Homework: 1

Due: September 22.

1. Let A be an orthogonal matrix. Prove that $|\det(A)|=1$. Show that if B is also orthogonal and $\det(A)=-\det(B)$ then A+B is singular.
2. Trefethen 2.5, 3.2, 3.3
3. Prove that $\|xy^*\|_F=\|xy^*\|_2=\|x\|_2\|y\|_2$ for any x,y in C^n

1 Homework Solutions

18.335 - Fall 2004

1.1 **Let A be an orthogonal matrix. Prove that $|det(A)| = 1$. Show that if B is also orthogonal and $det(A) = -det(B)$, then $A + B$ is singular.**

$(\det A)^2 = \det A \det A = \det A \det A^T = \det AA^T = \det I = 1$

$A + B$ is singular iff $A^T(A + B) = I + A^T B$ is. $A^T B$ is orthogonal so all its eigenvalues are 1 or -1. Since their product is equal to $\det A^T B = -1$ then at least one of the eigenvalues of $A^T B$ must be -1. Let the corresponding vector be x. Then $(I + A^T B)x = x - x = 0$, so $I + A^T B$ is singular and so is $A + B$.

Second proof: $\det(A + B) = -\det(A^T)\det(A + B)\det(B^T) =$

$-\det(A^T AB^T + A^T BB^T) = -\det(A^T + B^T) = -\det(A + B)$, so $\det(A + B) = 0$.

1.2 Trefethen 2.5

(a) Let λ be an eigenvalue of S and v its corresponding eigenvector so that $Sv = \lambda v \Rightarrow v^* Sv = \lambda v^* v = \lambda \|v\|^2$. We also have $\overline{v^* Sv} = v^* S^* v = -v^* Sv$. This implies that $\bar{\lambda} = -\lambda \Rightarrow \lambda$ is imaginary.

(b) If $(I - S)v = 0$ for $v \neq 0$ then $Sv = v$ and this means that 1 is an eigenvalue of S, a contradiction to (a).

(c) We have:

$$\begin{aligned} Q^*Q &= \left[(I-S)^{-1}(I+S)\right]^* (I-S)^{-1}(I+S) \\ &= (I+S^*)(I-S^*)^{-1}(I-S)^{-1}(I+S) \\ &= (I-S)(I+S)^{-1}(I-S)^{-1}(I+S) \\ &= (I+S)^{-1}(I-S)(I-S)^{-1}(I+S) = I \end{aligned}$$

where we have used that if $AB = BA$ and B is invertible that $AB^{-1} = B^{-1}A$

1.3 Trefethen 3.2

We know that $\|A\| = \sup_{x \neq 0} \dfrac{\|Ax\|}{\|x\|}$. Choose an eigenvalue λ of A and let $x_\lambda \neq 0$ such that $Ax_\lambda = \lambda x_\lambda$. Then $\dfrac{\|Ax_\lambda\|}{\|x_\lambda\|} = \dfrac{\|\lambda x_\lambda\|}{\|x_\lambda\|} = \dfrac{|\lambda|\,\|x_\lambda\|}{\|x_\lambda\|} = \lambda$. Thus we have $\|A\| = \sup_{x \neq 0} \dfrac{\|Ax\|}{\|x\|} \geq |\lambda|$. So $\|A\| \geq |\lambda|$ and since this is true for any eigenvalue of A we get $\|A\| \geq \sup\{|\lambda|, \lambda \text{ eigenvalue of } A\} = \rho(A)$.

1.4 Trefethen 3.3

(a) By definition $\|x\|_\infty = \max_{1\le i\le m} |x_i| \le \sqrt{\sum_{j=1}^{m} \left|x_j^2\right|} = \|x\|_2$. Equality is achieved when we have a vector with only one non-zero component.

(b) Again, using the definition $\|x\|_2 = \sqrt{\sum_{j=1}^{m} \left|x_j^2\right|} \le \sqrt{m \max_{1\le i\le m} |x_i|} = \sqrt{m}\,\|x\|_\infty$. We have equality for a vector whose components are equal to each other.

(c) Denoting by r_j the j-th row of A we have $\|A\|_\infty = \max_{1\le j\le m} \|r_j\|_1$. For some vector $v \in \mathbb{C}^n$, $v^* = (1, ..., 1)/\sqrt{n}$ and using the 2-norm definition we get $\|A\|_2 = \sup_{|x|=1} \|Ax\|_2 \ge \|Av\|_2 = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{m} \|r_j\|_1^2}$. These yield $\|A\|_\infty = \max_{1\le j\le m} \|r_j\|_1 \le \sqrt{\sum_{j=1}^{m} \|r_j\|_1^2} \le \sqrt{n}\,\|A\|_2$. Equality is achieved for a matrix which is zero everywhere except along a row of ones.

(d) Using the notation from part (c), $\|A\|_2 = \frac{1}{\sqrt{n}} \sqrt{\sum_{j=1}^{m} \|r_j\|_1^2} \le \sqrt{\sum_{j=1}^{m} \|r_j\|_1^2} \le \sqrt{m} \max_{1\le j\le m} \|r_j\|_1 = \sqrt{m}\,\|A\|_\infty$. We get equality for a square matrix which is zero everywhere except along a column of ones.

1.5 Prove that $\|xy^*\|_F = \|xy^*\|_2 = \|x\|_2 \|y\|_2$ for any x and $y \in \mathbb{C}^n$.

$\|xy^*\|_F = \sqrt{\sum_{j=1}^{n} \sum_{i=1}^{n} |x_i \bar{y}_j|^2} = \sqrt{\sum_{i=1}^{n} |x_i|^2} \sqrt{\sum_{j=1}^{n} |\bar{y}_j|^2} = \|x\|_2 \|y\|_2$

$\|xy^*\|_2 = \sup_{z\in\,\mathbb{C}^n} \frac{\|xy^*z\|_2}{\|z\|_2} = \sup_{z\in\,\mathbb{C}^n} \frac{\|x\|_2 |y^*z|}{\|z\|_2}$. This ratio is maximized if $z//y$, so that $|y^*z| = \|y\|_2^2$, thus completing the proof.

Homework: 2

Due: September 29.

1. Count the number of floating point operations required to compute the QR decomposition of an m-by-n matrix using (a) Householder reflectors (b) Givens rotations.
2. Trefethen 5.4
3. If $A=R+uv^*$, where R is upper triangular matrix and u and v are (column) vectors, describe an algorithm to compute the QR decomposition of A in $O(n^2)$ time.
4. Given the SVD of A, compute the SVD of $(A^*A)^{-1}$, $(A^*A)^{-1}A^*$, $A(A^*A)^{-1}$, $A(A^*A)^{-1}A^*$ in terms of U, Sigma and V.

2 Homework Solutions

18.335 - Fall 2004

2.1 **Count the number of floating point operations required to compute the QR decomposition of an m-by-n matrix using (a) Householder reflectors (b) Givens rotations.**

(a) See Trefethen p. 74-75. Answer: $\sim 2mn^2 - \frac{2}{3}n^3$ flops.

(b) Following the same procedure as in part (a) we get the same 'volume', namely $\sim \frac{1}{2}mn^2 - \frac{1}{6}n^3$. The only difference we have here comes from the number of flops required for calculating the Givens matrix. This operation requires 6 flops (instead of 4 for the Householder reflectors) and hence in total we need $\sim 3mn^2 - n^3$ flops.

2.2 **Trefethen 5.4**

Let the SVD of $A = U\Sigma V^*$. Denote with v_i the columns of V , u_i the columns of U and σ_i the singular values of A. We want to find $x = (x_1; x_2)$ and λ such that:

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \lambda \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

This gives $A^*x_2 = \lambda x_1$ and $Ax_1 = \lambda x_2$. Multiplying the 1st equation with A and substitution of the 2nd equation gives $AA^*x_2 = \lambda^2 x_2$. From this we may conclude that x_2 is a left singular vector of A. The same can be done to see that x_1 is a right singular vector of A. From this the $2m$ eigenvectors are found to be:

$$x_\pm = \frac{1}{\sqrt{2}} \begin{pmatrix} v_i \\ \pm u_i \end{pmatrix}, \; i = 1...m$$

corresponding to the eigenvalues $\lambda = \pm\sigma_i$. Therefore we get the eigenvalue decomposition:

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} V & V \\ U & -U \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} V & V \\ U & -U \end{pmatrix}^{-1}$$

2.3 **If $A = R + uv^*$, where R is upper triangular matrix and u and v are (column) vectors, describe an algorithm to compute the QR decomposition of A in $\mathcal{O}(n^2)$ time.**

The matrix A is of the form

$$A = \begin{pmatrix} * & * & * & \cdots & * & * \\ u_2v_1 & * & & & & * \\ u_3v_1 & u_3v_2 & * & & & \vdots \\ u_4v_1 & u_4v_2 & u_4v_3 & * & & \vdots \\ \vdots & & & & \ddots & * \\ u_nv_1 & u_nv_2 & \dots & \cdots & u_nv_{n-1} & * \end{pmatrix}$$

We exploit the fact that the matrix uv^* is rank one. By applying a sequence of Givens rotations starting from the bottom row, we notice that the rotation that zeroes the entry $A_{k,1}$ also zeroes out all the entries $A_{k,2}$, $A_{k,3}$, ...$A_{k,2n-k-2}$. Thus the $n-1$ Givens rotations that kill the first column also kill all the entries below the first subdiagonal:

$$\begin{pmatrix} * & * & \cdots & * & * \\ \times & * & & & * \\ 0 & \times & * & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & \times & * \end{pmatrix}$$

Thus we need another $n-1$ Givens rotations to kill the first subdiagonal entries (shown with $\times$'s above). We have a total cost $2n-2$ rotations at no more than $6n$ operations per Givens rotation. Hence this algorithm requires $\mathcal{O}(n^2)$ flops.

2.4 **Given the SVD of A, compute the SVD of $(A^*A)^{-1}$, $(A^*A)^{-1}A^*$, $A(A^*A)^{-1}$, $A(A^*A)^{-1}A^*$ in terms of U, Σ and V.**

Answers:

- $(A^*A)^{-1} = V\Sigma^{-2}V^*$
- $(A^*A)^{-1}A^* = V\Sigma^{-1}U^*$
- $A(A^*A)^{-1} = U\Sigma^{-1}V^*$
- $A(A^*A)^{-1}A^* = UU^*$ (note that UU^* may not be equal to I, unless U is square in the reduced SVD)

Homework: 3

Due October 6.

1. Trefethen 10.1
2. Let B be an n-by-n upper bidiagonal matrix. Describe an algorithm for computing the condition number of B measured in the infinity norm in O(n) time.

3 Homework Solutions

18.335 - Fall 2004

3.1 Trefethen 10.1

(a) $H = I - 2vv^*$ where $\|v\| = 1$. If $v^*u = 0$ (u is perpendicular to v), then $Hu = u - 2vv^*u = u$. So 1 is an eigenvalue with multiplicity $n-1$ (there are $n-1$ linearly independent eigenvectors perpendicular to v). Also $Hv = v - 2vv^*v = v - 2v = -v$, so -1 is an eigenvalue of H. The geometric interpretation is given by the fact that reflection of v is $-v$, and reflection of any vector perpendicular to v is v itself.

(b) $\det H = \prod_{i=1}^{n} \lambda_i = (-1)\, 1^{n-1} = -1.$

(c) $H^*H = (I - 2vv^*)^* (I - 2vv^*) = I - 4vv^* + 4vv^*vv^* = I$. So the singular values are all 1's.

3.2 Let B be an $n \times n$ upper bidiagonal matrix. Describe an algorithm for computing the condition number of B measured in the infinity norm in $\mathcal{O}(n)$ time.

The condition number of B in the infinity norm is defined as:

$$\kappa_\infty (B) = \left\|B^{-1}\right\|_\infty \|B\|_\infty$$

We have to compute these two matrix norms separately. For $\|B\|_\infty$, the operation count is $\mathcal{O}(n)$ since only $n-1$ operations (corresponding to row sums for the first $n-1$ rows) are required. In order to calculate $\left\|B^{-1}\right\|_\infty$ we need to compute B^{-1} first. To do so, let $C = B^{-1}$. Performing the matrix multiplication $BC = I$, one can see that the inverse is an upper triangular matrix whose entries are given by:

$$c_{i,j} = \begin{cases} 0 & , i > j \\ \dfrac{1}{b_{i,i}} & , i = j \\ \dfrac{1}{b_{i,i}} \displaystyle\prod_{k=i}^{j-1} \left(-\frac{b_{k,k+1}}{b_{k+1,k+1}} \right) & , i < j \end{cases}$$

This enables us to compute the infinity norm of B^{-1} :

$$\left\|B^{-1}\right\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |c_{i,j}| = \max_i \underbrace{\frac{1}{|b_{i,i}|} \left(1 + \sum_{j=i+1}^{n} \left| \prod_{k=i}^{j-1} \left(-\frac{b_{k,k+1}}{b_{k+1,k+1}} \right) \right| \right)}_{P_i}$$

Normally, doing this directly would require $\mathcal{O}\left(n^2\right)$ flops. We can avoid this many flops by making some simplifications. Let:

$$d_k = \left| \frac{b_{k,k+1}}{b_{k+1,k+1}} \right|$$

and notice that the row sum, P_i can be written as:

$$\begin{aligned} P_i &= \frac{1}{|b_{i,i}|}\left(1 + \sum_{j=i+1}^{n}\prod_{k=i}^{j-1} d_k\right) = \frac{1}{|b_{i,i}|}\left(1 + d_i + \sum_{j=i+2}^{n}\prod_{k=i}^{j-1} d_k\right) \\ &= \frac{1}{|b_{i,i}|}\left[1 + d_i\left(1 + \sum_{j=i+2}^{n}\prod_{k=i+1}^{j-1} d_k\right)\right] = \frac{1}{|b_{i,i}|}\left(1 + d_i\left|b_{i+1,i+1}\right| P_{i+1}\right) \\ &= \frac{1}{|b_{i,i}|}\left(1 + \left|b_{i,i+1}\right| P_{i+1}\right) \end{aligned}$$

Thus knowing P_{i+1} we can calculate P_i in 5 operations. Hence we have to start from the n-th row and proceed backwards. So our algorithm to compute $\left\|B^{-1}\right\|_\infty$ is:

$$\left.\begin{array}{l} P_n = \dfrac{1}{|b_{n,n}|} = \left\|B^{-1}\right\|_\infty \end{array}\right\} 2 \text{ flops}$$

$$\left.\begin{array}{ll} \textbf{for} & \text{i=n-1 } \textbf{to } 1 \\ & P_i = \dfrac{1 + \left|b_{i,i+1}\right| P_{i+1}}{|b_{i,i}|} \\ & \left\|B^{-1}\right\|_\infty = \max\left(P_i, P_{i+1}\right) \\ \textbf{end} & \end{array}\right\} \mathcal{O}\left(n\right) \text{ flops}$$

Since both $\left\|B^{-1}\right\|_\infty$ and $\|B\|_\infty$ require $\mathcal{O}\left(n\right)$ flops, so does $\kappa_\infty\left(B\right)$.

Homework: 4

Due October 20. Solve at least 4 of the problems below.

1. Trefethen 11.1
2. Trefethen 13.3
3. Trefethen 13.4
4. Prove that (13.7) in Trefethen is valid for complex arithmetic (all four arithmetic operations) with |epsilon| now bounded by a modest multiple of epsilon_machine.
5. Prove that in IEEE binary floating point arithmetic sqrt(x^2) returns x exactly.
6. Let a and b be positive IEEE binary floating point numbers such that a < b < 2a. Prove that fl(b-a)=b-a exactly.

4 Homework Solutions

18.335 - Fall 2004

4.1 Trefethen 11.1

First note that any $x \in \mathbb{C}^m$ can be writtten as $x = x_R + x_R^{\perp}$ where $x_R \in \mathcal{R}(A)$, $x_R^{\perp} \in \mathcal{R}(A)^{\perp}$. Now since:

$$\left(x_R^{\perp}\right)^* \underbrace{Ay}_{\in \mathcal{R}(A)} = 0,\ \forall y \in \mathbb{C}^n \Longrightarrow y^* \left(A^* x_R^{\perp}\right) = 0,\ \forall y \in \mathbb{C}^n \Longrightarrow A^* x_R^{\perp} = 0$$

we have by definition:

$$\begin{aligned}
\left\|A^+\right\| &= \max_{\substack{z \in \mathbb{C}^m \\ z \neq 0}} \frac{\left\|\left(A^*A\right)^{-1}\left(A^* x_R + A^* x_R^{\perp}\right)\right\|}{\|x\|} \\
&\leq \max_x \frac{\left\|\left(A^*A\right)^{-1} A^* x_R\right\|}{\|x_R\|} = \max_w \frac{\left\|\left(A^*A\right)^{-1} A^* A w\right\|}{\|Aw\|} = \max_w \frac{\|w\|}{\|Aw\|} \\
&\leq \max_w \frac{\|w\|}{\left\|\sqrt{\|A_1 w\|^2 + \|A_2 w\|^2}\right\|} \leq \max_w \frac{\|w\|}{\|A_1 w\|} = \max_w \frac{\left\|A_1^{-1} w\right\|}{\|w\|} = \left\|A_1^{-1}\right\|
\end{aligned}$$

4.2 Prove that (13.7) in Trefethen is valid for complex arithmetic (all four arithmetic operations) with ε now bounded by a modest multiple of $\varepsilon_{\text{machine}}$.

For addition and subtraction we have

$$(a + ib) \pm (c + id) := (a \pm c) + i(b \pm d).$$

Let δ_i be small numbers bounded in absolute value by ϵ. We have

$$\begin{aligned}
fl((a + ib) \pm (c + id)) &= (a \pm c + i(b \pm d)) \left(1 + \frac{(a \pm c)\delta_1 + i(b \pm d)\delta_2}{(a \pm c) + i(b \pm d)}\right) \\
&= (a \pm c + i(b \pm c))(1 + \delta)
\end{aligned}$$

where

$$|\delta|^2 = \frac{(a \pm c)^2 \delta_1^2 + (b \pm d)\delta_2^2}{(a \pm c)^2 + (b \pm d)^2} \leq 2\epsilon^2$$

so $|\delta| \leq \sqrt{2}\epsilon$. For multiplication we have:

$$(a + ib)(c + id) := (ac - bd) + i(ad + bc)$$

For some $|\delta_i| \leq 2\epsilon$ we have:

$$\begin{aligned} fl((a+ib)(c+id)) &= (ac(1+\delta_1) - bd(1+\delta_2)) + i(ad(1+\delta_3) + bc(1+\delta_4)) \\ &= [(ac-bd) + i(ad+bc)] + [(ac\delta_1 - bd\delta_2) + i(ad\delta_3 + bc\delta_4)] \end{aligned}$$

We will use the fact that $|u+iv| \leq |u| + |v| \leq \sqrt{2}|u+iv|$ to write

$$fl((a+ib)(c+id)) = (a+ib)(c+id)(1+\beta)$$

where

$$\beta = \frac{(ac\delta_1 - bd\delta_2) + i(ad\delta_3 + bc\delta_4)}{(a+ib)(c+id)}$$

$$\begin{aligned} |\beta| &\leq \frac{(|ac\delta_1| + |bd\delta_2|) + (|ad\delta_3| + |bc\delta_4|)}{\frac{1}{2}(|a|+|b|)(|c|+|d|)} \\ &\leq 4\epsilon \frac{(|a|+|b|)(|c|+|d|)}{(|a|+|b|)(|c|+|d|)} = 4\epsilon \end{aligned}$$

This result does not guarantee high relative accuracy in the individual components of the product. For example if we take two numbers whose product is nearly real, the imaginary part will be the result of cancellation and so be small but probably not accurate. But the real part will be large, so the bound holds. Another way to look at it is that the true product lies in a little ball in the complex plane centered at the true product p and with radius $4\epsilon|p|$. If this ball interects the real (or imaginary) axis, then we can't even guarantee the sign of the real (or imaginary) part.

For division we have the following algorithm for computing $(a+bi)/(c+di)$:

- $\alpha = \max(|c|, |d|)$
- $c_1 = c/\alpha$
- $d_1 = d/\alpha$, ... therefore $c_1 + d_1 i = \frac{1}{\alpha}(c+di)$
- $s = \alpha(c_1^2 + d_1^2)$, ... same as $\alpha(c_1 + d_1 i)(c_1 - d_1 i)$
- $w = (a+bi)(c_1 - d_1 i)$
- $z = w/s$, ... same as $z = \frac{1}{s} \cdot w$
- return z

This clearly produces the right answer in exact arithmetic. These operations can be interpreted as complex multiplications and forming inverses of real numbers. Each of those operation preserves the relative accuracy and the overall error bound is a product of all $(1+\delta)$ terms from each complex multiply. Over all we get a relative error bounded by 22ϵ.

4.3 Prove that in IEEE binary floating point arithmetic $\sqrt{x}$ returns x exactly.

Recall that any IEEE number, x, can be written as

$$x = 2^a \left(1 + \frac{m}{2^{53}}\right), \text{ with } 0 \le m < 2^{53}$$

Note that here we assumed double precision, even though this is not necessary. Since we are not concerned with overflow or underflow, no limits were placed on a. Then we have

$$x^2 = 2^{2a} \left(1 + \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}}\right)$$

To show that $\text{fl}(\sqrt{x^2}) = x$ we need to verify that:

$$2^a \left(1 + \frac{m - \frac{1}{2}}{2^{53}}\right) \le \sqrt{2^{2a} \left(1 + \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}}\right)} < 2^a \left(1 + \frac{m + \frac{1}{2}}{2^{53}}\right) \tag{1}$$

In order to do that we have to distinguish 2 cases

- $\frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}} < 1$

  This implies that for

  $$-\frac{1}{2} + \frac{m^2}{2^{53}} \le y \le \frac{1}{2} + \frac{m^2}{2^{53}} \tag{2}$$

  we get that

  $$\text{fl}(x^2) = 2^{2a} \left(1 + \frac{2m + y}{2^{53}}\right)$$

  Thus (1) becomes

  $$1 + \frac{m - \frac{1}{2}}{2^{53}} \le \sqrt{1 + \frac{2m + y}{2^{53}}} < 1 + \frac{m + \frac{1}{2}}{2^{53}}$$
  $$\Rightarrow -1 + \frac{\left(m - \frac{1}{2}\right)^2}{2^{53}} \le y < 1 + \frac{\left(m + \frac{1}{2}\right)^2}{2^{53}}$$

  which is obiously true because of (2) .

- $1 < \frac{2m}{2^{53}} + \frac{1}{2^{53}} \frac{m^2}{2^{53}} < 3$

  In this case we have that for some $0 \le k < 2^{52}$

  $$1 + \frac{2k - \frac{1}{2}}{2^{53}} \le \frac{2m}{2^{53}} + \frac{m^2}{2^{106}} < 1 + \frac{2k + \frac{1}{2}}{2^{53}} \tag{3}$$

and therefore

$$\text{fl}(x^2) = 2^{2a}\left(1 + 1 + \frac{2k}{2^{53}}\right) = 2^{2a+1}\left(1 + \frac{k}{2^{53}}\right)$$

To show that $\text{fl}(\sqrt{x^2}) = x$ we need to verify from (1) that

$$1 + \frac{m - \frac{1}{2}}{2^{53}} \leq \sqrt{2\left(1 + \frac{k}{2^{53}}\right)} < 1 + \frac{m + \frac{1}{2}}{2^{53}}$$

This follows directly from (3) since by re-arranging terms in (3) we get:

$$\left(1 + \frac{m}{2^{53}}\right)^2 - \frac{1}{2^{54}} \leq 2\left(1 + \frac{k}{2^{53}}\right) < \left(1 + \frac{m}{2^{53}}\right)^2 + \frac{1}{2^{54}}$$

Combining these 2 cases we complete the proof.

4.4 Let a and b be positive IEEE binary floating point numbers such that $a < b < 2a$. Prove that $\text{fl}(b - a) = b - a$ exactly.

Proof: Assume $a = 1.a_1a_2...a_n \times 2^k, b = 1.b_1b_2...b_n \times 2^r$ $(a_i, b_i \in \{0, 1\})$. Also we may assume $k = 0$. $b \geq a$ implies $r \leq 0$ and $2a \geq b$ implies $r + 1 \geq 0$, so we have either $r = 0$ or $r = -1$.

For $r = 0$ we have $b - a = 1.b_1b_2...b_n - 1.a_1a_2...a_n = 0.c_1c_2...c_n$, which is an exact floating point number, since it has less than $n + 1$ fraction bits.

For $r = -1$ we have $b - a = 1.b_1b_2...b_n - 1.a_1a_2...a_n \times 2^{-1} = 1.b_1b_2...b_n0 - 0.1a_1a_2...a_n = c_{-1}.c_0c_1c_2...c_n$. Since $2a \geq b$ we have $c_{-1} = 0$. So the result has at most $n + 1$ fraction bits which is again an exact floating point number.

Homework: 5

Due October 27. Solve 4 of the following 5 problems:

1. Trefethen 20.1
2. Trefethen 21.6
3. Trefethen 22.1
4. Let A be symmetric and positive definite. Show that $|a_{ij}|^2 < a_{ii}\, a_{jj}$.
5. Let A and A^{-1} be given real n-by-n matrices. Let $B=A+xy^T$ be a rank-one perturbation of A. Find an $O(n^2)$ algorithm for computing B^{-1}. Hint: B^{-1} is a rank-one perturbation of A^{-1}.

5 Homework Solutions

18.335 - Fall 2004

5.1 Trefethen 20.1

$\Rightarrow$If A has an LU factorization, then all diagonal elements of U are not zero. Since $A = LU$ implies that $A_{1:k,1:k} = L_{1:k,1:k}U_{1:k,1:k}$ we get that $A_{1:k,1:k}$ is invertible.

$\Leftarrow$We prove by induction that $A_{1:k,1:k} = L_{1:k,1:k}U_{1:k,1:k}$ with

$$L_{1:k+1,1:k+1} = \begin{pmatrix} L_{1:k,1:k} & 0 \\ * \; * \; * & 1 \end{pmatrix} \text{ and } U_{1:k+1,1:k+1} = \begin{pmatrix} U_{1:k,1:k} & * \\ 0 & u_{k+1} \end{pmatrix}$$

with all the elements on the diagonal of $U_{1:k,1:k}$ are non-zero for any k.

Step 1 For $k = 1$ we have $A_{1:1,1:1} = L_{1:1,1:1}U_{1:1,1:1}$ with $L_{1:1,1:1} = 1$, $U_{1:1,1:1} = A_{1:1,1:1} \neq 0$.

Step 2 If that is true for $k \leq m$ we prove it for $m + 1$. Simply choose:

$$A_{1:m+1,1:m+1} = \underbrace{\begin{pmatrix} L_{1:m,1:m} & 0 \\ X_m & 1 \end{pmatrix}}_{L_{1:m+1,1:m+1}} \underbrace{\begin{pmatrix} U_{1:m,1:m} & Y_m \\ 0 & u_{m+1} \end{pmatrix}}_{U_{1:m+1,1:m+1}}$$

with

$$\begin{aligned} X_m &= [a_{m+1,1} \ldots a_{m+1,m}]U^{-1}_{1:m,1:m} \\ Y_m &= L^{-1}_{1:m,1:m} \begin{bmatrix} a_{1,m+1} \\ \vdots \\ a_{m,m+1} \end{bmatrix} \\ u_{m+1} &= -X_m Y_m \end{aligned}$$

Now we have $u_{m+1} \neq 0$ since $\det(A_{1:m+1,1:m+1}) = \det(U_{1:m,1:m})\, u_{m+1} \neq 0$. Now since $A = A_{1:n,1:n} = L_{1:n,1:n}U_{1:n,1:n}$ and $L_{1:n,1:n}$ is unit lower diagonal, $U_{1:n,1:n}$ is upper diagonal and we complete the proof.

5.2 Trefethen 21.6

Write

$$A = \begin{pmatrix} a_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

Proceed with the first step of Gaussian elimination:

$$\begin{pmatrix} a_{11} & A_{12} \\ 0 & A_{22} - \dfrac{A_{21}}{a_{11}} A_{12} \end{pmatrix}$$

Now for $A_{22} - \dfrac{A_{21}}{a_{11}} A_{12}$ we show that it has the property of strictly diagonally dominant matrices.

$$\sum_{j \neq k} \left| \left(A_{22} - \frac{A_{21}}{a_{11}} A_{12} \right)_{jk} \right| \leq \sum_{j \neq k} \left| (A_{22})_{jk} \right| + \sum_{j \neq k} \left| \frac{1}{a_{11}} (A_{21})_j (A_{12})_k \right|$$

A is strictly diagonally dominant, so we may write

$$\sum_{j \neq k} \left| (A_{22})_{jk} \right| < |(A_{22})_{kk}| - |(A_{12})_k| \text{ and } \sum_{j \neq k} \left| (A_{21})_j \right| < |a_{11}| - |(A_{21})_k|$$

so that in the end we get:

$$\begin{aligned} \sum_{j \neq k} \left| \left(A_{22} - \frac{A_{21}}{a_{11}} A_{12} \right)_{jk} \right| &< |(A_{22})_{kk}| - |(A_{12})_k| + \frac{|(A_{12})_k|}{|a_{11}|} (|a_{11}| - |(A_{21})_k|) \\ &< |(A_{22})_{kk}| - \frac{|(A_{12})_k| \, |(A_{21})_k|}{|a_{11}|} \leq \left| (A_{22})_{kk} - \frac{(A_{21})_k (A_{12})_k}{a_{11}} \right| \\ &\leq \left| \left(A_{22} - \frac{A_{21} A_{12}}{a_{11}} \right)_{kk} \right| \end{aligned}$$

Hence by induction if the property is true for any matrix of dimension $\leq m-1$ then it is true for any matrix A of dim $A = n$. This means that the submatrices that are created by successive steps of Gaussian elimination are also strictly diagonally dominant and hence we have no need for row swappings.

5.3 Trefethen 22.1

Apply 1 step of Gaussian elimination to A :

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mm} \end{pmatrix} \xrightarrow[\text{of GE}]{\text{1 Step}} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ 0 & a_{22}^{(1)} & \cdots & a_{2m}^{(1)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{m2}^{(1)} & \cdots & a_{mm}^{(1)} \end{pmatrix}$$

, where the entries $a_{ij}^{(1)} = a_{ij} - l_{ik} a_{kj}$. Since we used partial pivoting in our calculation, we must have $|l_{ik}| \leq 1$,

$$|\tilde{a}_{ij}| = |a_{ij} - l_{ik} a_{kj}| \leq |a_{ij}| + |l_{ik}| \, |a_{kj}| \leq |a_{ij}| + |a_{kj}| \leq 2 \max_{i,j} |a_{i,j}|$$

In order to form A we need $m-1$ such steps, so in the end we have:

$$|u_{ij}| \leq 2 \max_{i,j} \left| a_{i,j}^{(m-2)} \right| \leq 2 \max_{i,j} \left| a_{i,j}^{(m-3)} \right| \leq \ldots \leq 2 \max_{i,j} |a_{i,j}|$$

so that we obtain $|u_{ij}| \leq 2^{m-1} \max_{i,j} |a_{i,j}|$. Therefore

$$\rho = \frac{\max_{i,j} |u_{i,j}|}{\max_{i,j} |a_{i,j}|} \leq 2^{m-1}$$

5.4 **Let A be symmetric and positive definite. Show that $|a_{ij}|^2 < a_{ii}\, a_{jj}$.**

Since A is symmetric and positive definite, it has all a_{ii} positive and for any vector x we have $x^T Ax > 0$. Choose x such that $x_k = \delta_{jk} a_{jj} - \delta_{ik} a_{ij}$, where δ_{lm} is the Kronecker delta, meaning that all the entries of x are zero except the i-th and the j-th entries which equal to $-a_{ij}$ and a_{jj} respectively . Carrying out the calculation gives $x^T Ax = a_{ii}\left(a_{ii}a_{jj} - a_{ij}^2\right) > 0$ thus completing the proof.

5.5 **Let A and A^{-1} be given real n-by-n matrices. Let $B = A + xy^T$ be a rank-one perturbation of A. Find an $O(n^2)$ algorithm for computing B^{-1}. Hint: B^{-1} is a rank-one perturbation of A^{-1}.**

Since B^{-1} is a rank-one perturbation of A^{-1} we may write $B^{-1} = A^{-1} + uv^T$. Then

$$\begin{aligned} BB^{-1} &= \left(A + xy^T\right)\left(A^{-1} + uv^T\right) \\ I &= I + Auv^T + xy^T A^{-1} + xy^T uv^T \\ 0 &= Auv^T + xy^T A^{-1} + xy^T uv^T \end{aligned}$$

Choosing $u = A^{-1}x$, allows us to write:

$$\begin{aligned} 0 &= xv^T + xy^T A^{-1} + xy^T uv^T \\ 0 &= v^T + y^T A^{-1} + y^T uv^T \\ 0 &= v^T(1 + y^T u) + y^T A^{-1} \\ v^T &= -\frac{y^T A^{-1}}{1 + y^T A^{-1} x} \end{aligned}$$

Hence B^{-1} is given by:

$$B^{-1} = A^{-1} - \frac{A^{-1}xy^T A^{-1}}{1 + y^T A^{-1} x}$$

It is easy to see that the inverse can be computed in $O(n^2)$ operations

Homework: 6

Due November 3.

1. Let A be skew Hermitian, i.e. $A^{*}=-A$. Show that $(I-A)^{-1}(I+A)$ is unitary.
2. Trefethen 25.1

6 Homework Solutions

18.335 - Fall 2004

6.1 **Let A be skew Hermitian, i.e. $A^*= -A$. Show that $(I - A)^{-1}(I + A)$ is unitary.**

See solutions for the first Homework, problem 2.

6.2 **Trefethen 25.1**

(a) Let λ be an eigenvalue of A. Therefore $B = A - \lambda I$ is singular and hence

$$\text{rank}(A - \lambda I) \leq m - 1$$

The $m - 1 \times m$ submatrix $B_{2:m,1:m}$ is upper triangular whose diagonal entries are non-zero by our assumptions on A. Hence $B_{2:m,1:m}$ has $m - 1$ linearly independent columns which implies

$$\text{rank}(B_{2:m,1:m}) = m - 1$$

Therefore we must also have $\text{rank}(A-\lambda I) = m-1$, and hence the null space of B is spanned by one vector, a unique eigenvector of A correspoding to λ. Since A is Hermitian, which requires m linearly independent eigenvectors, all λ must be distinct.

(b) $\begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$

Homework: 7

Due November 17.

1. Compute the smallest eigenvalue of the 100-by-100 Hilbert matrix H=1/(i+j-1). (Hint: The Hilbert matrix is also Cauchy. The determinant of a Cauchy matrix $C(i,j)=1/(x_i+y_j)$ is
   $\det C = [\prod_{i<j}(x_j-x_i)(y_j-y_i)]/ [\prod_{i,j}(x_i+y_j)]$.
   Any submatrix of a Cauchy matrix is also Cauchy. You can use Cramer's rule in order to compute accurate formulas for H^{-1} and then compute its largest eigenvalue.
2. Trefethen 30.2

7 Homework Solutions

18.335 - Fall 2004

7.1 **Compute the smallest eigenvalue of the 100×100 Hilbert matrix $H_{ij} = 1/(i+j-1)$. (Hint: The Hilbert matrix is also Cauchy. The determinant of a Cauchy matrix $C(i,j) = 1/(x_i + y_j)$ is $\det C = \prod_{i<j}(x_j - x_i)(y_j - y_i) \Big/ \prod_{i,j}(x_i + y_j)$. Any submatrix of a Cauchy matrix is also Cauchy. You can use Cramer's rule in order to compute accurate formulas for H^{-1} and then compute its largest eigenvalue)**

We use Cramer's rule

$$H_{ij}^{-1} = (-1)^{i+j} \frac{\det(C_{ij})}{\det(H)}$$

together with the formula given for the determinant with $x_i = i$ and $y_j = j - 1$ to get H_{ij}^{-1} :

$$\begin{aligned} H_{ij}^{-1} &= (-1)^{i+j} \frac{\prod_{r<s} (x_s - x_r)(y_s - y_r)}{\prod_{r \neq i, s \neq j} (y_s + x_r)} \frac{\prod_{i,j} (x_i + y_j)}{\prod_{i<j} (x_j - x_i)(y_j - y_i)} \\ &= \cdots \\ &= (-1)^{i+j} (i+j-1) \binom{n+i-1}{n-j} \binom{n+j-1}{n-i} \binom{i+j-2}{i-1}^2 \end{aligned}$$

Having computed the coefficients of H^{-1} we may use any iterative scheme to estimate the largest eigenvalue which can be inverted to obtain the smallest eigenvale of H. Alternatively one could use a simple matlab command:

$$\lambda_{\min}(H) = \frac{1}{\lambda_{\max}(H^{-1})} = \texttt{1/max(eig(invhilb(100)))=5.779700862834800e-151}$$

7.2 **Trefethen 30.2**

- Jacobi algorithm
  Calculation of J : $\mathcal{O}(1)$ flops
  $J^T A$ alters 2 rows of A only $\Rightarrow$ 3 ops $\times$ $2m$ elements $\Rightarrow \mathcal{O}(6m)$ flops
  $\left(J^T A\right) J$ alters 2 columns $\Rightarrow \mathcal{O}(6m)$ flops.
  In total we need $\mathcal{O}(12m)$ flops for a single step of Jacobi algorithm (Half in case A is symmetric)
  In a single sweep we need $\sim m^2 \mathcal{O}(12m)/2 = \mathcal{O}\left(6m^3\right)$ flops (not counting convergence iterations).

- QR
  Requires $\mathcal{O}\left(4m^3/3\right)$, a much better algorithm!

Homework: 8

Due November 24. The writeups should be brief (preferably < 1 page) and include 1) the answer 2) the method you used and why you think it works 3) one paragraph on how you implemented the method. You need to compute the answer with at least 9 correct decimal digits.

1. Compute the smallest eigenvalue of the 100-by-100 matrix H=1/(i+j).
   Solution: See the solution to Problem 7.1.
2. The infinite matrix A has entries A(1,1)=1, A(1,2)=1/2, A(2,1)=1/3, A(1,3)=1/4, A(2,2)=1/5, A(3,1)=1/6, etc. Compute $||A||_2$.
   Solution: See this website, and problem 3 (and its solution) there.

Homework 8 Solutions:

1. Solution: See the solution to Problem 7.1
2. See problem 3 (and its solution) at http://mathworld.wolfram.com/Hundred-DollarHundred-DigitChallengeProblems.html

Homework: 9

Due December 1.

1. A is a square matrix of size 19,000. All entries of A are zero except for the primes 2,3,5,7,...,212369 along the main diagonal and the number 1 in all the positions a_{ij} with |i-j|=1,2,4,8,...,16384. Compute the (1,1) entry of A^{-1}.
   Solution: See this website, and problem 7 (and its solution) there.
2. The 200-by-200 (diagonally dominant) matrix A has offdiagonal entries -1/i-1/j and row sums s(i)=sum(A(i,:))=2^{2i}. Compute the smallest in magnitude eigenvalue of A. (Hint: The diagonal entries A(i,i) can then be computed as sum of positive (why?) numbers, thus to high relative accuracy. Abandoning the row sums s(i), however, robs you of any chance of computing the smallest eigenvalue accurately). You will encounter no (true) subtractions when running Gaussian elimination, obtaining the LU decomposition of A to high relative accuracy componentwise. Using the LU factors to compute the inverse of A one column at a time will not result in any subtractions either, yielding a positive A^{-1}.)
   Solution: mmatinverse.pdf, hw6.m, InverseMM.m.
   Answer: 1.207993236710136e+002

Homework 9 Solution:

1. See problem 7 (and its solution) at http://mathworld.wolfram.com/Hundred-DollarHundred-DigitChallengeProblems.html
2. See mmatinverse.pdf, hw6.m, InverseMM.m. Answer: 1.207993236710136e+002

Homework: 10

Due December 8. Compute the smallest eigenvalue of the 200-by-200 Pascal matrix. Details on the Pascal matrix can be obtained here.

Homework 10 Solution:

While there are many approaches to solve this problem, one way would be to compute the largest eigenvalue of the inverse. The inverse (which has checkerboard sign pattern) can (once again) be computed without performing any subtractions if one takes the correct approach. Instead of eliminating the matrix in the typical Gaussian elimination fashion, try to eliminate it by using only ADJACENT rows and columns. This process is called Neville elimination. Once you eliminate the first row and first column, you will see that the Schur complement is also a Pascal matrix of one size less. In matrix form this elimination can be written as

$L * P_n * L^T = P'_{n-1}$

where L is lower bidiagonal matrix with ones on the main diagonal and -1 on the first subdiagonal and P'_{n-1} is an n-by-n matrix with zeros in the first row and column, except for the (1,1) entry (which equals one), and the matrix P_{n-1} in the lower right hand corner. You can now observe (no need to prove) that if you have $(P_{n-1})^{-1}$ you can compute $(P_n)^{-1}$ using the above equality without performing any subtractions.

18.335 Practice Midterm

1. (5 points) Let A be real symmetric and positive semidefinite, i.e. $x^T Ax \geq 0$ for all $x \neq 0$. Show that if the diagonal of A is zero, then A is zero.

2. (5 points) Show that if
$$Y = \begin{bmatrix} I & Z \\ 0 & I \end{bmatrix}$$
then $\kappa_F(Y) = 2n + \|Z\|_F^2$.

3. Let
$$T = \begin{bmatrix} a_1 & b_1 & & \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{n-1} \\ & & c_{n-1} & a_n \end{bmatrix}$$
be a real, n-by-n, nonsymmetric tridiagonal matrix where $c_i b_i > 0$ for all $1 \leq i \leq n-1$. Show that the eigenvalues of T are real (5 points) and distinct (5 points).

   Hint: Find a diagonal matrix D such that $C = DTD^{-1}$ is symmetric. Then argue about the rank of $C - \lambda I$.

4. (5 points) Let A be symmetric positive definite matrix with Cholesky factor C, i.e. $A = C^T C$. Show that $\|A\|_2 = \|C\|_2^2$.

5. (5 points) If A and B are real symmetric positive definite matrices then decide whether the following are true, justifying your results:

   - $A + B$ is symmetric positive definite.
   - $A \cdot B$ is symmetric positive definite.

6. (5 points) Prove that $\det(I + xy^T) = 1 + x^T y$.

18.335 Midterm. November 3, 2004

Name:

Problem 1	
Problem 2	
Problem 3	
Problem 4	
Problem 5	
Problem 6	
Total	

In all problems, all matrices are real and square and all vectors are real.

1. (5 points) Assume (do not prove here)

$$\|x\|_\infty \le \|x\|_2 \le \sqrt{n}\|x\|_\infty, \text{ for all } x \in \mathbf{R}^n.$$

Show that for any matrix A

$$\|A\|_\infty \le \sqrt{n}\|A\|_2 \le n\|A\|_\infty.$$

2. (5 points) Let A be symmetric positive definite matrix with Cholesky factor C, i.e. $A = C^T C$. Show that $\|A\|_2 = \|C\|_2^2$ and that $\kappa_2(A) = (\kappa_2(C))^2$, where $\kappa_2(X)$ is the condition number of the matrix X measured in the two-norm.

3. A matrix A is called strictly column diagonally dominant if $|a_{ii}| > \sum_{j \ne i} |a_{ji}|$ for all i.

   (a) (5 points) Show that such an A is nonsingular.

   (b) (5 points) Show that no pivoting is needed when computing $A = LU$. In other words, if we did do partial pivoting to compute $PA = LU$, P a permutation matrix, then we would get $P = I$. Hint: Show that after one step of Gaussian elimination, the bottom right $n-1$ by $n-1$ submatrix is also strictly column diagonally dominant.

4. (5 points) Let A be skew Hermitian, i.e. $A^* = -A$. Prove that the eigenvalues of A are purely imaginary and that $I - A$ is nonsingular.

5. (5 points) Let $\|\cdot\|$ be an operator norm. Prove that if $\|A\| < 1$ then $I - A$ is invertible.

6. (5 points) Let $x = (1, 2, 3, 4, 5, 6, 7)^T$ and $y = (-7.5, 2, -4, -4, 2, 3, 0.5)^T$. In double precision IEEE binary floating point arithmetic, true or false:

$$\text{fl}(x^T y) = 0?$$

Explain.

Smallest Eigenvalue of M-Matrices

- Def: M-Matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}; \qquad \begin{array}{l} a_{ii} \geq 0 \\ a_{ij} \leq 0, \ i \neq j \\ \text{Row Sums } s_i = \sum_{j=1}^{n} a_{ij} \geq 0 \end{array}$$

- Given: Row sums s_i and off diagonals $a_{ij}, i \neq j$.
- Diagonal elements computable accurately, sum of positives

$$a_{ii} = s_i - \sum_{j \neq i} a_{ij}$$

GE on Weakly Diagonally Dominant M-Matrices

- Pivoting, if needed, is diagonal, preserves structure
- One step of GE:
  - Off diagonals: $a_{ij} = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}}$
  - Row sums: $s_i = s_i - \frac{a_{ik}}{a_{kk}} s_k$
- Everything is preserved in Schur complementation
  - Weak diagonal dominance
  - M-matrix structure
  - High relative accuracy in a_{ij} and s_i
- Yields Cholesky factors

Getting the inverse

- Again no subtractions in solving

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} & c_{14} \\ & c_{22} & c_{23} & c_{24} \\ & & c_{33} & c_{44} \\ & & & c_{44} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}$$

- Think of b as e_i or > 0 in general.

$$\begin{aligned} x_4 &= b_4/c_{44} \\ x_3 &= (b_3 - c_{34}x_4)/c_{33} \\ x_2 &= (b_2 - c_{24}x_4 - c_{23}x_3)/c_{22} \\ x_1 &= (b_1 - c_{14}x_4 - c_{13}x_3c_{12}x_2)/c_{11} \end{aligned}$$

- Solving with C^T analogous $\Rightarrow A^{-1}$ – positive.
- Accurate (Positive) Inverse = Accurate smallest eigenvalue (even in the nonsymmetric case)