omega, \qquad \forall x \in G;$$

4) let $N \triangleq \{x \in G | \; W(x) \leq 0\}$ and $b \triangleq \sup_{x \in N} V(x) < \infty$;
5) let $A \triangleq \{x \in \boldsymbol{R}^d | \; V(x) \leq b + \omega\}$; $b + \omega < \gamma$.

Let

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

Under these conditions,

a) $N \subset A \subset G$, N is closed and $\delta \geq 0$;
b) $\forall x_0 \in G$, $x(k; x_0) \in G$, $\forall k \in \boldsymbol{Z}_+$;
c) $\forall x_0 \in G$, $x(k; x_0) \rightarrow A$ as $k \rightarrow \infty$ and A is an invariant set of (27). 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) $\forall x_0 \in G$, the positive limit set $M(x_0)$ of $\{x(k; x_0)\}_0^\infty$, is a subset of A and $M(x_0)$ is an invariant set of (27).

Convergence with Roundoff Errors Present

Let $\epsilon(x_k)$ denote the local roundoff error, then the difference equation becomes

$$x_{k+1} = x_k - [Df(x_k)]^{-1}[f(x_k) - y] + \epsilon(x_k). \tag{35}$$

We assume that there is an $\epsilon_\infty > 0$ such that

$$|\epsilon(x_k)| \leq \epsilon_\infty, \qquad \text{for all } x_k \text{ under consideration.} \tag{36}$$

Theorem D

Assume that f satisfies conditions 1)–3) of Theorem C. Assume that the roundoff error is bounded as in (36) and that

$$\epsilon_\infty < r^*/5.$$

Under these conditions, if x_0 is such that $|x_0 - x^*| \leq r^* - 2\epsilon_\infty$, then the sequence $\{x_k\}_0^\infty$ computed by (35) remains in $\overline{B}(x^*; r^* - 2\epsilon_\infty)$ and enters the closed ball $\overline{B}(x^*; 3\epsilon_\infty)$ after a finite number of steps and remains in it forever after.

If the local roundoff error ϵ_∞ is sufficiently small, then the radius of the convergence region is $2\epsilon_\infty$ 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 around x^* with a radius $3\epsilon_\infty$ in a finite number of steps.

Proof: From (35), we can derive a relation analogous to (29):

$$e_{k+1} = [Df(x^* - e_k)]^{-1} \cdot \int_0^1 [Df(x^* - e_k) - Df(x^* - \tau e_k)] d\tau \cdot e_k \Big\} + \epsilon(x^* - e_k).$$

As above, let $V(e) = |e|$ and obtain

$$\Delta V(e) \leq |e| \left(\frac{k^*(r)}{2m} |e| - 1 \right) + |\epsilon(x^* - e_k)|.$$

Hence for all $0 \leq |e| < r^*$

$$\Delta V(e) \leq |e| (|e| (r^*)^{-1} - 1) + \epsilon_\infty. \tag{37}$$

In order to apply the lemma, call $-W(e)$ the right-hand side of (37). W is then continuous and $\omega = \epsilon_\infty$. We choose $\gamma = r^* - 2\epsilon_\infty$. From the definition, b is the smallest zero of $|e|^2 (r^*)^{-1} - |e| + \epsilon_\infty = 0$; so

$$b = \epsilon_\infty + \epsilon_\infty^2 (r^*)^{-1} + \cdots < 2\epsilon_\infty.$$

Here $A = \{e \in \boldsymbol{R}^d | \; |e| \leq b + \omega < 3\epsilon_\infty\} \subset G$. Note that

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

Hence all the conditions of the lemma are satisfied and consequently for any $x_0 \in \overline{B}(x^*; r^* - 2\epsilon_\infty)$, the sequence $\{x_k\}$ remains in it and enters $B(x^*; 3\epsilon_\infty)$ after a finite number of steps and remains in it forever after.

Similar to Corollary A, we obtain the following corollary from Theorem D.

Corollary B

Assume that f satisfies conditions 1)–3) of Corollary A. Assume that the local roundoff error is bounded as in (36) and that $\epsilon_\infty < r_0^*/5$. Under these conditions, if x_0 is such that $|f(x_0) - y| \leq m(r_0^* - 2\epsilon_\infty)$, then the sequence $\{x_k\}_0^\infty$ computed by (35) remains in $\overline{B}(x^*; r_0^* - 2\epsilon_\infty)$, and enters the closed ball $\overline{B}(x^*; 3\epsilon_\infty)$ after a finite number of steps and remains in it forever after.

Conclusion

We have shown that the concept of the measure $\mu(\cdot)$ of a matrix unifies a number of known results on the analysis of dc solutions and transient response of circuits and exhibits the common structure between the dc equation and the dynamic equation. The simplicity and flexibility of the measure $\mu(\cdot)$ led us to more general results.

Appendix

Definition of Piecewise Continuity

A function f: $\boldsymbol{R}_+ \rightarrow \boldsymbol{R}^n$ is said to be piecewise continuous on $\boldsymbol{R}_+$ iff, on every compact subinterval $I = [a, b]$ of $\boldsymbol{R}_+$, 1) f is continuous on I except for at most a finite number of points; 2) if $c \in (a, b)$ is a point of discontinuity of f, then both $f(c-0)$ and $f(c+0)$ exist and are finite; and 3) $f(a+0)$ and $f(b-0)$ exist (Bartle [16], Desoer [17]).

Lemma A (slightly extended from Coppel, [2])

Let $A(t) \in \boldsymbol{R}^{d \times d}$ for all $t \in \boldsymbol{R}_+$ and $t \mapsto A(t)$ be piecewise continuous. Then bounds of the state transition matrix

$\Phi(t, t_0)$ associated with $A(\cdot)$ are given by

$$\exp\left[-\int_{t_0}^{t} \mu[-A(s)]ds\right] \leq \frac{1}{\|[\Phi(t, t_0)]^{-1}\|} \leq \|\Phi(t, t_0)\| \leq \exp\left[\int_{t_0}^{t} \mu[A(s)]ds\right], \quad \text{for all } t \geq t_0.$$

References

[1] G. Dahlquist, "Stability and error bounds in the numerical integration of ordinary differential equations," *Trans. Roy. Inst. Tech.* (Sweden), no. 130, 1959.

[2] W. A. Coppel, *Stability and Asymptotic Behavior of Differential Equations*. Boston, Mass.: D. C. Heath, 1965.

[3] T. E. Stern, *Theory of Nonlinear Networks and Systems*. Reading, Mass.: Addison-Wesley, 1965.

[4] A. N. Willson, Jr., "On the solutions of equations for nonlinear resistive networks," *Bell Syst. Tech. J.*, vol. 47, pp. 1755–1773, Oct. 1968.

[5] T. Ohtsuki and H. Watanabe, "State-variable analysis of RLC networks containing nonlinear coupling elements," *IEEE Trans. Circuit Theory*, vol. CT-16, pp. 26–38, Feb. 1969.

[6] E. S. Kuh and I. N. Hajj, "Nonlinear circuit theory: Resistive networks," *Proc. IEEE*, vol. 59, pp. 340–355, Mar. 1971.

[7] I. W. Sandberg, "Theorems on the computation of the transient response of nonlinear networks containing transistors and diodes," *Bell Syst. Tech. J.*, vol. 49, pp. 1739–1776, Oct. 1970.

[8] H. H. Rosenbrock, "A Lyapunov function for some naturally-occurring linear homogeneous time-dependent equations," *Automatica*, vol. 1, pp. 97–109, 1963.

[9] I. W. Sandberg and H. Shichman, "Numerical integration of systems of stiff nonlinear differential equations," *Bell Syst. Tech. J.*, vol. 47, pp. 511–527, Apr. 1968.

[10] I. W. Sandberg, "Some theorems on the dynamic response of nonlinear transistor networks," *Bell Syst. Tech. J.*, vol. 48, pp. 35–54, Jan. 1969.

[11] D. Mitra and H. C. So, "Linear inequalities and P matrices, with applications to stability theory," presented at the 5th Asilomar Conf. Circuits and Systems, Nov. 1971.

[12] R. S. Palais, "Natural operations on differential forms," *Trans. Am. Math. Soc.*, vol. 92, pp. 125–141, 1959.

[13] C. A. Holzman and R. W. Liu, "On the dynamical equations of nonlinear networks with n-coupled elements," in *1965 Proc. 3rd Annu. Allerton Conf. Circuit and Syst. Theory*, pp. 537–545, Oct. 1965.

[14] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*. New York: Academic Press, 1970.

[15] J. Hurt, "Some stability theorems for ordinary difference equations," *SIAM J. Numer. Anal.*, vol. 4, pp. 582–596, 1967.

[16] R. G. Bartle, *The Elements of Real Analysis*. New York: Wiley, 1964, pp. 311.

[17] C. A. Desoer, *Notes for a Second Course on Linear Systems*. New York: Van Nostrand, 1970, pp. 7.

[18] F. F. Wu and C. A. Desoer, "Global inverse function theorem," *IEEE Trans. Circuit Theory* (Corresp.), vol. CT-19, pp. 199–201, Mar. 1972.

Corrections to "The Measure of a Matrix as a Tool to Analyze Computer Algorithms for Circuit Analysis"

C. A. DESOER AND H. HANEDA

The authors of the above paper[1] would like to correct the following misprints.

481l, bottom line	replace 9) by 10).
481r, 3 lines below (5)	should read: $t \mapsto f(x, t)) \cdots$.
482l, line 3	delete line 3–5; should read: $D \triangleq \operatorname{diag}(d_1 d, d_2, \cdots, d_d) > 0$. Inequality (4) was obtained by Dahlquist [1], Coppel [2], Rosenbrock [8], using l^1 norms, Sandberg [10], and Mitra and So [11], both using weighted l^1 norms. Inequality (5) was obtained by Sandberg and Shichman [9], using l^2 norms and Sandberg [7], using weighted l^1 norms. Inequality (6) was obtained by Sandberg [7], using weighted l^1 norms.
483l, one line below (26)	should read: $\cdots$ follows from the global inverse function theorem.
483r, proof of Theorem C	should read: Proof: 2) $\cdots$.
484r, 4 lines above (34)	should read: 3') $\cdots$.
484r, one line below (34)	should read: $\cdots$, if $\lvert f(x_0) - y \rvert \leq m r_0^*, \cdots$.
484r, 10 lines below (34)	should read: Claim 3') → 3).
484r, 13 lines below (34)	should read: The condition 3') $\cdots$.
484r, 9 lines from bottom	should read: $\leq k_0 \; (r + (\lvert f(x_0) - y \rvert / m)) \cdot \lvert u - v \rvert, \cdots$.
485l, line 8	should read: $\cdots \leq -W(x) \cdots$.
485l, 2 lines from bottom	should read: $\cdot \; \{ \int_0^1 \cdots \}$.
485r, first line of Corollary B	should read: $\cdots$ 1) – 3') $\cdots$.

Manuscript received June 8, 1973.

C. A. Desoer is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Calif. 94720.

H. Haneda is with the Department of Electrical Engineering, Kobe University, Kyogo, Japan.

[1] C. A. Desoer and H. Haneda, *IEEE Trans. Circuit Theory*, vol. CT-19, pp. 480–486, Sept. 1972.