BIT 12 (1972), 172–181

CONDITIONING OF LINEAR BOUNDARY VALUE PROBLEMS[1]

JOHN H. GEORGE[2] and ROBERT W. GUNDERSON

Abstract.

The method of parallel shooting for the solution of two-point boundary value problems is investigated. Bounds are obtained for the norms of the fundamental matrices of the differential equations and their inverses. These bounds are used for estimation of the condition number and for determining the shooting intervals.

1. Introduction.

Several recent papers (e.g. Roberts and Shipman [1], Osborne [2], Bailey and Shampine [3]) have been concerned with shooting methods for the numerical solution of boundary value problems. In general, these methods solve the boundary value problems by reducing it to a corresponding initial value problem. In the linear case, the subject of this paper, this process results in a system of linear algebraic equations which must be solved for the initial value of the solution. Thus, computational difficulty will be encountered in the event these equations are poorly conditioned. In this paper, the method of parallel shooting (cf. Keller [4]) is shown to be useful in avoiding conditioning problems. Easily computed bounds are obtained for the norms of the fundamental matrices of the differential equations and their inverses. These bounds are then used to obtain estimates for the condition number of the system of linear equations. Finally, these estimates are used to determine the parallel shooting intervals.

2. Preliminaries.

Given a norm for the n-vector x, the corresponding induced norm of the $n \times n$ matrix A will be defined by

$$\|A\| = \sup_{\|x\|=1} \|Ax\|$$

and the measure of the matrix (Dahlquist [5]) by

Received February 12, 1972.

[1] This research was supported by NASA Grant No. NGR-002-016.

[2] Visiting Utah State University, Summer, 1971.

$$\mu[A] = \lim_{h \to +0} \frac{\|I + hA\| - 1}{h}$$

where I denotes the $n \times n$ unit matrix. In the sequel, the vector norms used will be those for which the induced norm of I is unity. For example, if $\|x\|_1 = \sum_{i=1}^{n} |x_i|$, then

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}| \quad \text{and} \quad \mu_1[A] = \max_j \left[a_{jj} + \sum_{\substack{i \neq j \\ i=1}}^{n} |a_{ij}| \right]$$

and, if $\|x\|_\infty = \max_i |x_i|$, then

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}| \quad \text{and} \quad \mu_\infty[A] = \max_i \left[a_{ii} + \sum_{\substack{j \neq i \\ j=1}}^{n} |a_{ij}| \right].$$

If a specific norm is not required then no subscript will be used.

THEOREM 2.1 (Noble [7]). *Consider the system of equations $Ax = b$, when A is an $n \times n$ matrix and x, b are n-vectors. Suppose that $A + \delta A$, $b + \delta b$ change the solution vector to $x + \delta x$. Then, if $\|\delta A\| \|A^{-1}\| \ll 1$,*

$$\frac{\|\delta x\|}{\|x\|} \leqq C M[A] \left[\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right]$$

where

$$M[A] = \|A\| \|A^{-1}\|$$

is called the condition number *of A and*

$$C = (1 - \|\delta A\| \|A^{-1}\|)^{-1}.$$

REMARK. In the standard references (e.g. Noble [7] and Wilkinson [8]), a system of equations, $Ax = b$, is said to be *well-conditioned* if the solution is relatively insensitive to small changes in the equation. Similarly, it is said to be *ill-conditioned* if small variations lead to relatively large changes in the solution. From the above result, it follows that if C is close to unity, then a small condition number indicates the equation is well-conditioned.

3. Conditioning and the linear boundary value problem.

Consider the boundary value problem given by

$$y' = A(t)y + F(t) \tag{3.1a}$$

$$B_1 y(0) + B_2 y(a) = c \tag{3.1b}$$

where $A(t)$ is a continuous matrix, defined on the interval $[0, a]$, and where $F(t)$ is a continuous n-vector on the same interval. The n-vector

c is fixed. The constant $n \times n$ matrices B_1, B_2 will be assumed to be such that the rank of the $n \times 2n$ matrix (B_1, B_2) equals n. Under this assumption, it is not difficult to show the existence of an elementary matrix, P, such that $B_1P + B_2$ is nonsingular (e.g. Keller [4], p. 60). We shall further assume $B_1P + B_2$ is well-conditioned.

Let $Y(t)$ be the fundamental matrix of (3.1a) satisfying $Y(0) = I$. According to the variation of parameters formula, any solution of (3.1a) can be written in the form

$$y(t) = Y(t)y(0) + Y(t)\int_0^t Y^{-1}(s)F(s)ds ,$$

which, when substituted in (3.1b), yields the algebraic equations

$$[B_1 + B_2Y(a)]y(0) = c - B_2Y(a)\int_0^a Y^{-1}(s)F(s)ds . \tag{3.2}$$

In order to obtain a unique solution to 3.1 then, it is necessary and sufficient to require $B_1 + B_2Y(a)$ to be nonsingular. It is also clear from (3.2) that successful numerical solution of (3.1) depends on the conditioning of $B_1 + B_2Y(a)$. In the following we shall be concerned with obtaining an estimate of the conditioning of the matrix.

THEOREM 3.1. *The fundamental matrix $Y(t)$ and its inverse $Y^{-1}(t)$ satisfy the following inequalities on $[0,a]$:*

$$\|Y(t)\| \leqq \exp\int_0^t \mu[A(s)]ds \tag{3.3}$$

$$\|Y^{-1}(t)\| \leqq \exp\int_0^t \mu[-A(s)]ds . \tag{3.4}$$

PROOF. The proof of (3.4) will be given here. The proof of (3.3) can be established in the same manner or obtained from a result of Dahlquist ([5], p. 14). To show (3.4), first set $V(t) = Y^{-1}(t)$ and differentiate $V(t)Y(t) = I$ to obtain the equation

$$V'(t) = -V(t)A(t)$$

satisfied by $V(t)$ on $[0,a]$. Now let $r(t) = \|V(t)\|$. Then, as shown in [5, p. 13], the right derivative of $r(t)$ satisfies

$$r_+'(t) = \lim_{h\to+0}\left[\frac{\|V(t)+hV'(t)\|-\|V(t)\|}{h}\right]$$
$$= \lim_{h\to+0}\left[\frac{\|V(t)-hV(t)A(t)\|-\|V(t)\|}{h}\right]$$
$$\leqq \left\{\lim_{h\to+0}\left[\frac{\|I-hA(t)\|-1}{h}\right]\right\}\|V(t)\|$$

i.e.

$$r_+'(t) \leqq \mu[-A(t)]r(t), \quad r(0) = 1.$$

Using a standard theorem on differential inequalities, it follows that

$$r(t) = \|V(t)\| \leqq \exp\int_0^t \mu[-A(s)]ds\,.$$

COROLLARY 3.1. *The condition number of the fundamental matrix $Y(t)$ satisfies*

$$M[Y(t)] \leqq \exp\int_0^t [\mu[A(s)]+\mu[-A(s)]]ds\,.$$

In order to relate the estimate established by the corollary to the conditioning of $B_1+B_2Y(a)$ we require the following result (cf. Noble [7]):

THEOREM 3.2. *If $\|S\|\leqq l<1$ and if $\|I\|=1$, then $I+S$ is non-singular and*

$$\|(I+S)^{-1}\| \leqq \frac{1}{1-\|S\|} \leqq \frac{1}{1-l}.$$

THEOREM 3.3. *Let B_1+B_2 be nonsingular. The condition number of $B_1+B_2Y(t)$ satisfies*

$$M[B_1+B_2Y(t)] \leqq M[B_1+B_2]M[Y(t)]\frac{1+l}{1-l} \tag{3.5}$$

for all $t\in[0,a]$ such that

$$\int_0^t \exp\left[\int_s^t \mu[-A(p)]dp\right]\|A(s)\|ds \leqq \frac{l}{\|(B_1+B_2)^{-1}\|\,\|B_1\|}$$

for $0<l<1$.

PROOF. Note that

$$B_1+B_2Y(t) = B_1(I-Y(t))+(B_1+B_2)Y(t) = [I+S(t)][B_1+B_2]Y(t)$$

where

$$S(t) = B_1(Y^{-1}(t)-I)(B_1+B_2)^{-1}.$$

From the preceding theorem, it follows immediately that

$$M[B_1+B_2Y(t)] \leqq M[B_1+B_2]M[Y(t)]\left[\frac{1+l}{1-l}\right]$$

for all $t \in [0,a]$ such that $\|S(t)\| \leqq l < 1$. Now let

$$R(t) = Y^{-1}(t)-I$$

and

$$r(t) = \|R(t)\|.$$

In the same manner as in the proof of Theorem 3.1

$$r'_+(t) \leqq \mu[-A(t)]r(t)+\|A(t)\|, \quad r(0)=0.$$

Using the differential inequality result,

$$r(t) \leqq \int_0^t \exp\left[\int_s^t \mu[-A(p)]dp\right]\|A(s)\|ds.$$

Then if

$$\int_0^t \exp\left[\int_s^t \mu[-A(p)]dp\right]\|A(s)\|ds \leqq \frac{l}{\|(B_1+B_2)^{-1}\|\,\|B_1\|}$$

$\|S(t)\| \leqq l$, proving the theorem.

A more convenient estimate is the following:

COROLLARY. (3.5) *is satisfied for all* $t \in [0,a]$ *such that*

$$\int_0^t \|A(s)\|ds \leqq \ln\left[\frac{l}{\|(B_1+B_2)^{-1}\|\,\|B_1\|}+1\right] \tag{3.6}$$

where $0 < l < 1$.

PROOF. Since $\mu[-A] \leqq \|A(t)\|$

$$r_+'(t) \leqq \|A(t)\|r(t)+\|A(t)\|,$$

and the proof follows as in the theorem.

The preceding results, while providing convenient estimates, require the matrix B_1+B_2 to be nonsingular. The following alternative is valid for singular B_1+B_2, as well as nonsingular. In it we make use of the observation that, if $\operatorname{rank}(B_1+B_2)=n$, then there exists an elementary matrix, P, such that B_1P+B_2 is nonsingular.

THEOREM 3.4. *Assume that* $B_1+B_2Y(t)$ *is nonsingular and that*

$$\operatorname{rank}(B_1,B_2) = n, \quad \operatorname{rank}(B_1+B_2) = m \leqq n .$$

Then

$$M[B_1+B_2Y(t)] \leqq M[B_1P+B_2]M[I+S_1(t)]M[Y(t)] \tag{3.7}$$

where

$$S_1(t) = B_1[Y^{-1}(t)-P][B_1P+B_2]^{-1} .$$

PROOF. The proof follows immediately by writing

$$B_1+B_2Y(t) = (I+S_1(t))(B_1P+B_2)Y(t) .$$

REMARK 1. In the case that B_1+B_2 is singular, the following result may be useful in determining when $B_1+B_2Y(t)$ is nonsingular. Here I_k denotes the $k\times k$ identity matrix.

THEOREM 3.5. (George and Gunderson [6]). *Let the matrices* B_1,B_2 *satisfy the assumptions of Theorem* 3.4, *with* $\operatorname{rank}(B_1+B_2)=m<n$. *Assume also, with no loss of generality, that the matrix sum* B_1+B_2 *is of the form*

$$B_1+B_2 = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where B_{11} *is a nonsingular* $m\times m$ *matrix. If*

$$T_1 = \begin{bmatrix} I_m & 0 \\ -B_{21}B_{11}^{-1} & I_{n-m} \end{bmatrix}, \qquad T_2 = \begin{bmatrix} I_m & -B_{11}^{-1}B_{12} \\ 0 & I_{n-m} \end{bmatrix}$$

and the matrices T_1UT_2 *and* $T_1B_1A(0)T_2$ *have at least one non-zero element* α_{ij}, *for* $m+1\leqq i,\ j\leqq n$, *then there exists a unique solution to the boundary value problem* (3.1) *for sufficiently small interval* $[0,a]$.

REMARK 2. Theorem 3.4 suffers in comparison to theorem 3.3 mainly in that it is not possible to obtain a convenient estimate for the condition number of the matrix $I+S_1(t)$. However, since it displays the dependence of conditioning on the condition number of the fundamental matrix, $Y(t)$, it is nevertheless useful, as will be illustrated in the following section.

4. The shooting interval.

The results of the preceding section can be illustrated by considering the following two boundary value problems:

(4.1a) $$y' = Ay = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} y$$

(4.1b) $$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

(4.1c) $$B_1 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$$

Considering first the boundary value problem [(4.1a), (4.1b)], it can be seen that $B_1+B_2=I$ so that Theorem 3.3 applies. Here

$$\mu_1[-A] = \|A\|_1 = 1$$

and, choosing $l=.65$, the corollary may be used to show the estimate (3.5) is valid over the interval $[0,.5]$. From Corollary 3.1, $M[Y(.5)] \leqq 2.8$ and it follows $M[B_1+B_2Y(.5)] \leqq 13$. Thus, if the interval $[0,a]$ is of length approximately one-half, then the shooting method formulation of the problem should be well conditioned. In the next section it will be shown that, in the case the interval $[0,a]$ is of length appreciably larger than one-half, the estimate above may still be used to partition the interval and the method of parallel shooting employed.

Theorem 3.4 could actually be used to determine an interval for either of the boundary value problems and will usually lead to a less conservative value for the right end-point than Theorem 3.3, but at the expense of some additional labor. To apply the theorem, the assumption is made that the primary contribution to the condition number is from $Y(t)$. This will usually be the case if $B_1+B_2Y(t)$ is nonsingular, but after establishing an interval $[0,a]$ the condition number of $I+S_1(a)$ should be checked. For the problem [(4.1a), (4.1c)]

$$P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

$$S_1(t) = \begin{bmatrix} e^{-t} & -1 \\ 0 & 0 \end{bmatrix}$$

and, from Corollary 3.1,

$$M[Y(t)] \leqq e^{2t}$$

Placing an upper bound of 100 on $M[Y(t)]$ yields

$$e^{2t} \leqq 100 \quad \text{or} \quad t \leqq 2.3 = a\,,$$

where

$$M[I+S_1(a)] = 3.82\,.$$

Again, if $a \gg 2.3$, the same procedure can be used to establish a partition of the interval for parallel shooting, discussed below.

5. Parallel shooting.

Following Keller [4], we establish a partition $0=t_0<t_1<\ldots<t_k=a$ of the interval $[0,a]$ and define $\Delta_j=t_j-t_{j-1}$, $s=t-t_{j-1}/\Delta_j$. By this change of variables, a system of k linear equations are obtained from (3.1),

$$\frac{dy_j}{ds} = A_j(s)y_j+f_j(s), \quad j=1,2,\ldots,k, \tag{5.1a}$$

each defined on $[0,1]$. Note that the boundary condition (3.1b) becomes

$$B_1y_1(0)+B_2y_k(1) = c \tag{5.1b}$$

If the additional conditions

$$y_{j+1}(0) = y_j(1), \quad j=1,2,\ldots,k-1, \tag{5.1c}$$

are imposed, then any solution to (5.1) will define a solution to (3.1). Now let $Y_j(s)$ be the fundamental matrix for (5.1a) satisfying $Y_j(0)=I$, $j=1,2,\ldots,k$. Substituting into (5.1b) and (5.1c), these conditions can be written

$$[\bar{B}_1+\bar{B}_2\bar{Y}(1)]\bar{d} = \bar{c}-\bar{B}_2\bar{Y}(1)\int_0^1 \bar{Y}^{-1}(s)\bar{f}(s)\,ds \tag{5.2}$$

where

$$\bar{f}(s) = \begin{bmatrix} f_1(s) \\ \vdots \\ f_n(s) \end{bmatrix}$$

$$\bar{B}_1 = \begin{bmatrix} B_1 & 0 & 0 & \ldots & 0 \\ 0 & I & 0 & \ldots & 0 \\ 0 & 0 & I & \ldots & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 & \ldots & I \end{bmatrix}$$

$$\bar{B}_2 = \begin{bmatrix} 0 & 0 & 0 \ldots & & B_2 \\ -I & 0 & 0 \ldots & & 0 \\ 0 & -I & 0 \ldots & & 0 \\ \vdots & & & & \vdots \\ 0 & 0 & 0 \ldots & -I & 0 \end{bmatrix}$$

$$\bar{Y}(1) = \begin{bmatrix} Y_1(1) & & 0 \\ & \ddots & \\ & & \ddots \\ 0 & & Y_k(1) \end{bmatrix}$$

$$\bar{c} = \begin{bmatrix} c \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Elementary matrix operations show that $\bar{B}_1+\bar{B}_2$ is nonsingular if B_1+B_2 is nonsingular. Also if $\operatorname{rank}(B_1+B_2)=m<n$, and the hypotheses of Theorem 3.5 are satisfied, it can be shown that, if a is sufficiently small, (5.1) has a solution.

The convenience of extending the results of the previous sections to the parallel shooting problem comes from observing that for any of the norms $(\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty)$ the condition number

$$M[\overline{Y}(1)] = \max_k \|Y_k(1)\| \cdot \max_j \|Y_j^{-1}(1)\| .$$

Hence, by Theorem 3.1,

$$M[\overline{Y}(1)] \leqq \exp\left\{ \max_k \int_{t_{k-1}}^{t_k} \mu[A(s)]ds + \max_j \int_{t_{j-1}}^{t_j} \mu[-A(s)]ds \right\}.$$

Thus equation (3.5) becomes

$$M[\bar{B}_1+\bar{B}_2\overline{Y}(1)] \leqq M[\bar{B}_1+\bar{B}_2]M[\overline{Y}(1)]\frac{1+l}{1-l}$$

and is valid so long as

$$\int_{t_{k-1}}^{t_k} \|A(s)\|ds \leqq \ln\left[\frac{l}{\|(\bar{B}_1+\bar{B}_2)^{-1}\|\,\|\bar{B}_1\|}\right] + 1$$

for all $k=1,2,\ldots,n$. Similarly we have in analogy to (3.7)

$$M[\bar{B}_1+\bar{B}_2\overline{Y}(1)] \leqq M[\bar{I}+\bar{S}_1(1)]M[\bar{B}_1\bar{P}+\bar{B}_2]M[\overline{Y}(1)] .$$

Returning to the examples of the previous section and assuming $a \gg 2.3$, it follows that the interval $[0,a]$ could be partitioned either by intervals $\Delta_j=.5$ or $\Delta_j=2.3$ and the corresponding parallel shooting formulation of the problems should be well conditioned (a simple calculation shows $M[\bar{I}+\bar{S}_1(1)] \leqq 100$).

Another interesting example is one given by Holt [9] and also considered by Osborne [2].

$$\frac{d^2y}{dt^2} - (1+t^2)y = 0$$

$$y(0) = 1, \quad y(L) = 0 .$$

According to Holt, the solution could not be obtained by conventional shooting methods for $L > 3.5$. Osborne employed the parallel shooting method with $\Delta_j = .2$ to obtain accurate solutions over the interval $[0, 10]$. Using our results, (with $B_1 + B_2$ singular) over the interval $[0, 5]$ and with $M[\overline{Y}(1)] \leqq 150$ yields the partition tabulated in table 4.1.

Table 4.1. *Partition for Holt's Example.*

$t_1 = 2$	$t_6 = 4.25$
$t_2 = 2.75$	$t_7 = 4.50$
$t_3 = 3.25$	$t_8 = 4.70$
$t_4 = 3.65$	$t_9 = 4.90$
$t_5 = 4.0$	$t_{10} = 5.00$

The solution obtained with this partition was equally accurate and reflected an economical partition choice. This indicates Holt's example is reasonably well conditioned for small values of t and increasingly poorly conditioned (hence a finer partition) for larger values of t.

REFERENCES

1. S. M. Roberts and J. S. Shipman, *Continuation in shooting methods for two-point boundary value problems*, J. Math. Anal. Appl. 18 (1967), 45–58.
2. M. R. Osborne, *On shooting methods for boundary value problems*, J. Math. Anal. Appl. 27 (1969), 417–433.
3. P. B. Bailey and L. F. Shampine, *On shooting methods for two-point boundary value problems*, J. Math. Anal. Appl. 23 (1968), 235–249.
4. H. B. Keller, *Numerical Methods for Two-Point Boundary Value Problems*, Blaisdell, Waltham, 1968.
5. G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Trans. Roy. Inst. Tech., Stockholm, No. 130 (1959).
6. J. H. George and R. W. Gunderson, *An existence theorem for linear boundary value problems*. (Preprint).
7. B. Noble, *Applied Linear Algebra*, Prentice–Hall, Englewood Cliffs, 1969.
8. J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon, Oxford, 1965.
9. J. F. Holt, *Numerical solution of two-point boundary value problems by finite difference methods*, Comm. ACM 7 (1964), 366–373.

MATHEMATICS DEPARTMENT
UNIVERSITY OF WYOMING
LARAMIE, WYOMING 82070

AND

MATHEMATICS DEPARTMENT
UTAH STATE UNIVERSITY
LOGAN, UTAH 84321