Random Matrices Generating Large Growth in LU Factorization with Pivoting

Higham, Desmond J. and Higham, Nicholas J. and Pranesh, Srikara

2020

MIMS EPrint: **2020.13**

ISSN 1749-9097

RANDOM MATRICES GENERATING LARGE GROWTH IN LU FACTORIZATION WITH PIVOTING[*]

DESMOND J. HIGHAM[†], NICHOLAS J. HIGHAM[‡], AND SRIKARA PRANESH[†]

Abstract. We identify a class of random, dense $n \times n$ matrices for which LU factorization with any form of pivoting produces a growth factor of at least $n/(4 \log n)$ for large n with high probability. The condition number of the matrices can be arbitrarily chosen and large growth also happens for the transpose. No previous matrices with all these properties were known. The matrices can be generated by the MATLAB function `gallery('randsvd',..)`, and they are formed as the product of two random orthogonal matrices from the Haar distribution with a diagonal matrix having only one diagonal entry different from 1, which lies between 0 and 1 (the "one small singular value" case). Our explanation for the large growth uses the fact that the maximum absolute value of any element of a Haar distributed orthogonal matrix tends to be relatively small for large n. We verify the behavior numerically, finding that for partial pivoting the actual growth is significantly larger than the lower bound, and much larger than the growth observed for random matrices with elements from the uniform $[0, 1]$ or standard normal distributions. We show more generally that a rank-1 perturbation to an orthogonal matrix producing large growth for any form of pivoting also generates large growth under reasonable assumptions. Finally, we demonstrate that GMRES-based iterative refinement can provide stable solutions to $Ax = b$ when large growth occurs in low precision LU factors, even when standard iterative refinement cannot.

Key words. LU factorization, Gaussian elimination, large growth factor, pivoting, random orthogonal matrix, Haar distribution, MATLAB, randsvd, GMRES-based iterative refinement

AMS subject classifications. 65F05

1. Introduction.

The MATLAB code

```
rng(1), n = 750; kappa = 1e8; mode = 2;
A = gallery('randsvd',n,kappa,mode,[],[],1);
[L,U,P,Q,growth] = gep(A,'p'); growth % Partial pivoting
```

produces the output

```
growth =
  103.7971
```

The code uses the function `gep` from the Matrix Computation Toolbox [15] to compute the growth factor for LU factorization with partial pivoting on a random $n \times n$ matrix A with $n = 750$. The growth factor is defined by

$$\rho_n(A) = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|},$$

where $a_{ij}^{(k)}$ ($k = 1\colon n$) are the elements at the kth stage of the factorization [16, sect. 9.3], [35]. Growth of over 100 for a matrix of this size with partial pivoting is very unusual. Unusually large growth is also obtained for the same matrix with rook pivoting and complete pivoting:

[*] Draft version of May 14, 2020.

Funding: This work was supported by Engineering and Physical Sciences Research Council grant EP/P020720/1 and the Royal Society.

[†] School of Mathematics, University of Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9, 3FD, UK (`d.j.higham@ed.ac.uk`)

[‡] Department of Mathematics, University of Manchester, Manchester, M13 9PL, UK (`nick.higham@manchester.ac.uk`, `srikara.pranesh@manchester.ac.uk`).

```
>> [L,U,P,Q,growth] = gep(A,'r'); growth % Rook pivoting
growth =
   57.1362
>> [L,U,P,Q,growth] = gep(A,'c'); growth % Complete pivoting
growth =
   43.2643
```

(See [16, sect. 9.1], [30], [35] for details of all these pivoting strategies.) Large growth factors are undesirable because they are a warning that numerical instability is likely in the LU factorization, as originally shown by Wilkinson [35].

Several classes of matrices generating large growth factors for partial pivoting are known. Wilkinson [35, p. 327], [36, p. 212] showed that the $n \times n$ matrix of the form illustrated for $n = 4$ by

$$A_n = \begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{bmatrix}$$

gives $\rho_n = 2^{n-1}$, which is the worst case for partial pivoting. Higham and Higham [17] give examples of practically occurring $n \times n$ matrices for which $\rho(A) \gtrsim n/2$ for any pivoting strategy; they are all orthogonal matrices or well-conditioned diagonal scalings of orthogonal matrices. Wright [37] describes a class of two-point boundary value problems for which the multiple shooting method leads to a linear system on which partial pivoting suffers exponential growth. The matrix is block lower bidiagonal, except for a nonzero block in the top right-hand corner. Foster [9] shows that a quadrature method for solving a practically occurring Volterra integral equation gives rise to dense linear systems for which partial pivoting again gives growth factors exponential in the dimension. In all these examples the matrices are well conditioned.

The matrix in our example has 2-norm condition number $\kappa_2(A) = \sigma_1/\sigma_n = 10^8$, with singular value decomposition (SVD) of the form

$$A = P\Sigma Q^T \in \mathbb{R}^{n\times n}, \quad P^TP = Q^TQ = I, \tag{1.1a}$$

$$\Sigma = \mathrm{diag}(1,\dots,1,\sigma_n), \quad 1 \ge \sigma_n \ge 0. \tag{1.1b}$$

Here, $n-1$ of the singular values of A are equal to 1 and the last one is less than or equal to 1. The matrices P and Q are orthogonal matrices from the Haar distribution, that is, they are distributed according to the Haar measure, which is the unique measure on the orthogonal matrices that is invariant under multiplication on the left and right by orthogonal matrices [27]. A Haar distributed random orthogonal matrix can be obtained as the orthogonal QR factor of a matrix with elements from the normal (0,1) distribution, provided that the factorization is normalized so that the diagonal elements of R are nonnegative [3], [32].

Matrices of the form (1.1) are generated by a MATLAB function call of the form `gallery('randsvd',n,kappa,mode)` with `kappa` $= \sigma_n^{-1} \ge 1$ and `mode = 2` (the default value of `mode` is 3, which produces geometrically distributed singular values). Figure 1.1 shows the results of an experiment in which we generated matrices this way for dimensions $n = 100\colon 100\colon 2500$ and computed the growth factors for partial pivoting, rook pivoting, and complete pivoting. For each dimension we generated 12 matrices and took the mean growth factor. The figure illustrates the results for $\kappa_2(A) = 10^2, 10^6, 10^{10}$. As above, we used the the `gep` function, which computes the exact growth factor (as opposed to the lower bound $\max_{i,j}|u_{ij}|/\max_{i,j}|a_{ij}|$ that

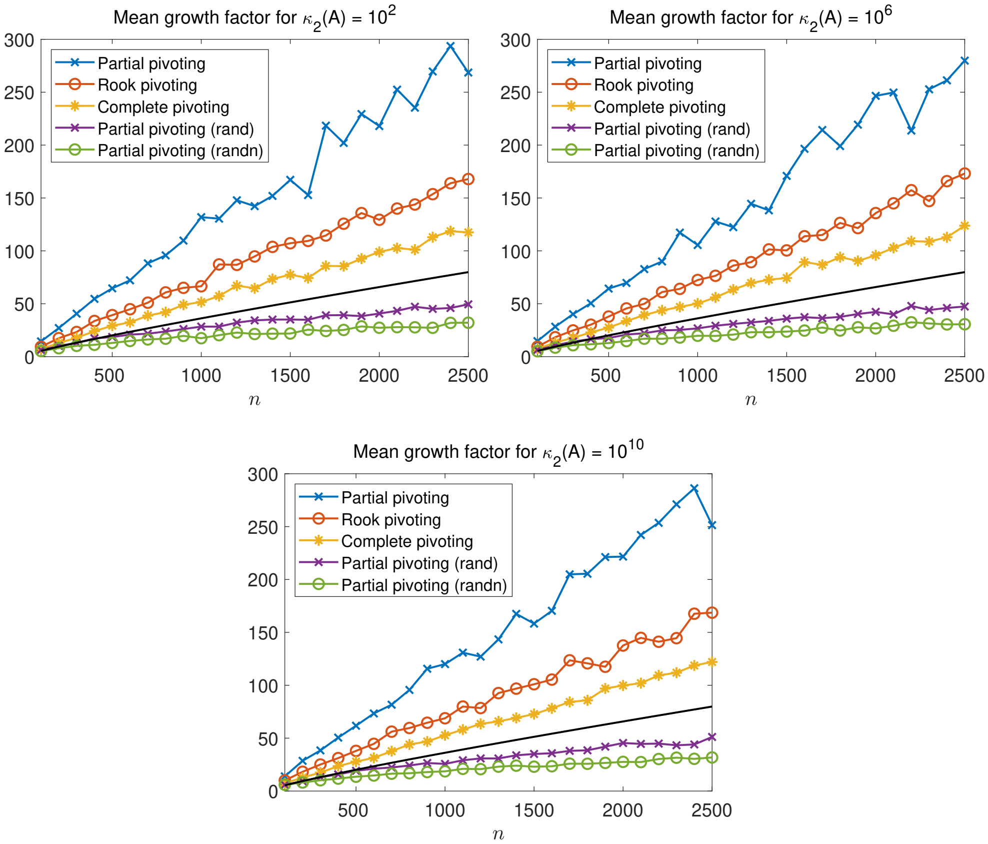

Fig. 1.1. *Mean growth factors for matrices* (1.1) *with* $\kappa(A) = 10^2, 10^6, 10^{10}$ *and for* `rand` *and* `randn` *matrices, with* 12 *samples for each* n*. The black curve is* $n/(4 \log n)$*.*

must be used if we have access to the LU factors but not the intermediate quantities). We used the Parallel Computing Toolbox [29] to speed up the computations. We see that *irrespective of the condition number*, the growth factor increases with n at a rate roughly proportional to n for all three pivoting strategies. The largest growth factor observed in this experiment was 483. By contrast, for random matrices with elements from the uniform $[0, 1]$ distribution (`rand` in MATLAB) or the normal $(0, 1)$ distribution (`randn` in MATLAB) the figure shows that the growth factor for partial pivoting grows more slowly than linearly in n (as previously observed in [34]).

The significance of the matrices (1.1) is that they provide a new class of dense matrices A for which

- A generates large growth for any pivoting strategy,
- A^T also generates large growth for any pivoting strategy,
- $\kappa_2(A)$ is arbitrary and is easily assigned by choosing σ_n in (1.1).

The existing examples of large growth mentioned above are all well conditioned, some produce large growth only for partial pivoting, and not all of them produce large growth for A^T.

A growth factor of order αn for some constant $\alpha < 1$ with $\alpha > 1/10$ (say) may not seem to be a serious problem, given that the worst-case growth for partial pivoting is 2^{n-1}. But matrix dimensions in practical problems are increasing, with dense linear systems of order 10^7 being solved on today's largest machines [4]. The backward error bound for solution of a linear system $Ax = b$ by LU factorization is proportional to $\rho_n u$, where u is the unit roundoff [16, Thm, 9.5], so growth of order n can be problematic. Matters are exacerbated by the growing use of low precision arithmetics such as IEEE half precision ($u \approx 5 \times 10^{-4}$) [21] and bfloat16 ($u \approx 4 \times 10^{-3}$) [22]. Low precision LU factorizations are being used in combination with iterative refinement to achieve faster solution times [11], [12], [13], [20] and the new HPL-AI benchmark uses this approach [8]. In low precision arithmetic large growth can even cause overflow. Indeed we spotted the relatively large growth factor for randsvd matrices with mode 2 because it led to overflow in LU factorization on these matrices in IEEE half precision arithmetic, for which the largest finite number is of order 6×10^4.

In the next section we prove that for large n, large growth occurs with high probability for the matrices (1.1) with $\kappa_2(A) = 1$. This is the case of Haar distributed orthogonal matrices. In section 3 we show that if an orthogonal matrix generates large growth for any pivoting strategy then large growth persists after a rank-1 perturbation, under reasonable assumptions. In section 4 we specialize the results to a rank-1 perturbation of a Haar distributed orthogonal matrix, that is, matrices of the form (1.1) with an arbitrary $\kappa_2(A)$. In section 5 we provide an alternative analysis for the growth factor of a rank-1 perturbation of an orthogonal matrix based on the Sherman–Morrison formula. In section 6 we investigate the ability of iterative refinement to overcome the instability in LU factorization caused by large growth factors.

2. Orthogonal matrices from the Haar distribution.

We first consider the case where $\sigma_n = 1$ in (1.1), so that $A = PQ^T$ with P and Q orthogonal matrices from the Haar distribution. Since the Haar distribution is invariant under left or right multiplication by an orthogonal matrix, A is also Haar distributed, so we are effectively taking a single sample from the Haar distribution.

We need the following result from [17].

THEOREM 2.1. *Let $A \in \mathbb{C}^{n\times n}$ be nonsingular and set $\alpha = \max_{i,j} |a_{ij}|$, $\beta = \max_{i,j} \big|\big(A^{-1}\big)_{ij}\big|$, and $\theta = (\alpha\beta)^{-1}$. Then $\theta \le n$, and for any permutation matrices Π_r and Π_c such that $\Pi_r A \Pi_c$ has an LU factorization, the growth factor for GE without pivoting on $\Pi_r A \Pi_c$ satisfies $\rho(A) \ge \theta$.*

Theorem 2.1 is used in [17] to show that for certain specific matrices that are orthogonal, or are well conditioned diagonal scalings of orthogonal matrices, the inequality $\rho_n(A) \gtrsim n/2$ holds for any pivoting strategy.

Jiang [24] shows that for $n \times n$ matrices A drawn from the Haar distribution, $\Pr\big(\max_{i,j} |a_{ij}| > 2\sqrt{\log(n)/n}(1+\epsilon)\big) \to 0$ as $n \to \infty$ for any $\epsilon > 0$. Hence $\max_{i,j} |a_{ij}| \lesssim 2\sqrt{\log(n)/n}$ for large n with high probability. Since $A^{-1} = A^T$, we can take $\alpha = \beta = 2\sqrt{\log(n)/n}$ in Theorem 2.1 and conclude that

$$\rho_n(A) \gtrsim \frac{n}{4\log n} \tag{2.1}$$

for large n with high probability for *any* pivoting strategy.

The lower bound in (2.1) is not as large as those for the (scaled) orthogonal matrices in [17], but those matrices are non-random. Orthogonal matrices from the Haar distribution are the first class of random orthogonal matrices to be shown to give large growth.

Figure 2.1 shows the results of an experiment in which we generated Haar distributed orthogonal matrices of dimensions $n = 100\colon 100\colon 2500$ and computed the

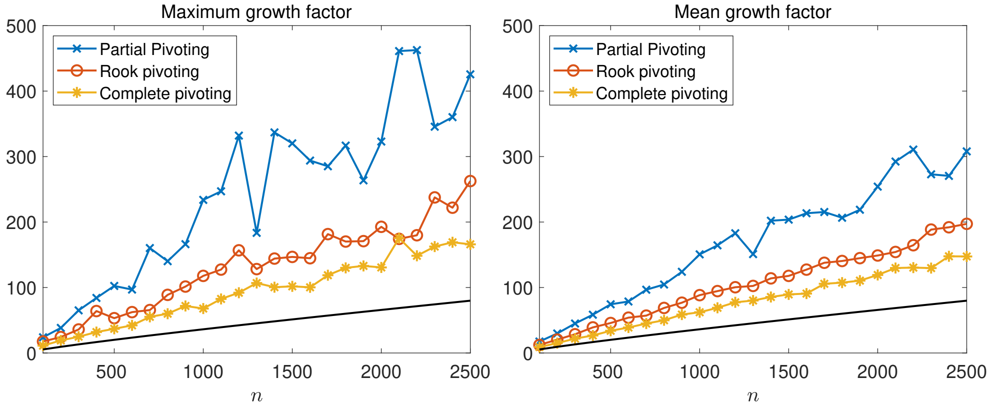

FIG. 2.1. *Growth factors for orthogonal matrices from the Haar distribution: maximum growth factor (left) and mean growth factor (right) over 12 samples for each n. The black curve is $n/(4 \log n)$.*

growth factors for partial pivoting, rook pivoting, and complete pivoting. For each dimension we generated 12 matrices and we show the maximum and average growth factors. We see that all the growth factors exceed the approximate lower bound and that they increase with n a little more rapidly than the lower bound. As expected, the growth factor for partial pivoting exceeds that for rook pivoting which in turn exceeds that for complete pivoting.

3. Rank-1 perturbations of orthogonal matrices. Before we treat general matrices of the form (1.1), we consider the larger class of matrices

$$A = W + xy^T, \tag{3.1}$$

where W is orthogonal. Later, we will assume that $\|x\|_2 \le 1$ and $\|y\|_2 \le 1$, but for now we make no assumptions on the norms.

For the moment, we take W to be any nonsingular matrix that has an LU factorization. We know that the U factor in the LU factorization of $W \in \mathbb{R}^{n\times n}$ is given explicitly by [16, sect. 9.2]

$$u_{ij} = \frac{\det\bigl(W(1\colon i, [1\colon i-1,\, j])\bigr)}{\det(W_{i-1})}, \quad i \le j, \tag{3.2}$$

where $W_j = W(1\colon j, 1\colon j)$. Suppose A in (3.1) has the LU factorization $A = \widetilde{L}\widetilde{U}$. It is easy to show that

$$\det(A) = \det(W)(1 + y^T W^{-1} x). \tag{3.3}$$

Now

$$A(1\colon i, [1\colon i-1,\, j]) = W(1\colon i, [1\colon i-1,\, j]) + x(1\colon i) y([1\colon i-1,\, j])^T$$

and hence, analogously to (3.3), assuming $W(1\colon i, [1\colon i-1,\, j])$ is nonsingular,

$$\begin{aligned}\det(A(1\colon i, [1\colon i-1,\, j]) &= \det\bigl(W(1\colon i, [1\colon i-1,\, j])\bigr)\\ &\qquad \times \bigl(1 + y([1\colon i-1,\, j])^T W(1\colon i, [1\colon i-1,\, j])^{-1} x(1\colon i)\bigr).\end{aligned}$$

Similarly,

$$\det(A_{i-1}) = \det\bigl(W_{i-1}\bigr)\bigl(1 + y(1{:}\,i-1)^T W_{i-1}^{-1}\, x(1{:}\,i-1)\bigr),$$

so by (3.2) applied to A we have

$$\widetilde{u}_{ij} = \frac{\det\bigl(W(1{:}\,i, [1{:}\,i-1,\, j])\bigr)\bigl(1 + y([1{:}\,i-1,\, j])^T W(1{:}\,i, [1{:}\,i-1,\, j])^{-1} x(1{:}\,i)\bigr)}{\det\bigl(W_{i-1}\bigr)\bigl(1 + y(1{:}\,i-1)^T W_{i-1}^{-1}\, x(1{:}\,i-1)\bigr)}.$$

Combining this equation with (3.2) we obtain

$$\frac{\widetilde{u}_{ij}}{u_{ij}} = \frac{1 + y([1{:}\,i-1,\, j])^T W(1{:}\,i, [1{:}\,i-1,\, j])^{-1} x(1{:}\,i)}{1 + y(1{:}\,i-1)^T W_{i-1}^{-1}\, x(1{:}\,i-1)}. \tag{3.4}$$

Assume now that W is orthogonal and that there is large growth in its LU factorization. Then $|u_{ij}|$ must be large for some indices i and j. In practice we would expect these i and j to be large (essentially because otherwise a much smaller matrix can be constructed that gives the same growth), and experiments confirm that this is usually the case. For example, with $n = 1000$ and partial pivoting we found that for the orthogonal MATLAB matrices `gallery('orthog',n,j)` with $j = 1, 2, 5, 6$, all of which have growth factors at least 500, the largest element of U was in the (n, n) position in every case, while for one hundred 1000×1000 matrices of the form (1.1) with $\sigma_{\min} = 10^{-8}$, the largest element of U was always in row 953 or higher.

From (3.4), we have the lower bound

$$\left|\frac{\widetilde{u}_{ij}}{u_{ij}}\right| \ge \frac{1 - \sigma_{\min}(W(1{:}\,i, [1{:}\,i-1,\, j]))^{-1}\|x\|_2\|y\|_2}{1 + \sigma_{\min}(W_{i-1})^{-1}\|x\|_2\|y\|_2}, \tag{3.5}$$

where $\sigma_{\min}$ denotes the smallest singular value, but this bound is too weak to be informative because the $\sigma_{\min}^{-1}$ terms can both be larger than 1.

Let B be an $(n-k) \times (n-k)$ submatrix of W, which we identify with either of the submatrices on the right-hand side of (3.4). We will argue that the second terms in the numerator and the denominator of (3.4) should be safely less than 1 for large n and small k. By Lemma A.1, B has $n - 2k$ singular values equal to 1 as long as $k < n/2$. Let B have the SVD $B = \widetilde{U} D \widetilde{V}^T$. Assume $\|x\|_2 \le 1$ and $\|y\|_2 \le 1$ and let $g, h \in \mathbb{R}^{n-k}$ denote subvectors of x and y. Then, assuming B is nonsingular,

$$g^T B^{-1} h = g^T \widetilde{V} D^{-1} \widetilde{U}^T h \tag{3.6}$$

$$= \widetilde{g}^T D^{-1} \widetilde{h} \quad \bigl(\widetilde{g} = \widetilde{V}^T g, \quad \widetilde{h} = \widetilde{U}^T h\bigr) \tag{3.7}$$

$$= \sum_{i=1}^{n-2k} \widetilde{g}_i \widetilde{h}_i + \sum_{i=n-2k+1}^{n-k} \widetilde{g}_i \widetilde{h}_i \sigma_i(B)^{-1}. \tag{3.8}$$

If we assume that $\widetilde{g}$ and $\widetilde{h}$ have elements of similar magnitudes, so that the elements are of order at most $(n-k)^{-1/2}$, then (3.8) expresses $g^T B^{-1} h$ as the sum of a term of order $(n-2k)/(n-k)$ and a term of order $\sum_{n-2k+1}^{n-k} \sigma_i(B)^{-1}/(n-k)$. For large n and $k \ll n$, the second term will be less than 1 as long as the sum of the reciprocals of the k smallest singular values of B is smaller than $n - k$; since B is a large submatrix of an orthogonal matrix this is likely to be the case unless B is very special. Under these conditions we will have $|g^T B^{-1} h| < 1$.

We conclude that as long as $\|x\|_2 \le 1$ and $\|y\|_2 \le 1$ we can expect the right-hand side of (3.4) to be of order 1 and hence a large $|u_{ij}|$ to imply a large $|\widetilde{u}_{ij}|$, that is, A will have a large growth factor if W does.

This analysis is for LU factorization without pivoting. The effect of pivoting is easily incorporated by multiplying on the left and the right of (3.1) by permutation matrices. It is possible that different pivot sequences are used in the LU factorizations of A and W, but this is not a concern because we are interested in orthogonal W for which large growth is obtained for any pivot sequence.

To summarize, we have argued that if the LU factorization of the orthogonal matrix W suffers large growth with any form of pivoting then the same will be true of a rank-1 perturbation of 2-norm bounded by 1 provided that (a) large elements of U occur in the bottom right of the matrix, (b) the corresponding submatrices of W are not too ill conditioned, and (c) the relevant subvectors of the vectors making up the rank-1 perturbation have roughly equal elements after transformation by the singular vector matrices of the submatrices in (b).

A key point is that none of this analysis is particular to W being a Haar distributed matrix. As an illustration, consider the following MATLAB code, which uses an orthogonal matrix [17, Eq. (2.3)] that is known to give growth factor at least $(n+1)/2$.

```
rng(1), n = 2000;
W = gallery('orthog',n);          % Orthog. matrix giving large growth.
[L,U,P,Q,growth_W] = gep(W,'p'); % Partial pivoting
growth_W
for k = 1:100
  x = randn(n,1); x = x/norm(x); y = randn(n,1); y = y/norm(y);
  A = W + x*y';
  [L,U,P,Q,growth_A(k)] = gep(A,'p');  % Partial pivoting
  fprintf('%3.0f %9.2e\n', k, growth_A(k))
end
[min(growth_A) mean(growth_A) max(growth_A)]
```

The growth factor for W is 1046 versus a lower bound of 1000 (rounding all values to the nearest integer). The minimum, mean, and maximum growth factors for $A = W + xy^T$ are 797, 997, and 1485, confirming that large growth persists (and can even increase) under these rank-1 perturbations of unit 2-norm. The conditions (a), (b), and (c) of the previous paragraph are satisfied in every case in this example.

4. Randsvd matrices.

We now consider matrices A of the form (1.1) with P and Q from the Haar distribution. We have

$$A = PQ^T + (\sigma_n - 1)p_n q_n^T, \tag{4.1}$$

where p_n and q_n are the last columns of P and Q, respectively. Since $W = PQ^T$ is Haar distributed, it gives a large growth factor with high probability, as shown in section 2. We have verified experimentally that for matrices of the form (4.1) the conditions (a), (b), and (c) in the penultimate paragraph of the previous section are usually satisfied. Therefore the analysis of the previous section applies and provides an explanation for the behavior seen in Figure 1.1.

It is worth emphasizing that from the argument in section 3 we know that Haar distributed orthogonal matrices maintain large growth under a wider class of rank-1 perturbations than (4.1). Figure 4.1 plots growth factors for partial pivoting for $A = W + xy^T$ with W an orthogonal matrix from the Haar distribution and x and

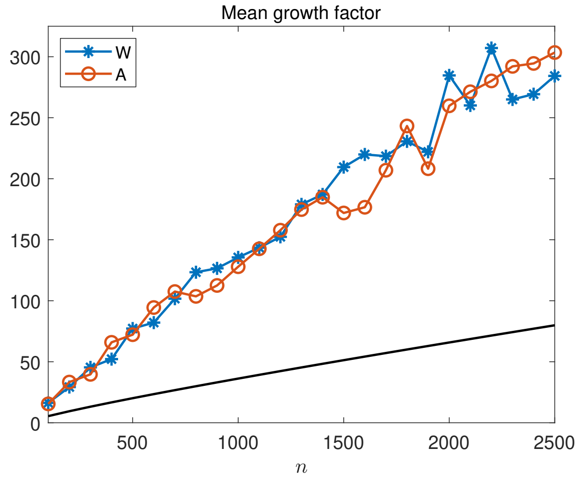

FIG. 4.1. *Mean growth factors for partial pivoting for orthogonal matrices W from the Haar distribution and for $A = W + xy^T$, where x and y are generated with elements from the uniform distribution on $[0, 1]$ and then scaled so that $\|x\|_2 = \|y\|_2 = 1$. The mean is over 12 matrices A and W for each n. The black curve is $n/(4 \log n)$.*

y generated with elements from the uniform distribution on $[0, 1]$ and then scaled so that $\|x\|_2 = \|y\|_2 = 1$. For each $n = 100\colon 100\colon 2500$ we generated 12 random A and took the mean growth factor. We see that the growth factors for A are very similar to those for W.

It is interesting to note that, unlike for (4.1) with small σ_n, $A = W + uv^T$ is very well conditioned when u and v are random unit 2-norm vectors with independent entries from the same distribution. Indeed Benaych-Georges and Nadakuditi [1, sec. 3.2] show that, almost surely

$$\sigma_1(A) \to \frac{1+\sqrt{5}}{2}, \quad \sigma_n(A) \to \frac{-1+\sqrt{5}}{2} \quad \text{as } n \to \infty,$$

and we know that the other $n - 2$ singular values remain at 1 (because they are the square roots of the eigenvalues of A^TA, which is the identity plus a rank-2 matrix). Hence $\kappa_2(A) \approx (1+\sqrt{5})/(-1+\sqrt{5}) \approx 2.8$ for large n. In the experiment just mentioned the values of $\kappa_2(A)$ were all on the interval $[2.49, 2.93]$.

We have also observed experimentally that large growth is preserved under rank-k perturbations of Haar distributed orthogonal matrices for $k \geq 1$, with the growth factor decreasing as k increases.

5. Analysis via the Sherman–Morrison formula. In section 3 we used the explicit characterization (3.2) of U in order to study growth factors for rank-1 perturbations xy^T of orthogonal matrices, focusing on the case where $\|x\|_2 \leq 1$ and $\|y\|_2 \leq 1$.

In this section we look at rank-1 perturbations of orthogonal matrices from a different perspective, applying the Sherman–Morrison formula and then making use of the indirect bound from Theorem 2.1. We will show that that growth of order $n/(4\log(n))$ arises for any rank-1 perturbation xy^T of a Haar distributed orthogonal matrix whenever the vectors x and y have 1-norm bounded by 1 and have elements of roughly uniform magnitude.

Theorem 5.1. *Let $A \in \mathbb{R}^{n\times n}$ have the form*

$$A = W + txy^T, \tag{5.1}$$

where $W \in \mathbb{R}^{n\times n}$ is orthogonal, $t \in [0,1]$, and $x, y \in \mathbb{R}^n$ satisfy $\|x\|_1 \le 1$ and $\|y\|_1 \le 1$. Let

$$\alpha_w = \max_{i,j} |w_{ij}|,$$

and suppose that $\alpha_w < 1$. Then A is nonsingular and, for any pivoting strategy producing an LU factorization for A, we have the lower bound

$$\rho_n(A) \ge \frac{1 - t\alpha_w}{\alpha_w\,(\alpha_w + t\|x\|_\infty \|y\|_\infty)}. \tag{5.2}$$

Proof. The Sherman–Morrison formula [14] gives, since W is orthogonal,

$$A^{-1} = W^T - \frac{tW^T xy^T W^T}{1 + ty^T W^T x}. \tag{5.3}$$

Using the Hölder inequality ($|f^T g| \le \|f\|_\infty \|g\|_1$) we have

$$\max_k |W^T x|_k \le \alpha_w, \quad \max_k |Wy|_k \le \alpha_w.$$

Also, $|y^T W^T x| \le \alpha_w < 1$, which confirms that the denominator in (5.3) is nonzero and hence that A is nonsingular, and indeed

$$\left| \frac{tW^T xy^T W^T}{1 + ty^T W^T x} \right|_{ij} \le \frac{t\alpha_w^2}{1 - t\alpha_w}.$$

Hence, in (5.3),

$$\max_{i,j} |A^{-1}|_{ij} \le \alpha_w + \frac{t\alpha_w^2}{1 - t\alpha_w} = \frac{\alpha_w}{1 - t\alpha_w}.$$

Using this bound in Theorem 2.1, along with $\max_{i,j} |a_{ij}| \le \alpha_w + t\|x\|_\infty \|y\|_\infty$, we arrive at (5.2). □

In the case where W is an orthogonal matrix from the Haar distribution, we have $\alpha_w \lesssim 2\sqrt{\log(n)/n}$ for large n with high probability, as noted in section 2. In this case, Theorem 5.1 gives

$$\rho(A) \gtrsim \frac{1}{4\log(n)/n + 2t\sqrt{\log(n)/n}\,\|x\|_\infty \|y\|_\infty}. \tag{5.4}$$

So if

$$t\|x\|_\infty \|y\|_\infty = o(\sqrt{\log(n)/n}) \tag{5.5}$$

we obtain

$$\rho_n(A) \gtrsim \frac{n}{4\log n}, \tag{5.6}$$

which matches the bound (2.1) for the unperturbed case. Under the constraints $\|x\|_1 \le 1$ and $\|y\|_1 \le 1$, the additional requirement (5.5) will hold when the vectors x and y have elements of roughly equal magnitude, because then $\|x\|_\infty \approx \|y\|_\infty \approx 1/n$.

If u and v are constructed by drawing elements independently from the uniform [0,1] distribution then each element has mean value $1/2$, so $\|u\|_1 \approx n/2$, and likewise for v. Let $x = u/\|u\|_1$, $y = v/\|v\|_1$, and $t = 1$. Then $\|x\|_1 = \|y\|_1 = 1$ and $\|x\|_\infty \approx \|y\|_\infty \approx 2/n$, so (5.5) is satisfied and hence (5.6) holds.

Now let u and v be columns of Haar distributed orthogonal matrices and consider a perturbation uv^T. For large n, the vectors u and v have components that are approximately independent normally distributed random variables with mean 0 and standard deviation $n^{-1/2}$ [24, Cor. 1], [25, Thm. 3]. Since the mean of the absolute value of a standard normal random variable is $(2/\pi)^{1/2}$ [26, Eq. (3)], the 1-norms of u and v have means approximately $(2/\pi)^{1/2}n$. Moreover, the ∞-norm of a random vector $z \in \mathbb{R}^n$ with independent standard normal components has mean and variance bounded above by terms of order $\sqrt{\log n}$ and $\log n$, respectively [7, Appendix A]; an application of the Chebyshev inequality [2, p. 80] then allows us to bound $\|z\|_\infty$ by order $\sqrt{n}\log n$ with high probability. Identifying u and v with $z/\sqrt{n}$, we find that $x = u/\|u\|_1$ and $y = v/\|v\|_1$ each have ∞-norms bounded above by order $\log n/n$ with high probability whence, with $t = 1$, (5.5) and (5.6) follow.

Theorem 5.1 also shows that existing growth factor bounds obtained for orthogonal matrices, such as those in [17], are essentially unchanged under appropriate rank-1 perturbations.

We conclude that Theorem 5.1 can guarantee the preservation of large growth factors under rank-1 perturbations. However, the theorem constrains the 1-norms of x and y. Therefore the result obtained with the Sherman–Morrison formula is weaker than that from section 3, though it requires fewer assumptions.

6. Curing instability with mixed precision iterative refinement. When an LU factorization of A suffers large growth and we use the factorization to solve $Ax = b$, the solution usually (but not always [17]) has a correspondingly large backward error. Suppose A is one of the types of matrix identified in this paper that has an LU factorization with a large growth factor; how can we obtain a backward stable solution to $Ax = b$ using this factorization? The natural answer is to apply iterative refinement. Indeed, it has been known since the 1970s that iterative refinement can cure instability in LU factorization [23], [31].

A recent usage of iterative refinement is with the LU factorization computed at a lower precision than the working precision, with residuals possibly computed in extra precision, and with the refinement equation solved either by substitution using the LU factors (denoted LU-IR) or by GMRES using the LU factors as preconditioners (known as GMRES-IR). GMRES-IR was proposed by Carson and Higham in [5], [6] and the analysis therein (notably [6, Thm. 4.1]) implies that it can tolerate instability in the factorization provided that the convergence of GMRES is not hindered by a lower quality preconditioner. Element growth is likely to reduce the quality of the preconditioner, so it is of interest to test experimentally what is the effect of a large growth factor on the convergence of GMRES.

We present an experiment in which we used mode 2 `gallery('randsvd')` matrices (that is, matrices of the form (1.1)) of varying dimensions, and $\kappa_2(A) = 10^2$ and $\kappa_2(A) = 10^7$. The iterative refinement algorithms that we use are characterized by a triple of precisions: (p_1, p_2, p_3), where p_1 is the precision at which the LU factorization is computed, p_2 is the working precision, and p_3 is the precision at which the residual is computed. We consider three precision combinations: (H, S, D), (H, D, D), and (S, D, D), where H, S, and D denote half precision ($u \approx 4.88 \times 10^{-4}$), single precision ($u \approx 5.96 \times 10^{-8}$), and double precision ($u \approx 1.11 \times 10^{-16}$), respectively. Half pre-

Table 6.1

Total number of iterative refinement steps in standard iterative refinement (LU-IR), and GMRES-IR for different precision combinations for $\kappa_2(A) = 10^2$. Numbers in parentheses denote the total number of GMRES iterations.

	(H, S, D)		(H, D, D)		(S, D, D)	
n	LU-IR	GMRES-IR	LU-IR	GMRES-IR	LU-IR	GMRES-IR
500	1	2 (2)	5	3 (6)	2	2 (2)
750	1	1 (1)	5	3 (6)	2	2 (2)
1000	1	1 (1)	6	3 (6)	2	2 (2)
1250	1	2 (2)	6	3 (6)	2	2 (2)
1500	1	1 (1)	5	3 (6)	2	2 (2)
1750	1	1 (2)	5	3 (6)	2	2 (2)
2000	1	1 (1)	6	3 (6)	2	2 (2)
2250	1	1 (2)	6	3 (7)	2	2 (2)
2500	1	1 (2)	6	2 (6)	2	2 (2)

Table 6.2

Growth factors for partial pivoting and condition number of left preconditioned matrix for $\kappa_2(A) = 10^2$.

		$\kappa_\infty(\widehat{U}^{-1}\widehat{L}^{-1}A)$		
n	ρ_n	(H, S, D)	(H, D, D)	(S, D, D)
500	44.61	6.44e+00	6.44e+00	1.00
750	92.63	8.43e+00	8.43e+00	1.00
1000	221.61	1.59e+01	1.59e+01	1.00
1250	125.77	2.13e+01	2.13e+01	1.00
1500	167.26	2.09e+01	2.09e+01	1.00
1750	349.38	3.31e+01	3.31e+01	1.00
2000	170.52	3.91e+01	3.91e+01	1.01
2250	256.11	5.41e+01	5.41e+01	1.01
2500	248.20	6.45e+01	6.45e+01	1.01

cision computations are performed using the `chop` function[1] of Higham and Pranesh [19]. The right-hand side vector is generated using `randn`. Iterative refinement is terminated when

$$\frac{\|b - A\widehat{x}\|_\infty}{\|b\|_\infty + \|A\|_\infty \|\widehat{x}\|_\infty} \leq nu,$$

where u is the unit roundoff of the working precision. The inner GMRES iterations are terminated based on a backward error criterion for the preconditioned system with tolerance 10^{-2} and 10^{-4} for working precisions of single and double respectively, and a maximum of 20 iterative refinement steps are performed. In practice, we hope for convergence in a handful of iterative refinement steps, but we allow more in order to explore the speed of convergence for different problems and the two methods.

Table 6.1 shows the convergence for $\kappa_2(A) = 10^2$ and Table 6.2 shows the growth factors and condition numbers. Tables 6.3 and 6.4 give the corresponding information for $\kappa_2(A) = 10^7$. We need $\kappa_\infty(A)u$ sufficiently less than 1 to guarantee convergence of LU-IR and $\kappa_\infty(\widehat{U}^{-1}\widehat{L}^{-1}A)u$ sufficiently less than 1 to guarantee convergence of GMRES-IR [6].

Both LU-IR and GMRES-IR successfully solve the problems with $\kappa_2(A) = 10^2$. For $\kappa_2(A) = 10^7$, LU-IR fails to converge in several instances whereas GMRES-IR

[1] https://github.com/higham/chop

Table 6.3
Total number of iterative refinement steps in standard iterative refinement (LU-IR), and GMRES-IR for different precision combinations for $\kappa_2(A) = 10^7$. Numbers in parentheses denote the total number of GMRES iterations. "–" denotes that iterative refinement failed to converge.

	(H, S, D)		(H, D, D)		(S, D, D)	
n	LU-IR	GMRES-IR	LU-IR	GMRES-IR	LU-IR	GMRES-IR
500	–	2 (3)	–	3 (10)	12	3 (5)
750	4	1 (2)	–	3 (10)	–	3 (5)
1000	6	3 (7)	17	3 (13)	–	2 (4)
1250	16	2 (3)	–	4 (16)	16	3 (5)
1500	2	1 (2)	19	3 (12)	12	3 (5)
1750	2	1 (2)	–	3 (12)	19	3 (5)
2000	2	1 (2)	18	3 (12)	19	3 (5)
2250	3	1 (2)	–	2 (8)	–	3 (5)
2500	–	3 (9)	–	3 (13)	–	2 (4)

Table 6.4
Growth factors for partial pivoting and condition number of left preconditioned matrix for $\kappa_2(A) = 10^7$.

		$\kappa_\infty(\widehat{U}^{-1}\widehat{L}^{-1}A)$		
n	ρ_n	(H, S, D)	(H, D, D)	(S, D, D)
500	53.29	3.31e+10	3.30e+10	2.32e+03
750	103.80	3.60e+10	3.62e+10	3.33e+03
1000	90.27	8.96e+10	9.07e+10	4.84e+03
1250	102.03	1.58e+11	1.57e+11	1.72e+04
1500	178.48	1.24e+11	1.23e+11	1.51e+04
1750	186.22	2.10e+11	2.11e+11	3.31e+04
2000	321.61	2.49e+11	2.50e+11	1.48e+04
2250	349.27	3.85e+11	3.84e+11	3.28e+04
2500	188.25	3.95e+11	3.97e+11	1.34e+05

always converges within three iterative refinement steps, even though the condition guaranteeing convergence is not satisfied for (H, S, D). This behavior is consistent with the theory [6]. The important finding is that the inner GMRES solves converge in a modest number iterations, which shows that the large growth does not inhibit the ability of the computed low precision LU factors to act as effective preconditioners for GMRES.

We note that the convergence of the refinement could be enhanced by improving the preconditioner using a correction term based on an inexpensive estimate of the error in the factorization, as proposed by Higham and Mary [18].

7. Conclusions. The matrices (1.1) produce growth factors in LU factorization of order $n/\log n$ for any pivoting strategy, with high probability. Although these matrices are readily generated by the MATLAB `randsvd` function (albeit not with the default value of the `mode` parameter), this property appears to have gone unnoticed. The large growth stems from two properties. First, a random orthogonal matrix from the Haar distribution has relatively small elements with high probability for large n, which implies that the growth factor must be large for any pivoting strategy by a result from [17]. Second, if Q is an orthogonal matrix that gives large growth for any pivoting strategy then a rank-1 perturbation of 2-norm at most 1 to Q tends to preserve large growth. We have given two explanations for this second property, one based on a determinantal formula for the elements of U and one based on the Sherman–Morrison formula. The rank-1 perturbation allows the matrix to be given any 2-norm condition number, resulting in the class (1.1) of matrices with large growth and an arbitrary condition number.

With matrix dimensions in practical problems growing ever larger and low precision arithmetic becoming increasingly prevalent, growth of order $n/\log n$ in LU factorization can render the solution to a linear system unstable. Fortunately, iterative refinement is able to cure the instability, and we found that the performance of GMRES-IR, which uses the low precision computed LU factors as preconditioners for a GMRES-based solution to the correction equations, is unaffected by the lower quality computed LU factors.

Acknowledgments. We thank Tim Davis, Cleve Moler, Rob Schreiber, and Nick Trefethen for their comments on a draft manuscript.

Appendix A. Singular values of submatrix of an orthogonal matrix. Let σ_i denote the ith largest singular value.

LEMMA A.1. *Let B be an $(n-k)\times(n-k)$ submatrix of an orthogonal matrix $W\in\mathbb{R}^{n\times n}$, where $k<n/2$. Then B has at least $n-2k$ singular values equal to* 1 *and the remaining singular values are bounded above by* 1.

Proof. Assume without loss of generality that B is the leading principal submatrix of order $n-k$ of W. Then the CS decomposition of the orthogonal matrix W is (see, e.g., Golub and Van Loan [10, p. 85], Paige and Wei [28], or Stewart [33, p. 75])

$$\begin{bmatrix} V_1^T & 0 \\ 0 & U_2^T \end{bmatrix} \begin{bmatrix} B & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \begin{bmatrix} U_1 & 0 \\ 0 & V_2 \end{bmatrix} = \left[\begin{array}{cc|c} I_{n-2k} & 0 & 0 \\ 0 & C & S \\ \hline 0 & S & -C \end{array}\right]$$

for orthogonal $U_1, V_1 \in \mathbb{R}^{(n-k)\times(n-k)}$ and $U_2, V_2\in\mathbb{R}^{k\times k}$, where $C=\mathrm{diag}(c_1,\dots,c_k)$ and $S=\mathrm{diag}(s_1,\dots,s_k)$ are diagonal and nonnegative with $C^2+S^2=I$. It follows that the singular values of B are 1 repeated $n-2k$ times and $c_1,\dots,c_k$. □

REFERENCES

[1] Florent Benaych-Georges and Raj Rao Nadakuditi. The singular values and vectors of low rank perturbations of large rectangular random matrices. *J. Multivariate Anal.*., 111:120–135, 2012.

[2] Patrick Billingsley. *Probability and Measure*. Third edition, Wiley, New York, 1995. ISBN 0-471-00710-2.

[3] Garrett Birkhoff and Surender Gulati. Isotropic distributions of test matrices. *J. Appl. Math. Phys.* (*ZAMP*), 30:148–158, 1979.

[4] Ian Buck. World's fastest supercomputer triples its performance record. https://blogs.nvidia.com/blog/2019/06/17/hpc-ai-performance-record-summit/, June 2019. Accessed June 24, 2019.

[5] Erin Carson and Nicholas J. Higham. A new analysis of iterative refinement and its application to accurate solution of ill-conditioned sparse linear systems. *SIAM J. Sci. Comput.*, 39(6): A2834–A2856, 2017.

[6] Erin Carson and Nicholas J. Higham. Accelerating the solution of linear systems by iterative refinement in three precisions. *SIAM J. Sci. Comput.*, 40(2):A817–A847, 2018.

[7] Sourav Chatterjee. *Superconcentration and Related Topics*. Springer-Verlag, Cham, Switzerland, 2014. ix+156 pp.

[8] Jack J. Dongarra, Piotr Luszczek, and Yaohung M. Tsai. HPL-AI mixed-precision benchmark. https://icl.bitbucket.io/hpl-ai/.

[9] Leslie V. Foster. Gaussian elimination with partial pivoting can fail in practice. *SIAM J. Matrix Anal. Appl.*, 15(4):1354–1362, 1994.

[10] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. Fourth edition, Johns Hopkins University Press, Baltimore, MD, USA, 2013. xxi+756 pp. ISBN 978-1-4214-0794-4.

[11] Azzam Haidar, Ahmad Abdelfattah, Mawussi Zounon, Panruo Wu, Srikara Pranesh, Stanimire Tomov, and Jack Dongarra. The design of fast and energy-efficient linear solvers: On the potential of half-precision arithmetic and iterative refinement techniques. In *Computational Science—ICCS* 2018, Yong Shi, Haohuan Fu, Yingjie Tian, Valeria V. Krzhizhanovskaya, Michael Harold Lees, Jack Dongarra, and Peter M. A. Sloot, editors, Springer International Publishing, Cham, 2018, pages 586–600.

[12] Azzam Haidar, Harun Bayraktar, Stanimire Tomov, Jack Dongarra, and Nicholas J. Higham. Mixed-precision solution of linear systems using accelerator-based computing. Technical Report ICL-UT-20-05, Innovative Computing Laboratory, University of Tennessee, Knoxville, TN, USA, May 2020. 30 pp.

[13] Azzam Haidar, Stanimire Tomov, Jack Dongarra, and Nicholas J. Higham. Harnessing GPU tensor cores for fast FP16 arithmetic to speed up mixed-precision iterative refinement solvers. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, SC '18 (Dallas, TX), Piscataway, NJ, USA, 2018, pages 47:1–47:11. IEEE Press.

[14] H. V. Henderson and S. R. Searle. On deriving the inverse of a sum of matrices. *SIAM Rev.*, 23(1):53–60, 1981.

[15] Nicholas J. Higham. The Matrix Computation Toolbox. http://www.maths.manchester.ac.uk/~higham/mctoolbox.

[16] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Second edition, Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2002. xxx+680 pp. ISBN 0-89871-521-0.

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

[18] Nicholas J. Higham and Theo Mary. A new preconditioner that exploits low-rank approximations to factorization error. *SIAM J. Sci. Comput.*, 41(1):A59–A82, 2019.

[19] Nicholas J. Higham and Srikara Pranesh. Simulating low precision floating-point arithmetic. *SIAM J. Sci. Comput.*, 41(5):C585–C602, 2019.

[20] Nicholas J. Higham, Srikara Pranesh, and Mawussi Zounon. Squeezing a matrix into half precision, with an application to solving linear systems. *SIAM J. Sci. Comput.*, 41(4): A2536–A2551, 2019.

[21] *IEEE Standard for Floating-Point Arithmetic, IEEE Std* 754-2019 (*Revision of IEEE* 754-2008). The Institute of Electrical and Electronics Engineers, New York, USA, 2019. 82 pp. ISBN 978-1-5044-5924-2.

[22] Intel Corporation. BFLOAT16—hardware numerics definition, November 2018. White paper. Document number 338302-001US.

[23] M. Jankowski and H. Woźniakowski. Iterative refinement implies numerical stability. *BIT*, 17: 303–311, 1977.

[24] Tiefeng Jiang. Maxima of entries of Haar distributed matrices. *Prob. Theory Relat. Fields*, 131 (1):121–144, 2005.

[25] Tiefeng Jiang. How many entries of a typical orthogonal matrix can be approximated by independent normals? *Ann. Probab.*, 34:1497–1529, 2006.

[26] F. C. Leone, L. S. Nelson, and R. B. Nottingham. The folded normal distribution. *Technometrics*, 3(4):543–550, 1961.

[27] Francesco Mezzadri. How to generate random matrices from the classical compact groups. *Notices Amer. Math. Soc.*, 54(5):592–604, 2007.

[28] C. C. Paige and M. Wei. History and generality of the CS decomposition. *Linear Algebra Appl.*, 208/209:303–326, 1994.

[29] Parallel Computing Toolbox. The MathWorks, Inc., Natick, MA, USA. http://www.mathworks.co.uk/products/parallel-computing/.

[30] George Poole and Larry Neal. The rook's pivoting strategy. *J. Comput. Appl. Math.*, 123(1–2): 353–369, 2000.

[31] Robert D. Skeel. Iterative refinement implies numerical stability for Gaussian elimination. *Math. Comp.*, 35(151):817–832, 1980.

[32] G. W. Stewart. The efficient generation of random orthogonal matrices with an application to condition estimators. *SIAM J. Numer. Anal.*, 17(3):403–409, 1980.

[33] G. W. Stewart. *Matrix Algorithms. Volume II: Eigensystems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2001. xix+469 pp. ISBN 0-89871-503-2.

[34] Lloyd N. Trefethen and Robert S. Schreiber. Average-case stability of Gaussian elimination. *SIAM J. Matrix Anal. Appl.*, 11(3):335–360, 1990.

[35] J. H. Wilkinson. Error analysis of direct methods of matrix inversion. *J. Assoc. Comput. Mach.*, 8:281–330, 1961.

[36] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Oxford University Press, Oxford, UK, 1965. xviii+662 pp. ISBN 0-19-853403-5 (hardback), 0-19-853418-3 (paperback).

[37] Stephen J. Wright. A collection of problems for which Gaussian elimination with partial pivoting is unstable. *SIAM J. Sci. Statist. Comput.*, 14(1):231–238, 1993.