# Gaussian Elimination with Partial Pivoting Can Fail in Practice

Leslie V. Foster
Department of Mathematics and Computer Science
San Jose State University
San Jose, California 95192
foster@sjsumcs.sjsu.edu

(Preprint of pp. 1354-1362, Vol. 15, SIAM J. Matrix Anal. Appl., 1994.)

**Abstract**

Even though Gaussian elimination with partial pivoting is very widely used, $n \times n$ matrices can be constructed where the error growth in the algorithm is proportional to $2^{n-1}$. Thus for moderate or large $n$, in theory, there is a potential for disastrous error growth. However, prior to 1993 no reports of such an example in a practical application had appeared in the literature. Examples are presented that arise naturally from integral and differential equations and that lead to disastrous error growth in Gaussian elimination with partial pivoting.

**Key words.** Gaussian elimination, numerical stability, integral equations

**AMS(MOS) subject classification.** 65F05, 65R20, 65G05

## 1 Introduction

Gaussian elimination with partial pivoting (GEPP) is one of the most widely used algorithms in scientific computing. When applied to an $n \times n$ matrix $A$ it results in a factorization $PA = LU$, where $P$ is a permutation matrix, $L$ is lower triangular, and $U$ is upper triangular. Let $\hat{\boldsymbol{x}}$ represent the solution to $A\boldsymbol{x} = \boldsymbol{b}$ computed in floating point arithmetic on a computer with relative machine precision $\epsilon$. Then it is known [Wil] that

$$\frac{\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} \leq 4n^2 \mathrm{cond}_\infty(A)\rho\,\epsilon,$$

where $\boldsymbol{x}$ is the exact solution, $\mathrm{cond}_\infty(A)$ is the condition number of $A$ in the supremum norm and $\rho$ is the growth factor,

$$\rho = \frac{\max_{i,j,k} |a_{i,j}^{(k)}|}{\max_{i,j} |a_{i,j}|} \tag{1.1}$$

with $a_{i,j}^{(k)}$ denoting the $i,j$ element after the $k$th step of elimination. Thus GEPP is considered numerically stable unless $\rho$ is large.

The theory for GEPP suggests that $\rho$ can be very large. The sharpest bound is $\rho \leq 2^{n-1}$

and this is attained, for example, for matrices $A_n$ of the form [HH], [Wil], [GVL]

$$A_5 = \mathrm{diag}(\pm 1) \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 & 1 \\ -1 & -1 & 1 & 0 & 1 \\ -1 & -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} & & & 0 \\ & T & & \vdots \\ & & & 0 \\ 0 & \cdots & 0 & \theta \end{pmatrix},$$

where $T$ is an $n-1 \times n-1$ nonsingular upper triangular matrix and $\theta = \max |a_{ij}|$. Thus for moderate or large $n$ the growth factor can be large. However, more than 25 years ago Wilkinson reported:

"It is our experience that any substantial increase in size of elements of successive $A_n$ is extremely uncommon even with partial pivoting. ... No example which has arisen naturally has in my experience given an increase by a factor as large as 16."

Since Wilkinson made his remarks, Dongarra et al. [DBMS] report an example where $\rho$ is 23 and Higham and Higham [HH] report several natural, noncontrived examples where the growth factor is between $n/2$ and $n$. Although the growth factors reported in these papers are larger than those mentioned by Wilkinson, they are much less than the theoretical limit of $2^{n-1}$. For random matrices Trefethen and Schreiber [TS] show that the average growth factor does not grow exponentially. In this paper we present a class of practical examples where the growth factors do grow exponentially. Recently Wright [Wri] also presented such a class. Wright's paper and ours are different in that we consider Volterra integral equations, which are not discussed by Wright, and the growth factors for our matrices can be closer to the theoretical limit than the growth factors for Wright's matrices. Also the matrices in our examples are dense whereas Wright's are sparse. The papers are related in that results in both papers apply to boundary value problems.

In the next section we show that when the quadrature method [Bak], [DM], [Linz] is used to numerically solve certain Volterra integral equations, large growth factors can result. In §3 we illustrate the theory of §2 with a population growth model and with a two-point boundary value problem. In the last section we present a brief discussion of the implications of such examples.

Software, in the form of Matlab m files for constructing the above examples is available via the gopher system. Type "gopher sundance.sjsu.edu" on a computer with a gopher client and follow the menus.

# 2 A class of Volterra integral equations that lead to large growth factors

A linear Volterra integral equation of the second kind is of the form:

$$x(s) - \int_0^s k(s,t)x(t)\,dt = G(s). \tag{2.1}$$

Such equations show up in a wide variety of applications [Bur], [Jer], [Linz]. We consider for known $k(s,t)$, $\beta(s)$, and $G(s)$ the following modification of (2.1):

$$x(s) - \int_0^s k(s,t)x(t)\,dt + \beta(s)x(L) = G(s). \tag{2.2}$$

In general it is not possible to find an exact solution to (2.1) or (2.2). However, a variety of approximation techniques [Bak], [DM], [Linz] can be used, including the commonly used quadrature method. This is the method that we use, "starting" our procedure with the block quadrature method [DM]. We use Newton–Cotes quadrature formulas of order $p \geq 1$. To be specific for any $n$ we divide $0 \leq s \leq L$ into $n-1$ equal subintervals of length $h = L/(n-1)$.

For $i = 1, \ldots, n$, $j = 1, \ldots, n$, let $s_i = (i-1)h$, $t_j = (j-1)h$, $\beta_i = \beta(s_i)$, $b_i = G(s_i)$, $k_{ij} = k(s_i, t_j)$, and let $x_i$ be the numerical approximation to $x(s_i)$. Let $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_n)^T$, where the superscript $T$ indicates transpose. We approximate $\int_0^{s_i} k(s,t)x(t)\,dt$. Our approximation depends on $i$. For $1 \le i \le p+1$ we integrate an interpolating polynomial of degree $p$ through $(t_j, k_{ij}\, x_j), j = 1, \ldots, p+1$. For $i \ge p+1$ we use composite integration. For example if $i = kp + l$ we use standard $p$th order closed Newton–Cotes composite integration for the integral from 0 to $s_{kp+1}$. For the integral from $s_{kp+1}$ to $s_{kp+l}$, we integrate an interpolating polynomial of degree $p+1$ through $(t_j, k_{ij}\, x_j), j = kp + l - p - 1, \ldots, kp + l$.

With these approximations the vector $\boldsymbol{x}$ satisfies

$$A\boldsymbol{x} = \boldsymbol{b}. \tag{2.3}$$

For example, for $p = 2$ and $n = 7$ the matrix $A$ is

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & \beta_1 \\
-\frac{5hk_{21}}{12} & 1 - \frac{2hk_{22}}{3} & \frac{hk_{23}}{12} & 0 & 0 & 0 & \beta_2 \\
-\frac{hk_{31}}{3} & -\frac{4hk_{32}}{3} & 1 - \frac{hk_{33}}{3} & 0 & 0 & 0 & \beta_3 \\
-\frac{3hk_{41}}{8} & -\frac{9hk_{42}}{8} & -\frac{9hk_{43}}{8} & 1 - \frac{3hk_{44}}{8} & 0 & 0 & \beta_4 \\
-\frac{hk_{51}}{3} & -\frac{4hk_{52}}{3} & -\frac{2hk_{53}}{3} & -\frac{4hk_{54}}{3} & 1 - \frac{hk_{55}}{3} & 0 & \beta_5 \\
-\frac{hk_{61}}{3} & -\frac{4hk_{62}}{3} & -\frac{17hk_{63}}{24} & -\frac{9hk_{64}}{8} & -\frac{9hk_{65}}{8} & 1 - \frac{3hk_{66}}{8} & \beta_6 \\
-\frac{hk_{71}}{3} & -\frac{4hk_{72}}{3} & -\frac{2hk_{73}}{3} & -\frac{4hk_{74}}{3} & -\frac{2hk_{75}}{3} & -\frac{4hk_{76}}{3} & 1 - \frac{hk_{77}}{3} + \beta_7
\end{pmatrix}$$

In this case rows three, five, and seven are produced by using Simpson's rule for integration. Rows four and six have an odd number of intervals of integration and involve the usual Simpson rule and Simpson 3/8 rule [Linz, p. 99]. Row two arises from the block quadrature method [DM].

This approach has the attractive feature that involves higher order approximations for $p > 1$ and that, except for the last column, the matrix $A$ is block lower triangular and so, in principle, (2.3) can be solved in $O(n^2)$ not $O(n^3)$ operations.

For large $n$ we have the following results.

**Theorem 2.4** Assume that $k(s,t)$ is bounded for $0 \le s \le L$, $0 \le t \le L$. For any fixed order of integration $p \ge 1$ and for sufficiently large $n$ no row interchanges are required when GEPP is applied to the matrix $A$ in (2.3).

**Proof** Let $w_{ij}$ be the weights in the numerical integration formula corresponding to the $i, j$ element of $A$ and choose $n$ such that $\delta \equiv \max_{1 \le i,\, j \le n} |w_{ij} k_{ij} h| \le 1/(p+1)$. The first $n-1$ columns of $A$ are lower triangular except for a $p \times p$ diagonal block in columns 2 through $p+1$. Outside these columns no row interchanges are required in GEPP since $\delta \le 1/(p+1) \le \frac{1}{2}$. Let $\hat{A}$ consist of columns 2 to $p+1$ of $A$ and $\hat{I}$ be an $n \times p$ matrix with ones on the diagonal and zeros elsewhere. Then $\hat{A} = \hat{I} - B$, where $\max_{1 \le i \le n,\, 1 \le j \le p} |b_{ij}| \le \delta$. Let $\hat{A}^k$ be $\hat{A}$ after $k$ steps of GEPP, let $B^k = \hat{I} - \hat{A}^k$, and let $x_k = \max_{1 \le i \le n,\, 1 \le j \le p} |b_{ij}^k|$. Clearly, no row interchanges are required at the first step of GEPP applied to $\hat{A}$.

We now use induction. For some $k$, $1 \le k \le p-1$ assume no row interchanges are required through step $k$ of GEPP and that $x_k \le \delta/(1 - k\delta)$. Then for $i = k+1, \ldots, n$, $|\hat{a}_{ik}^k| \le x_k \le [1/(p+1)]/[(1 - k/(p+1)] \le \frac{1}{2}$ and $|\hat{a}_{kk}^k| \ge 1 - x_k \ge \frac{1}{2}$. Therefore no pivoting will be required at step $k+1$. Also since $\hat{a}_{ij}^{k+1} = \hat{a}_{ij}^k - \hat{a}_{kj}^k \hat{a}_{ik}^k / \hat{a}_{kk}^k$ then for $i \ne j$, $|\hat{a}_{ij}^{k+1}| \le x_k + x_k x_k/(1 - x_k) = x_k/(1 - x_k)$. For $i = k+1, \ldots, p$ we then have $|\hat{a}_{ii}^{k+1} - 1| \le x_k + x_k x_k/(1 - x_k) = x_k/(1 - x_k)$. Consequently $x_{k+1} \le x_k/(1 - x_k)$. This and $x_k \le \delta/(1 - \delta k)$ imply that $x_{k+1} \le \delta/[1 - (k+1)\delta]$. This completes the proof and also shows that $\max_{1 \le i \le n,\, 1 \le j \le p} |b_{ij}^p| \le \delta/(1 - p\delta)$. □

**Theorem 2.5** Assume that $k(s,t)$ is continuous over $0 \le s \le L$, $0 \le t \le L$, and $\beta(s)$ is continuous over $0 \le s \le L$. Let $\gamma(s)$ be the solution to the integral equation

$$\gamma(s) - \int_0^s k(s,t)\gamma(t)\,dt = \beta(s). \tag{2.6}$$

For any fixed $p$ the growth factor $\rho$ for GEPP applied to (2.3) satisfies

$$\lim_{n\to\infty} \rho = \frac{\max_{0\le\tau\le s\le L}\{1,\ |\beta(s) + \int_0^\tau k(s,t)\gamma(t)\,dt|,\ |1+\beta(L) + \int_0^\tau k(s,t)\gamma(t)\,dt|\}}{\max_{0\le s\le L}\{1,\ |\beta(s)|,\ |1+\beta(L)|\}}. \tag{2.7}$$

**Proof** Select a fixed $\tau$, $0 \le \tau \le L$ and suppose that $k$ is an integer, $p+1 \le k \le n-1$ such that $hk = \tau$. Assume that $h$ is sufficiently small so that no pivoting is required in GEPP applied to $A$ in (2.3). Then after step $k$ of GEPP, we have

$$\begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & I & 0 \\ L_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} U_{11} & 0 & u_1 \\ 0 & A_{22} & u_2 \\ 0 & A_{32} & u_n \end{pmatrix} = A = \begin{pmatrix} A_{11} & 0 & 0 \\ A_{21} & A_{22} & 0 \\ A_{31} & A_{32} & \tilde{a}_{nn} \end{pmatrix} + e_n^T \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \beta_n \end{pmatrix}. \tag{2.8}$$

In (2.8) there are $k$, $n-k-1$, and 1 elements, respectively, in the first, second, and third block rows and columns. Here $e_n = (0,0,\ldots,0,1)^T$.

Now let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T, \beta_n)^T$, $u = (u_1^T, u_2^T, u_n)^T$ and $v_1 \in R^k$ satisfy $U_{11}v_1 = u_1$. Then from (2.8) $L_{11}u_1 = \boldsymbol{\beta}_1$ and $L_{11}U_{11} = A_{11}$ so that $A_{11}v_1 = \boldsymbol{\beta}_1$. This last equation is the equation produced when the quadrature method is applied to (2.6) over the interval $0 \le s \le \tau = kh$. By standard results [DM, pp. 126–127], for $s_i = ih$ fixed and $i \le k$

$$v_i - \gamma(s_i) \to 0 \quad \text{as } h \to 0. \tag{2.9}$$

In the proof of Theorem 2.4 it was shown that $U_{11} = I + \tilde{B}$, where $\tilde{b}_{ij} = 0$, for $j > i > p+1$, and where $\max_{1\le i,j\le n} |\tilde{b}_{ij}| \le \delta/(1-p\delta)$, with $\delta \to 0$ as $h \to 0$. It then follows from (2.9) that for $ih$ fixed and $i \le k$

$$u_i - \gamma(s_i) \to 0 \quad \text{as } h \to 0. \tag{2.10}$$

From the second block row in (2.8), we have $L_{21}u_1 + u_2 = \boldsymbol{\beta}_2$ and $L_{21}U_{11} = A_{21}$. Therefore $u_2 = -A_{21}v_1 + \boldsymbol{\beta}_2$. Since elements in $-A_{21}$ come from discrete approximations to integrals and from (2.9) it follows that for $i$ such that $k+1 \le i \le n-1$ and for $s_i$ fixed then

$$u_i \to \int_0^\tau k(s,t)\gamma(t)\,dt + \beta(s_i) \text{ as } h \to 0. \tag{2.11}$$

From the third block row in (2.8), we also have $L_{31}u_1 + u_n = \tilde{a}_{nn} + \beta_n$ and $L_{31}U_{11} = A_{31}$ so that $u_n = \tilde{a}_{nn} - A_{31}v_1 + \beta_n$. Since $\tilde{a}_{nn} \to 1$ as $h \to 0$, we see that

$$u_n \to 1 + \beta(L) + \int_0^\tau k(s,t)\gamma(t)\,dt. \tag{2.12}$$

In proving (2.10), (2.11), and (2.12) we have assumed that $k \ge p+1$. These equations are also true for $k \le p$. For example, for $i \le k \le p$ as $h \to 0$ it follows easily that $u_i \to \beta(0)$ and $\gamma(s_i) \to \beta(0)$ so that (2.10) is true. The theorem follows from (2.10)–(2.12) and the definition of growth factor. □

**Corollary 2.13** With the assumptions of Theorem 2.5

$$\lim_{n\to\infty} \rho \ge \frac{\max_{0\le s\le L}\{1,\ |\gamma(s)|,\ |1+\gamma(L)|\}}{\max_{0\le s\le L}\{1,\ |\beta(s)|,\ |1+\beta(L)|\}}.$$

**Proof** The result follows by letting $\tau = s$ in (2.7) and using (2.6). □

Corollary 2.13 shows that for any of the numerical integration schemes that we have outlined, large growth occurs for sufficiently large $n$ if the solution $\gamma(s)$ to (2.6) is large relative to the coefficient $\beta(s)$ in (2.2). The next section shows that large growth can happen for practical problems where $A$ is well conditioned.

# 3 Examples

Our first example comes from a simple model for population dynamics. Let $x(s)$ represent the population of $a$ species at time $s$ and let $x_0$ be the initial population. For some fixed time $L$ assume for $0 \le s \le L$ that births occur at a rate $r(s)$ and deaths are governed by a survival function $f(s)$ where a fraction $f(s-t)$ of the organisms born at time $t$ are alive at time $s$, $t \le s \le L$. It follows [Jer] that

$$x(s) = x_0 f(s) + \int_0^s f(s-t) r(t)\, dt, \qquad 0 \le s \le L. \tag{3.1}$$

If we assume that the birth rate is proportional to the population, $r(t) = \kappa x(t)$ for some constant $\kappa$, then (3.1) becomes a Volterra integral equation of the second kind (2.1) with $k(s,t) = \kappa f(s-t)$ and $G(s) = x_0 f(s)$. On the other hand if we introduce a birth control policy where the birth rate is reduced by an amount proportional to the final population so that $r(t) = \kappa x(t) - \alpha x(L)$, for a constant $\alpha$, then (3.1) reduces to the form (2.2) with $k(s,t) = \kappa f(s-t)$, $G(s) = x_0 f(s)$, and $\beta(s) = \int_0^s \alpha f(s-t)\, dt$.

We can now illustrate the results of §2 by assuming, say, that $x_0 = 1$, $\kappa = 1$, $L = 50$, $\alpha = .5$, and $f(s) = e^{-cs}$ with $c = .25$ so that $\beta(s) = \alpha(1-e^{-cs})/c$. For most functions $f(s)$, (2.1) or (2.2) with $k(s,t) = \kappa f(s-t)$ do not have exact solutions in terms of the usual transcendental functions and are not equivalent to ordinary differential equations. However, to calculate errors we choose a simple $f(s)$ so that (2.2) has an exact solution $x(s) = x_0[\alpha + (\kappa - c - \alpha)e^{(\kappa-c)(s-L)}]/[\alpha + (\kappa - c - \alpha)e^{-(\kappa-c)L}]$. In this case the solution to (2.6) is $\gamma(s) = \alpha[1 - e^{(\kappa-c)s}]/(\kappa - c)$. Thus by Corollary 2.13 for large $n$ the growth factor should be approximately

$$\frac{1 + \alpha[1 - e^{(\kappa-c)L}]/(\kappa - c)}{1 + \alpha(1 - e^{-cL})/c} = 4.3 \times 10^{15},$$

or larger, and partial pivoting should have large error growth.

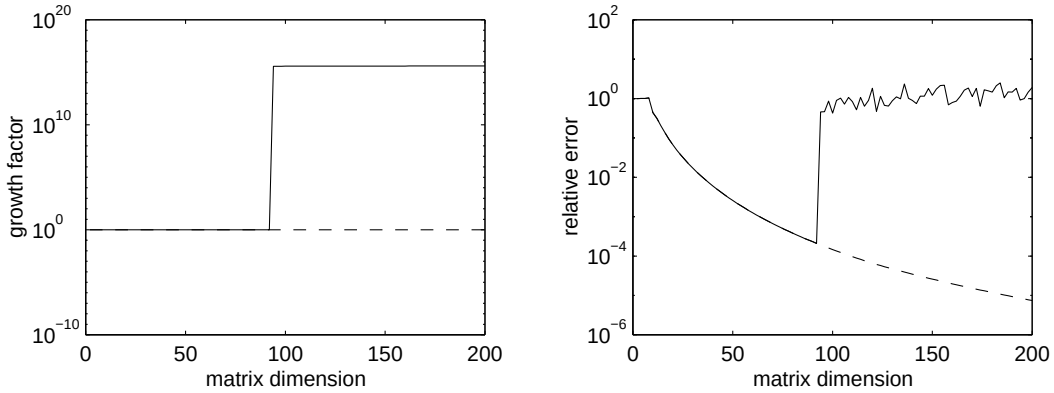


Fig. 3.2. *The growth factor (left), and the relative error (right), when solving* (2.3) *with partial pivoting (solid lines) and complete pivoting (dashed lines).*

In Fig. 3.2 we have plotted, versus $n$, the growth factors as computed by (1.1) when solving (2.3) for $p = 2$ by GEPP and complete pivoting. For partial pivoting row interchanges are required for $n \le 92$ and the growth factor remains moderate. However for $n \ge 93$ no row interchanges are required for partial pivoting and the growth factor becomes large. For $n = 200$ the computed growth factor is $4.02 \times 10^{15}$, close to the above theoretical estimate. Also in Fig. 3.2 we have plotted, versus $n$, the relative error in the approximate solution to (2.1) obtained by solving (2.3) by partial pivoting and complete pivoting. Here by relative error we mean

$\max_{i=1,\ldots,n} |x(s_i) - x_i| / \max_{i=1,\ldots,n} |x(s_i)|$, where $x(s)$ is the true solution. Gaussian elimination with complete pivoting is numerically stable and for this example the relative error decreases proportional to $h^4$ (since our numerical integration is based on the Simpson rule). On the other hand, for partial pivoting, the large growth factor leads a large relative error in the calculated answer. GEPP *is unstable and inaccurate for* $n \geq 93$.

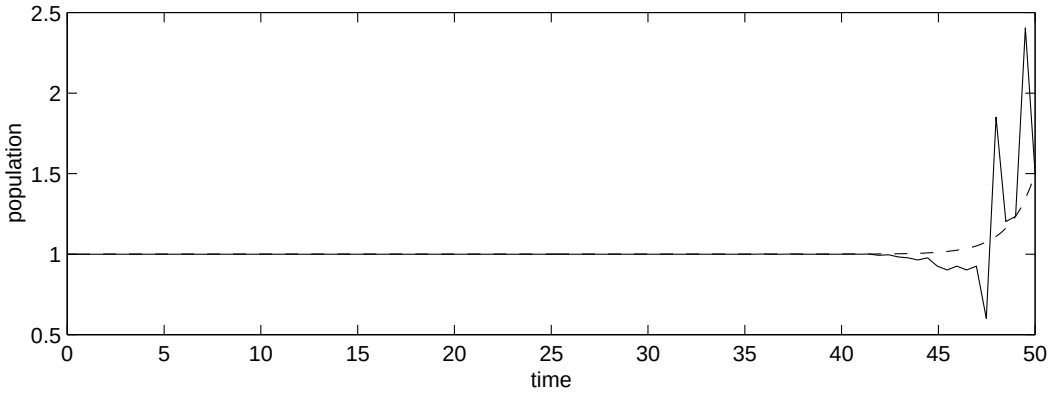


Fig. 3.3. *The true solution (dashed) to* (2.3) *and approximate solution (solid) calculated when using partial pivoting for* $n = 100$.

To further illustrate the difficulty with partial pivoting in Fig. 3.3 we have plotted the approximate solution to (2.2) obtained by solving (2.3) using partial pivoting with $n = 100$ and the true solution. In this case, to solve (2.3) we used the Matlab "\" operator, which is based on Linpack's implementation of GEPP, on a Sun Sparcstation. The large discrepancy for $s \geq 42$ is due to the instability of partial pivoting.

Finally we note that on a modern workstation it requires less than a second to set up and solve (2.3) for $n = 100$, say, and that the linear systems solved to produce the graphs in Figs. 3.2 and 3.3 are well conditioned. For example, the matrices used to produce Fig. 3.2 all have condition numbers less than 162.

For our second example we consider a boundary value problem. Suppose constants $L, k$, and $C$ and a function $g(t)$, $0 \leq t \leq L$, are known and that an unknown function $x(t)$, $0 \leq t \leq L$, satisfies

$$x'(t) = kx(t) + g(t), \quad 0 < t < L \quad \text{with } x(L) = Cx(0). \tag{3.4}$$

This example is simple enough so that we can find the solution exactly, but suppose that we wished to solve it numerically. We choose to first convert the differential equation (3.4) into an integral equation by integrating from zero to $s$ and substitute in $x(0) = x(L)/C$ to get

$$x(s) - \int_0^s kx(t)\,dt - x(L)/C = \int_0^s g(t)\,dt \equiv G(s). \tag{3.5}$$

This is of the form (2.2) with $k(s,t) = k$ and $\beta(s) = -1/C$. If we apply the quadrature method to (3.4) using the trapezoid rule to approximate the integrals, the resulting linear system $Ax = b$ is simple enough in this case so that we can exactly describe $L$ and $U$ in an $LU$

factorization of $A$. It is easy to check for $b = 1 - kh/2$ and $\omega = (1 + kh/2)/(1 - kh/2)$ that

$$A = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 & -1/C \\ -\frac{kh}{2} & 1-\frac{kh}{2} & 0 & \cdots & 0 & -1/C \\ -\frac{kh}{2} & -kh & 1-\frac{kh}{2} & \ddots & \vdots & \vdots \\ \vdots & \vdots & & \ddots & 0 & -1/C \\ -\frac{kh}{2} & -kh & \cdots & -kh & 1-\frac{kh}{2} & -1/C \\ -\frac{kh}{2} & -kh & \cdots & -kh & -kh & 1-1/C-\frac{kh}{2} \end{pmatrix} =$$

$$\begin{pmatrix} 1 & 0 & \cdots & & & 0 \\ -\frac{kh}{2} & 1 & \ddots & & & \vdots \\ -\frac{kh}{2} & -\frac{kh}{b} & 1 & & & \\ -\frac{kh}{2} & -\frac{kh}{b} & -\frac{kh}{b} & 1 & & \\ \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ -\frac{kh}{2} & -\frac{kh}{b} & -\frac{kh}{b} & \cdots & -\frac{kh}{b} & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & \cdots & & 0 & -\frac{1}{C} \\ 0 & b & \ddots & & \vdots & -\frac{b\omega}{C} \\ \vdots & \ddots & b & & & -\frac{b\omega^2}{C} \\ & & & \ddots & 0 & \vdots \\ & & & & b & -\frac{b\omega^{n-2}}{C} \\ 0 & \cdots & & & 0 & -\frac{b\omega^{n-1}}{C}+b \end{pmatrix}$$

From this factorization it follows easily that if $|kh| \leq 2/3$, no row interchanges are required in GEPP and that there will be large elements in $U$ if $b\omega^{n-1}/C$ is large.

As a specific example let $k = 1$, $L = 40$, and $C = 6$. An application producing such numbers would be a solution mixture problem over the time period $-40 \leq \hat{t} \leq 0$ where fluid with a solute concentration 1 enters a tank of volume 1 at a rate 1, where mixed fluid leaves at a rate 1, where the ratio of the final to the initial amount of solute in the tank is 6 and where we let $t = -\hat{t}$ to transform the domain to $0 \leq t \leq 40$. In this example, if $n \geq 61$ we have $kh \leq \frac{2}{3}$ and the growth factor is large. For $n = 61$, say, the condition number of $A$ is 88, the relative error in the calculated solution when solving (2.3) by a QR factorization is 1.1%. Yet if partial pivoting is used to solve (2.3) then due to a growth factor of $1.28 \times 10^{17}$ the relative error is 860%. GEPP *fails again for this concrete physical example.*

From the above decomposition it follows that when $C = 1$, say, and $kh = 2/3$ that the growth factor is $(2/3)(2^{n-1} - 1)$. This is quite close to the maximum theoretical growth factor of $2^{n-1}$. In comparison, the maximum growth factor reported in [Wri] is proportional to $(\sqrt{2})^n$. We might also note that we can generalize the results for our second example to boundary value problems for systems of $m$ differential equations. For such systems we can get growth factors, approximately, as big as $[2(1.5)^m - 1]^{(n/m)-1}/[3(1.5)^m - 3]$ where large growth occurs in the last $m$ columns of $U$. For example if $m = 5$ and $n = 90$, growth of $2 \times 10^{18}$ can occur in the last five columns of $U$.

# 4 Conclusions

The existence of practical examples where partial pivoting fails leads to a number of questions for the scientific computing community. To initiate a debate we propose the following answers.

1. *Why haven't practical examples where partial pivoting fails been reported earlier?* We expect that the primary reason is that such examples are indeed rare. Our problems and our approach to solving these problems were carefully selected. However, we should note that most software packages in numerical linear algebra do not report information about the growth factor and so it is possible that large growth factors do occur from time to time in practice, but have gone unreported.

2. *Should widely used packages in numerical linear algebra provide information about growth factors?* In view of our examples we believe that such information should be reported if, for example, a user requests an "expert" solution. The authors of Lapack are planning to incorporate this in future releases of Lapack [Dem]. For a dense matrix a bound on the effect of the growth

factor on the error in the calculated solution can be determined in $O(n^2)$ operations [CG], [ER], [GVL] which is small compared to the work in factoring the matrix.

Both Linpack [DBMS] and Lapack [ABB] had difficulty with the matrices in our examples. For the second example in §3, if $L = 60$, $n = 100$, $k = 1$, and $C = 6$ Linpack DGECO reports an estimated condition number of $1.5 \times 10^{12}$ when the actual condition of $A$ is 132. On the other hand, if $L = 40$ and $n = 61$ Linpack DGECO reports an estimated condition number of 534, which is much closer to 88, the true condition number. However, this might lead the user into thinking that the calculated solution is correct while, as we have seen in §3, it is not. With Lapack for our examples, if the growth factors were not too large, then the "expert" routine DGESVX successfully uses iterative refinement to overcome the inaccuracy due the large growth factor. However, if the growth factor is sufficiently large, iterative refinement does not converge. Also the Lapack condition estimator fails in some cases. Neither package warns the user that GEPP is unstable due to a large growth factor.

For our examples, the reason that the Linpack and Lapack condition estimators fail is that they rely on the ability to solve $Ax = y$. However due to large growth factors, the solutions to $Ax = y$ are not calculated correctly. The underlying condition estimators are not failing themselves, rather they are working with incorrect solutions to $Ax = y$.

3. *Are tests on random matrices useful for comparing and analyzing algorithms in numerical linear algebra?* Such tests are valuable in that they allow the quick generation of many examples. Also random matrices can be amenable to theoretical analysis. However tests with random matrices are not sufficient. For example, the matrices in our illustrations contained only negative numbers below the diagonal. If random matrices were generated so that signs of elements were random, then for $n = 50$, say, the probability of this sign pattern is negligible ($10^{-737}$). It would never show up in random sampling. We feel that tests with random matrices are often overused.

4. *Is there a need for a collection of test matrices arising from practical problems?* Since there are phenomena that occur in practice that do not show up in tests with random matrices, there is such a need. The collection could complement existing collections such as the one in [DGL]. Ideally the new collection would be in an easy-to-use format such as Matlab m files so that the collection is compact and flexible code could be included that would generate the matrices for different parameter choices, similar to the style used by Higham [Hig]. However, the focus would be on interesting matrices that can arise in practice. The collection of examples in Hansen's Regularization Tools [Han] would be a good beginning. Indeed the genesis of this paper came from observing some of the sign patterns for the matrices in Hansen's collection.

# References


[ABB] E. Anderson, Z. Bai, C. H. Bischof, J. W. Demmel, J. J. Dongarra, J. J. Du Croz, A. Greenbaum, S. J. Hammarling, A. M. McKenney, S. Ostrouchov, and D. Sorenson, *Lapack User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1992.

[Bak] C. T. H. Baker, *The Numerical Treatment of Integral Equations*, Oxford University Press, Oxford, 1977.

[Bur] T. A. Burton, *Volterra Integral and Differential Equations*, Academic Press, New York, 1983.

[CG] E. Chu and J. A. George, "An Algorithm to Estimate the Error in Gaussian Elimination Without Pivoting," *Tech. report CS-84-21*, University of Waterloo, Ontario, 1984.

[Dem] J. W. Demmel, private communication,June, 1993.

[DBMS] J. J. Dongarra, J. R. Bunch, C. B. Moler, and G. W. Stewart, *Linpack User's Guide*, Society for Industrial and Applied Mathematics, Philadelphia, 1979.

[DGL] I. S. Duff, R. G. Grimes, and J. G. Lewis, "Sparse matrix test problems," *ACM Trans. Math. Software*, vol. 15, pp. 1–14, 1989.

[DM] L. M. Delves and J. I. Mohamed, *Computational Methods for Integral Equations*, Cambridge University Press, Cambridge, 1985.

[ER] A. M. Erisman and J. K. Reid, "Monitoring the stability of the triangular factorization of a sparse matrix," *Numer. Math.*, vol. 22, pp. 183–186, 1974.

[GVL] G. H. Golub and C. F. Van Loan, *Matrix Computations,* 2nd ed., Johns Hopkins University Press, Baltimore, 1989.

[Han] P. C. Hansen, "Regularization Tools," *Report UNIC-92-03*, Danish Computing Center for Research and Education, Technical University of Denmark, 1992

[Hig] N. J. Higham, "Algorithm 694: A collection of test matrices in MATLAB", *ACM Trans. Math. Software*, vol. 17, pp. 289–305, 1991.

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

[Jer] A. J. Jerri, *Introduction to Integral Equations with Applications*, Marcel Dekker, New York, 1985.

[Linz] P. Linz, *Analytical and Numerical Methods for Volterra Equations*, Society for Industrial and Applied Mathematics, Philadelphia, 1985.

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

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

[Wri] S. J. Wright, "A collection of problems for which Gaussian elimination with partial pivoting is unstable," *SIAM J. Sci. Statist. Comput.*, vol 14, pp. 231–238, 1993.