SIAM J. SCI. COMPUT.
Vol. 14, No. 1, pp. 231–238, January 1993

013

A COLLECTION OF PROBLEMS FOR WHICH GAUSSIAN ELIMINATION WITH PARTIAL PIVOTING IS UNSTABLE*

STEPHEN J. WRIGHT†

Abstract. A significant collection of two-point boundary value problems is shown to give rise to linear systems of algebraic equations on which Gaussian elimination with row partial pivoting is unstable when standard solution techniques are used.

Key words. two-point boundary value problems, Gaussian elimination, partial pivoting

AMS(MOS) subject classifications. 65F05, 65L10, 65L20

1. Introduction.

It is well known that when Gaussian elimination with row partial pivoting is applied to a $k \times k$ real matrix, a growth factor of up to 2^{k-1} may be observed in the upper-triangular factor. A matrix which achieves this upper bound, due to Wilkinson [9, p. 212], has passed into the folklore of numerical linear algebra. However, matrices that exhibit growth factors which are exponential in their dimension have apparently not previously been observed in connection with any "practical applications." In the next section, we discuss two-point boundary value problems for which standard solution techniques give rise to matrices with this undesirable property. An extreme case is derived in §3. Some random trials, reported in §4, suggest that the collection of such problems may represent a significant fraction of the set of two-point boundary value problems with coupled end conditions.

2. Examples arising from two-point boundary value problems.

Consider the general two-point boundary value problem

$$y' = M(t)y + q(t), \qquad B_a y(a) + B_b y(b) = \beta, \qquad y(t) \in \mathbf{R}^n, \tag{1}$$

and the particular problem defined by

$$n = 2, \qquad M(t) \equiv \begin{bmatrix} -\frac{1}{6} & 1 \\ 1 & -\frac{1}{6} \end{bmatrix}, \qquad a = 0,\ b = 60, \qquad B_a = B_b = I. \tag{2}$$

(The values of $q(t)$ and β are not relevant to our present discussion.) The problem defined by the data (2) is well conditioned; that is, its solution y is insensitive to perturbations in the data $M(t)$, $q(t)$, B_a, B_b, and β which define it. To show this, we can construct the analytic solution as follows: First, define $Y(t) \in \mathbf{R}^{n\times n}$ as a fundamental solution of the homogeneous ordinary differential equation $y' = M(t)y$. Defining

$$Q = B_a Y(a) + B_b Y(b)$$

and

$$G(x,t) = \begin{cases} Y(x)Q^{-1}B_a Y(a) Y^{-1}(t) & t \le x, \\ -Y(x)Q^{-1}B_b Y(b) Y^{-1}(t) & t > x, \end{cases}$$

*Received by the editors October 2, 1991; accepted for publication (in revised form) February 6, 1992. This research was supported by the Applied Mathematical Sciences subprogram of the Office of Energy Research, U. S. Department of Energy, under contract W-31-109-Eng-38, and by a Raybould Fellowship from the University of Queensland, Australia.

†Argonne National Laboratory, 9700 South Cass Avenue, Argonne, Illinois 60439.

we can write

$$y(x) = Y(x)Q^{-1}\beta + \int_a^b G(x,t)q(t)\,dt \tag{3}$$

(see, for example, [1]). Well-conditioning is usually quantified in terms of norms of the two operators in (3). For the constant-coefficient problem (2) (with $M \equiv M(t)$) we have, choosing $Y(0) = I$, that

$$\begin{aligned} Y(t) = e^{Mt}, \qquad Q = I + e^{60M}&, \\ \kappa_1 \triangleq \sup\nolimits_{a\le x\le b} \|Y(x)Q^{-1}\|_\infty \approx 1&, \\ \kappa_2 \triangleq (b-a)\ \sup\nolimits_{a\le t,x\le b} \|G(x,t)\|_\infty \approx (b-a)&, \end{aligned} \tag{4}$$

and so our claim of well-conditioning is verified.

A standard algorithm for problems of the form (1) is multiple shooting [6]. In its simplest form, this algorithm proceeds by partitioning $[a,b]$ into N subintervals, that is, defining a mesh $a = t_1 < t_2 < \cdots < t_{N+1} = b$ by

$$t_i = a + ih, \qquad i = 1,\dots,N+1, \qquad h = (b-a)/N.$$

On each subinterval $[t_i, t_{i+1}]$, the following initial value problems are solved:

$$Y_i' = M(t)Y_i, \qquad Y_i(t_i) = I, \tag{5}$$

$$y_{pi}' = M(t)y_{pi} + q(t), \qquad y_{pi}(t_i) = 0. \tag{6}$$

Defining s_i as the value of the true solution y at t_i, we note that

$$y(t) = Y_i(t)s_i + y_{pi}(t), \qquad t \in [t_i, t_{i+1}], \qquad i = 1,\dots,N.$$

Since, clearly, $y(t)$ must be continuous across the mesh points, we have

$$Y_i(t_{i+1})s_i + y_{pi}(t_{i+1}) = s_{i+1}, \qquad i = 1,\dots,N. \tag{7}$$

Moreover, from the boundary conditions,

$$B_a s_1 + B_b s_{N+1} = \beta. \tag{8}$$

The equations (7) and (8) yield a system of linear equations whose solution is $s_1,\dots,$ s_{N+1}. The coefficient matrix has the general form

$$A = \begin{bmatrix} B_a & & & & B_b \\ -Y_1(t_2) & I & & & 0 \\ & -Y_2(t_3) & I & & \vdots \\ & & \ddots & \ddots & 0 \\ & & & -Y_N(t_{N+1}) & I \end{bmatrix}. \tag{9}$$

For the data (2) we have in particular that $Y_i(t_{i+1}) = e^{Mh}$, and so (9) becomes

$$\bar{A} = \begin{bmatrix} I & & & & I \\ -e^{Mh} & I & & & 0 \\ & -e^{Mh} & I & & \vdots \\ & & \ddots & \ddots & 0 \\ & & & -e^{Mh} & I \end{bmatrix}. \tag{10}$$

The conditioning of the "shooting matrix" A can be related to the conditioning of the original problem (1) by a theorem of Osborne and Mattheij, which appears as Theorem 4.11 in Lentini, Osborne, and Russell [3].

THEOREM 2.1. *Suppose that* $\|[B_a|B_b]\|_\infty \leq 1$ *and that* N *is chosen large enough that*

$$\|Y_i(t_{i+1})\|_\infty \leq K', \qquad i = 1, \ldots, N.$$

Then

$$\operatorname{cond}_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty \leq (K' + 1)(\kappa_1 + \kappa_2 N/(b-a)),$$

where κ_1 *and* κ_2 *are as defined in* (4).

For (10), this bound translates to

$$\operatorname{cond}_\infty(\bar{A}) \leq (\|e^{Mh}\|_\infty + 1)(1 + N)$$

for h sufficiently small. (For (2) with $N = 200$, the bound is about 459.) Suppose that h is chosen small enough that all elements of e^{Mh} are less than 1 in magnitude. This is certainly possible for (2), since

$$(11) \qquad e^{Mh} = I + Mh + O(h^2) \approx \begin{bmatrix} 1 - h/6 & h \\ h & 1 - h/6 \end{bmatrix}.$$

If Gaussian elimination with row partial pivoting is applied to the matrix $\bar{A}$, *no pivoting occurs*, and the following factorization is obtained:

$$(12) \quad \bar{A} = \begin{bmatrix} I & & & & \\ -e^{Mh} & I & & & \\ & -e^{Mh} & I & & \\ & & \ddots & \ddots & \\ & & & -e^{Mh} & \hat{L} \end{bmatrix} \begin{bmatrix} I & & & & & I \\ & I & & & & e^{Mh} \\ & & I & & & e^{2Mh} \\ & & & \ddots & & \vdots \\ & & & & I & e^{M(60-h)} \\ & & & & & \hat{U} \end{bmatrix},$$

where $\hat{L}\hat{U} = (I + e^{60M})$ with $\hat{L}$ and $\hat{U}$ lower and upper triangular, respectively. Clearly, exponential element growth has taken place in the last column of U. When $N = 200$ ($h = 0.3$), the largest element in the U factor is approximately 2.59×10^{21}.

Poor performance of Gaussian elimination is not limited to matrices derived from the multiple-shooting algorithm (5)–(8). Similar behavior also occurs when "backwards shooting" (from the right-hand endpoint of each subinterval) is used and when a midpoint-rule finite difference discretization is applied to (1), (2).

As expected, a complete pivoting strategy produces a stable factorization. For the matrix $\bar{A}$, the largest element in the computed U factor is just 1.284, and both L and U factors are sparse (density 1.41 percent for L, 1.18 percent for U), though the fill-in pattern is somewhat irregular.

The behavior exhibited in (12) will occur whenever the coefficient matrix M has negative diagonal elements, since then, provided h is sufficiently small, e^{Mh} will have all its elements less than one in magnitude and no pivoting will occur. Terms of order $e^{M(b-a)}$ will therefore appear in the U factor, and these will be large if any of the eigenvalues of M have positive real parts.

Mattheij [5] has observed that stability of the partial pivoting strategy is closely related to the following feature of the pivoting pattern: if the matrix A is regarded as a collection of $N+1$ "row blocks," each containing n rows, then *the number of rows that are pivoted between row block i and row block $i+1$ ($i = 1, \ldots, N-1$) should equal the number of eigenvalues of M whose real parts are positive.* An assumption that the pivoting pattern has this property is crucial to the analysis of Wright [11]. Although this property appears to hold for most of the standard two-point boundary test problems in the literature, the matrices M just discussed lead to shooting matrices for which it is not satisfied.

The trouble is not confined to matrices M with negative diagonal entries for which no pivoting occurs during the factorization of the shooting matrix. The coefficient matrix

$$M = \begin{bmatrix} -10 & -19 \\ 19 & 30 \end{bmatrix}$$

has two positive eigenvalues, at approximately 3.755 and 16.245. Suppose we take the remaining data as in (2), construct the shooting matrix as in (5)–(8), and apply row partial pivoting. Blowup is again observed for sufficiently small h; when the shooting matrix is factorized, only one row is pivoted between each successive pair of block rows.

Neither is the blowup behavior restricted to constant-coefficient problems (i.e., those for which $M = M(t)$ is constant on $[a, b]$). Examples similar to those in this note, but with nonconstant coefficients or nonlinear dynamics, can easily be constructed by modifying the examples above. Such examples are actually more "relevant" since, in practice, they are often solved by multiple-shooting and finite difference algorithms and therefore run the risk of exhibiting the instability just described, whereas constant-coefficient problems are usually solved by other means.

In order to produce the unstable behavior, it is necessary for the problem to have coupled end conditions. If instead the boundary conditions have the separated form

$$\bar{B}_0 x(a) = \beta_a, \qquad \bar{B}_1 x(b) = \beta_b, \qquad \bar{B}_0 \in \mathbf{R}^{n_a \times n}, \bar{B}_1 \in \mathbf{R}^{n_b \times n}, \qquad n_a + n_b = n,$$

the matrix (9) is a permutation of a banded matrix that has a bandwidth of $2n$. Since element growth in row partial pivoting is at worst exponential in the bandwidth, the type of growth depicted above cannot occur.

Alternative stable and efficient means for producing factorizations of the matrices (9) are available. Wright [10] has described a structured QR factorization scheme for these matrices and proved its stability. (A later modification of this scheme, based on Givens rotations rather than Householder transformations, is stable while being not much less efficient than the LU algorithm discussed above.) If speed is an important consideration, the LU factorization can be attempted in the first instance and backup strategies (such as QR factorization of A or LU factorization of A^T) can be called on only if instability is detected. LU factorization of A^T will work for the problems discussed above, though it can fail on other problems, for example, the problem

$$n = 2, \qquad M(t) \equiv \begin{bmatrix} \frac{1}{6} & -1 \\ -1 & \frac{1}{6} \end{bmatrix}, \qquad a = 0,\ b = 60, \qquad B_a = I, \qquad B_b = 2I.$$

Whether there exist problems for which LU factorization of both A and A^T is unstable is an open question.

Liu and Russell [4] have apparently observed the effects of lack of stability of LU factorization on a practical problem. They use a continuation code for parametrized ordinary differential equations (ODEs) to solve the Kuramoto–Sivashinsky equation and try various factorization techniques to perform the core operation of solving the linear equations that arise repeatedly during the computation. (The coefficient matrices of these linear systems are actually bordered versions of the shooting matrices above.) They find that the continuation algorithm is less robust when an LU algorithm with partial pivoting is used to perform the factorization than when a QR algorithm is used.

3. Fraction of maximum possible growth.

When $n = 1$, it is not possible to construct an example that leads to exponential blowup. A worst case appears to be the scalar problem

$$y' = q(t), \qquad y(a) + y(b) = \beta, \qquad y(t) \in \mathbf{R},$$

which, when algorithm (5)–(8) is applied, produces a matrix for which element growth of order N occurs in the U factor.

For $n = 2$, we can investigate how closely the upper bound on element growth for matrices of size $(N+1)n$, namely, $2^{(N+1)n-1}$, is approached when the matrix has the form (9). Consider a constant-coefficient problem defined by the data

$$(13) \quad n = 2, \qquad M = \lambda \begin{bmatrix} -\alpha & 1 \\ 1 & -\alpha \end{bmatrix}, \qquad a = 0, \;\; b = L, \qquad B_a = B_b = I,$$

where $\alpha \in (0,1)$. The matrix M has eigenvalues $\lambda(-\alpha \pm 1)$. By choosing N and setting $h = L/N$, we obtain a coefficient matrix similar to $\bar{A}$ in (10). To ensure that no pivoting occurs during the factorization, we must choose N large enough that no elements of e^{Mh} exceed 1. Explicit calculation of the matrix exponential shows that this requirement is equivalent to

$$e^{-\alpha\lambda h}[e^{\lambda h} + e^{-\lambda h}] \leq 2,$$

or, if we define $X = e^{\lambda h}$,

$$p(X) = X^2 - 2X^{1+\alpha} + 1 \leq 0.$$

Since, for $\alpha \in (0,1)$, $p(0) = 1$, $p(1) = 0$, $p'(1) < 0$, and $\lim_{X\to+\infty} p(X) = +\infty$, the equation

$$p(X) = 0$$

has a solution $X(\alpha)$ that is strictly greater than 1. Since $h = L/N > 0$, this is the solution of interest to us.

If we assume that no pivoting occurs during the Gaussian elimination, it follows from (12) that the largest element in the U factor is approximately equal to the largest element in e^{ML}. A little calculation shows that this is approximately

$$E = \tfrac{1}{2} e^{\lambda L(1-\alpha)}.$$

Given the upper bound on the growth for matrices of dimension $2(L/h + 1)$, namely,

$$E_{\max} = 2^{2(L/h+1)-1},$$

we have

$$\rho(\alpha) \stackrel{\triangle}{=} \frac{\log_2 E}{\log_2 E_{\max}} = \frac{\dfrac{\lambda L(1-\alpha)}{\ln 2} - 1}{2(\frac{L}{h}+1)-1}$$

$$\approx \frac{1}{\ln 2}\frac{\lambda h(1-\alpha)}{2}$$

$$= \frac{1}{2\ln 2}(1-\alpha)\ln X(\alpha).$$

Simple analysis of $p(X)$ shows that $X(\alpha) \to +\infty$ and $(1-\alpha)\ln X(\alpha) \to \ln 2$ as $\alpha \to 1^-$. Therefore,

$$\lim_{\alpha\to 1^-} \rho(\alpha) = \frac{1}{2}.$$

In fact, we can show that $\rho(\alpha)$ is monotonic increasing on the interval $(0,1)$, with $\lim_{\alpha\to 0^+} \rho(\alpha) = 0$.

We conclude that for multiple-shooting coefficient matrices arising from (13), the observed growth factor during Gaussian elimination may be as large as approximately 2^{N+1}.

4. Discussion. In [7], Trefethen writes with reference to real $k \times k$ matrices: "Perhaps large growth rates like 2^{k-1} correspond to unstable 'modes' that are themselves somehow unstable, in the sense that computations tend to drift away from them towards stabler configurations." The "drifting away" is precisely what *fails* to happen for the matrices described above. Rather, the unstable modes are propagated and reinforced because of the presence of a recurrence in which the relationship between any two successive terms is the same. The highly specialized structure of the matrices (9) and (10) accounts for the difference between our experience with randomly generated examples (reported below) and the experience of Trefethen and Schreiber [8], who worked with randomly generated *dense* matrices and observed an average growth rate of $k^{2/3}$ for real $k \times k$ matrices when partial pivoting was used.

Higham and Higham [2] present a number of dense matrices that arise naturally in applications for which the growth factors are between $n/2$ and n when both partial and complete pivoting are used. They point out that large growth factors do not necessarily imply large backward errors in the computed solution. However, for the problems described above, we would expect both the forward and backward errors in the computed solution to be large. Consider again the example (1), (2). Because $(1+e^{60M})$ is singular in double precision arithmetic, the computed U factor has a zero element in its bottom right-hand corner. Suppose we choose $q(t)$ and β in (1) such that the true solution has $y(t) = s_i = (1,1)^T$ for all t and i, and suppose that we avoid the singularity in U by setting $\hat{s}_{N+1} = s_{N+1}$, where the hat denotes a computed quantity. If we back-substitute for the remaining $\hat{s}_i$, we obtain a relative error of 2.88×10^5 in the computed solution of (7), (8). The backward error, which is defined as

$$\frac{\|c - A\hat{x}\|_2}{\|A\|_F\|\hat{x}\|_2},$$

TABLE 1

Number of "blowups" during each of five trials, each trial consisting of 100 *randomly generated problems.*

		$\mu \in (10^{10}, \infty)$ / $\mu \in (10^3, 10^{10}]$		
Trial	γ	$n = 2$	$n = 4$	$n = 6$
A	1	0/0	0/0	0/0
B	10	2/2	2/4	4/2
C	100	3/3	4/5	6/1
D	.1	0/0	0/0	0/0
E	2 ($B_a = B_b = I$)	2/2	5/5	10/6

where $\hat{x}$ is the computed solution of the system $Ax = c$, is found to be 0.0286. All computations here were performed in double precision.

An interesting open question is: Among all constant-coefficient problems of the form (1), is the set of triplets (M, B_a, B_b) that give rise to matrices on which Gaussian elimination fails truly a "nontrivial" subset of

$$\{(M, B_a, B_b) \mid M, B_a, B_b \in \mathbf{R}^{n \times n}, \quad \|[B_a|B_b]\|_\infty \leq 1\}?$$

Computational experiments with randomly generated constant-coefficient problems shed a little light on this question. We generated $n \times n$ matrices M $(n = 2, 4, 6)$ by choosing each matrix element from a uniform distribution on $[-10, 10]$. In the first four sets of trials, the elements of B_a and B_b were chosen likewise and then scaled so that $\|[B_a|B_b]\|_\infty = \gamma$. We chose $\gamma = 1$, 10, 100, and 0.1 in trials A, B, C, and D, respectively. In trial E, we set $B_a = B_b = I$. In each trial, 100 randomly generated problems were solved by using (5)–(8) with 1024 subintervals, with $a = 0$ and $b = 10$. For each trial we define the following quantity μ that measures the "blowup" in the U factor

$$\mu \triangleq \max(U) \;/\; \|[B_a|B_b]\|_\infty,$$

where max(U) is the magnitude of the largest element in the upper-triangular factor. Each entry in Table 1 has two components. The first is the number of problems (out of 100) for which μ exceeded 10^{10} and the second is the number of problems for which $\mu \in (10^3, 10^{10}]$. When $\mu > 10^{10}$, the resulting loss of precision is usually large enough to destroy the accuracy of the computed solution.

Three features of Table 1 are worth noting. First, the likelihood of blowup seems to increase as n increases. This is not a surprise—as n grows we would expect a mismatch between the number of positive eigenvalues and the number of rows that are pivoted between successive blocks (see §2) to become more likely. Second, the choice $B_a = I$ (which can be obtained by a simple transformation whenever the B_a in (1) is nonsingular) seems to be particularly inappropriate, though it is the "natural" choice in many circumstances. The reason for this should be clear from the discussion near the end of §2. Third, a wise strategy appears to be to scale B_a, B_b, and β by a moderately small constant. From Theorem 2.1 and the definitions of Q and κ_1, we see that this may cause some slight deterioration in the conditioning of the matrix (9), but trials A and D indicate that exponential blowup is much less likely to occur. A "too small" choice of γ may lead to Assumption 1 in [11] being violated during factorization of the first few row blocks, but after this initial phase, the pivoting pattern exhibits the "stable" pivoting feature described in §2.

Acknowledgments. The comments of Bob Russell, Nick Higham, and anonymous referees are gratefully acknowledged.

REFERENCES

[1] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[2] N. J. Higham and D. J. Higham, *Large growth factors in Gaussian elimination with pivoting*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 155–164.

[3] M. Lentini, M. R. Osborne, and R. D. Russell, *The close relationships between methods for solving two-point boundary value problems*, SIAM J. Numer. Anal., 22 (1985), pp. 280–309.

[4] L. Liu and R. D. Russell, *Linear system solvers for boundary value* ODEs, J. Comput. Appl. Math., to appear.

[5] R. M. M. Mattheij, *Decoupling and stability of algorithms for boundary value problems*, SIAM Rev., 27 (1985), pp. 1–44.

[6] M. R. Osborne, *On shooting methods for boundary value problems*, J. Math. Anal. Appl., 27 (1969), pp. 417–433.

[7] L. N. Trefethen, *Three mysteries of Gaussian elimination*, ACM SIGNUM Newsletter, 20 (1985), pp. 2–5.

[8] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination*, SIAM J. Matrix Anal. Appl., 11 (1990), pp. 335–360.

[9] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

[10] S. J. Wright, *Stable parallel algorithms for two-point boundary value problems*, SIAM J. Sci. Comput., 13 (1992), pp. 742–764.

[11] ———, *Stable parallel elimination for boundary value* ODEs, Tech. Rep. MCS–P229–0491, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, April 1991.