Foundations of Computational Math I Exam 1
Take-home Exam
Open Notes, Textbook, Homework Solutions Only
Calculators Allowed
Tuesday 19 October, 2010

Question	Points Possible	Points Awarded
1. Basics	15	
2. Bases and Orthogonality	20	
3. Factorization Complexity	30	
4. Backward Stability	30	
5. Conditioning and Backward Error	25	
Total Points	120	

Name:________________________________

Alias: ________________________________

to be used when posting anonymous grade list.

Problem 1

(15 points)

Each question below has a brief answer and justification.

1.a. **(5 points)** Explain the idea of a "hidden bit" in a floating point system with base $\beta = 2$ and the benefit achieved by using it.

1.b. **(5 points)** Can the idea of a "hidden bit" be usefully generalized to a floating point system with base $\beta \neq 2$?

1.c. **(5 points)** Suppose you have a problem whose condition number is $\kappa \approx 10^5$. Given that you want at least 2 digits of accuracy in the solution how many decimal digits would you recommend be used in the floating point system used to solve the problem?

Solution:

The hidden bit is taken to be a 1 for normalized floating point numbers in an FP system with $\beta = 2$ and 0 for subnormal floating point numbers. The benefit is that it allows t bits of precision while actually storing only $t - 1$.

It does not work with $\beta \neq 2$ since in that case there is more than one value for the first digit in a normalized FP number, i.e., $0, 1, \dots, \beta - 1$.

The relative forward error is related to the uncertainty in the problem data by the condition number, i.e.,

$$\frac{|f(d+e) - f(d)|}{|f(d)|} \leq \kappa \frac{|e|}{|d|}$$

For the forward error to be smaller than 10^{-2} with $\kappa = 10^5$ we need the uncertainty in the data to be 10^{-7}. So we need to compute with at least 7 digits.

Problem 2

(20 points)

Consider the vector space $\mathbb{R}^3$ and the subspace $\mathcal{S}$ of dimension 1 given by

$$\mathcal{S} = span[v_1], \qquad v_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

2.a. Determine a basis $\{v_2, \;\; v_3\}$ of the subspace $\mathcal{S}^\perp$ where the vectors v_2 and v_3 are **not** orthogonal.

2.b. Derive from the basis $\{v_2, \;\; v_3\}$ a second basis $\{q_2, \;\; q_3\}$ of the subspace $\mathcal{S}^\perp$ where the vectors q_2 and q_3 are orthonormal vectors.

Solution:

By definition $\mathcal{S}^\perp$ is a subspace with dimension 2 comprising all vectors $w \in \mathbb{R}^3$ such that $w^T v_1 = 0$. Given the structure of v_1 any vector that has a single 1 element and a single -1 element is in $\mathcal{S}^\perp$. So we can take, for example,

$$v_2 = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} \quad \text{and} \quad v_3 = \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix}$$

Clearly, v_2 and v_3 are linearly independent due to nonzero structure so they are a basis for $\mathcal{S}^\perp$. Also note that

$$v_2^T v_3 = 1$$

so they are not orthogonal as desired.

To derive $\{q_2, \;\; q_3\}$ from $\{v_2, \;\; v_3\}$ where $q_2^T q_3 = 0$, $q_2^T q_2 = q_3^T q_3 = 1$ we must use the results of the homework problem that considered the least squares approximation of one vector, x, by another, y, and produced an orthogonal decomposition of x, i.e.,

$$min_\alpha \|x - \alpha y\|_2 x = z + \alpha_{min} y, \quad \alpha_{min} = y^T x / y^T y$$

Applying this result yields

$$z = v_3 - \frac{v_2^T v_3}{v_2^T v_2} v_2 = \begin{pmatrix} 1/2 \\ 1/2 \\ -1 \end{pmatrix}$$

$$z^T v_2 = 0$$

So we can normalize to get orthonormal vectors

$$q_2 = \frac{1}{\|v_2\|_2} v_2 = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \\ 0 \end{pmatrix}$$

$$q_3 = \frac{1}{\|z\|_2} z = \frac{1}{\sqrt{3/2}} \begin{pmatrix} 1/2 \\ 1/2 \\ -1 \end{pmatrix} = \begin{pmatrix} 1/\sqrt{6} \\ 1/\sqrt{6} \\ -\sqrt{2/3} \end{pmatrix}$$

We then have the desired results

$$q_2^T q_3 = 0, \quad q_2^T v_1 = 0, \quad q_3^T v_1 = 0$$
$$q_2^T q_2 = q_3^T q_3 = 1$$

This is a simple example of a more general procedure called the Gram-Schmidt method to produce an orthonormal basis from a basis that is not orthonormal.

Note that you could also use the technique based on Householder reflectors presented for solving least squares problems to produce and orthogonal basis.

Problem 3

(30 points)

3.a

(15 points)

Consider $A \in \mathbb{R}^{n \times n}$ whose nonzero elements are restricted to the main diagonal, the strict upper triangular part, and the first subdiagonal. For $n = 8$ the locations that must be zero are indicated and the positions that may be nonzero are indicated by α_{ij}:

$$A = \begin{pmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} & \alpha_{14} & \alpha_{15} & \alpha_{16} & \alpha_{17} & \alpha_{18} \\ \alpha_{21} & \alpha_{22} & \alpha_{23} & \alpha_{24} & \alpha_{25} & \alpha_{26} & \alpha_{27} & \alpha_{28} \\ 0 & \alpha_{32} & \alpha_{33} & \alpha_{34} & \alpha_{35} & \alpha_{36} & \alpha_{37} & \alpha_{38} \\ 0 & 0 & \alpha_{43} & \alpha_{44} & \alpha_{45} & \alpha_{46} & \alpha_{47} & \alpha_{48} \\ 0 & 0 & 0 & \alpha_{54} & \alpha_{55} & \alpha_{56} & \alpha_{57} & \alpha_{58} \\ 0 & 0 & 0 & 0 & \alpha_{65} & \alpha_{66} & \alpha_{67} & \alpha_{68} \\ 0 & 0 & 0 & 0 & 0 & \alpha_{76} & \alpha_{77} & \alpha_{78} \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{87} & \alpha_{88} \end{pmatrix}$$

(i) **(5 points)** Suppose the subdiagonal elements $\alpha_{i+1,i} \neq 0$ (this is called an unreduced Hessenberg matrix). Determine a necessary and sufficient condition for A to be nonsingular.

(ii) **(10 points)** Describe an efficient algorithm to solve $Ax = b$ via factorization and determine the order computational complexity, i.e., give k in $O(n^k)$. Your solution should include a description of how you exploit the structure of the matrix and how it influences the structure of your factors.

Solution:

Due to the fact that the Hessenberg matrix is unreduced we have a nonzero element in each of the columns 1 to $n-1$ that corresponds to a zero element in all the previous columns. Therefore, columns 1 to $n-1$ are linearly independent. The rank of A is therefore at least $n-1$. It will have rank n if and only if the last column is linearly independent from the first $n-1$ columns, i.e., if it is not in *span* $[Ae_1, \dots, Ae_{n-1}]$. Similarly, rows 2 to n are linearly independent and A will have rank n if and only if the first row is not in *span* $\left[e_2^T A, \dots, e_n^T A\right]$.

Another way to see this is to permute the rows of A by

$$P = \begin{pmatrix} e_2^T \\ e_3^T \\ \vdots \\ e_n^T \\ e_1 \end{pmatrix}$$

This yields a matrix with the form

$$PA = \begin{pmatrix} U & v \\ r^T & \mu \end{pmatrix}$$

Since A is unreduced the diagonal elements of U are nonzero and therefore there are $n-1$ linearly independent rows. The last row's relationship with the span of rows $1 dotson-1$ determines the singularity or nonsingularity of A.

Note that the singularity or nonsingularity of the matrix

$$M = \begin{pmatrix} \alpha_{77} & \alpha_{78} \\ \alpha_{87} & \alpha_{88} \end{pmatrix}$$

does not determine the nonsingularity or singularity of A. To see this consider the special case of an unreduced Hessenberg matrix with $n = 8$:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{18} \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & \alpha_{28} \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \alpha_{38} \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \alpha_{48} \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & \alpha_{58} \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \alpha_{68} \\ 0 & 0 & 0 & 0 & 0 & 1 & \alpha_{77} & \alpha_{78} \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \alpha_{88} \end{pmatrix}$$

Taking the singular matrix

$$M = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} \alpha_{18} \\ \alpha_{28} \\ \alpha_{38} \\ \alpha_{48} \\ \alpha_{58} \\ \alpha_{68} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

yields

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

which is nonsingular since the columns are clearly linearly independent.

Taking the nonsingular matrix

$$M = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \text{and} \quad \begin{pmatrix} \alpha_{18} \\ \alpha_{28} \\ \alpha_{38} \\ \alpha_{48} \\ \alpha_{58} \\ \alpha_{68} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

yields

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{pmatrix}$$

which is singular since columns 6 and 4 are identical (and the first row is all 0s).

A factorization algorithm is easily constructed to eliminate the subdiagonal. This can be an LU factorization or a QR using methods we have discussed. Pivoting in the LU factorization need only consider rows i and $i+1$ during the i-th step. The Gauss transform M_i^{-1} has a single off-diagonal element in position $(i+1, i)$ so L is has 1's on the diagonal and all other potentially nozero elements are on the first subdiagonal. The complexity is dominated by the term

$$2\sum_{i=1}^{n-1}(n-i) = n(n-1) = n^2 + O(n)$$

that corresponds to the scaling of row $i+1$ and its addition to row i. The triangular solves with L and U are $O(n)$ and $O(n^2)$ resepectively so the total is $O(n^2)$.

An alternative LU is to permute A by the P defined above and then eliminate elements of r^T. The first element is eliminated by combining row n with row 1, the second is eliminated by combining the updated row n with row 2, etc. This is $n-1$ steps where the i-th step involves $2(n-i+1)$ operations thereby giving a complexity of $n^2 + O(n)$ as above.

A QR factorization is easily formed by by applying a 2×2 reflector (or plane rotation) to rows i and $i+1$ on the i-th step. No pivoting is required and the complexity is dominated by the term

$$6\sum_{i=1}^{n-1}(n-i) = 3n(n-1) = 3n^2 + O(n)$$

that corresponds to applying a 2×2 orthogonal matrix to rows i and $i+1$.

3.b

(15 points)

Consider $S \in \mathbb{R}^{n\times n}$ whose nonzero elements have the following pattern for $n = 8$:

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \mu_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \delta_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \delta_3 & 0 & 0 & 0 & 1 \end{pmatrix}$$

(i) **(5 points)** Determine a necessary and sufficient condition for S to be nonsingular.

(ii) **(10 points)** Describe an efficient algorithm to solve $Sx = b$ via factorization and determine the order computational complexity, i.e., give k in $O(n^k)$. Your solution should include a description of how you exploit the structure of the matrix and how it influences the structure of your factors.

Solution:

Consider $P^T SP$ where

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \mu_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \delta_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \delta_3 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$P = \begin{pmatrix} e_1 & e_2 & e_3 & e_6 & e_7 & e_8 & e_4 & e_5 \end{pmatrix}$$

$$\tilde{S} = P^T SP = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & \mu_2 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & \mu_3 \\ 0 & 0 & 0 & 1 & 0 & 0 & \delta_1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & \delta_2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & \delta_3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta \\ 0 & 0 & 0 & 0 & 0 & 0 & \gamma & \delta \end{pmatrix}$$

$\tilde{S} = P^T SP$ is a block upper triangular matrix with 1 in $n-2$ of the diagonal elements and a trailing 2×2 diagonal block

$$M = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$$

Therefore,$\tilde{S}$ is nonsingular if and only if M is nonsingular. The same follows for S.

The factorization algorithm to solve $Sx = b$ is therefore simply factoring M, with pivoting if necessary, and solving the permuted system $\tilde{S}P^T x = P^T b$. This is clearly $O(n)$. In particular we have the factorization for $n = 8$ that clearly generalizes to any n:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & \mu_2 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & \mu_3 \\
0 & 0 & 0 & 1 & 0 & 0 & \delta_1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & \delta_2 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & \delta_3 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & \gamma & \delta
\end{pmatrix} = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \gamma/\alpha & 1
\end{pmatrix} \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_1 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & \mu_2 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & \mu_3 \\
0 & 0 & 0 & 1 & 0 & 0 & \delta_1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & \delta_2 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & \delta_3 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & \alpha & \beta \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \omega
\end{pmatrix}$$

The 2×2 forward and backward solve is $O(1)$ in complexity. Computing the remaining $n-2$ elements of $P^T x$ requires a multiplication and addition for each giving a complexity of

$$2n + O(1)$$

for the entire solution procedure.

Note that using LU factorization on the original form of S causes fill-in to appear and the complexity goes to $O(n^2)$. This can be seen after one step (assume pivoting is not needed for stability) that eliminates $\gamma, \delta_1, \ldots, \delta_3$.

$$S = \begin{pmatrix} 1 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \mu_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & \gamma & \delta & 0 & 0 & 0 \\ 0 & 0 & 0 & \delta_1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \delta_2 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \delta_3 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$M_4^{-1}S = \begin{pmatrix} 1 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \mu_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & \omega_3 & 0 & 0 & 1 \end{pmatrix}$$

On the next step, the ω_i must be eliminated but this does not cause fill-in.

$$M_5^{-1}M_4^{-1}S = \begin{pmatrix} 1 & 0 & 0 & 0 & \mu_1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & \mu_2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \mu_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & \alpha & \beta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \delta' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

so the algorithm remains $O(n)$ as desired but slightly more work is performed for the factorization and solution.

Applying Householder reflectors to S starting with eliminating column 4 below the diagonal causes the trailing part of the matrix to fill in completely in general therefore requiring the subsequent reduction of a dense matrix which yields and $O(n^3)$ computation. This can be seen by noting that a relector designed to perform

$$H \begin{pmatrix} \alpha \\ \gamma \\ \delta_1 \\ \delta_2 \\ \delta_3 \end{pmatrix} = \begin{pmatrix} \alpha' \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

has the form

$$H = I - 2uu^T$$

where u is a dense vector in general. Therefore

$$H \begin{pmatrix} \beta & 0 & 0 & 0 \\ \delta' & 0 & 0 & 0 \\ \omega_1 & 1 & 0 & 0 \\ \omega_2 & 0 & 1 & 0 \\ \omega_3 & 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \\ * & * & * & * \end{pmatrix}$$

Problem 4

(25 points)

4.a

(10 points)

Let $x \in \mathbb{R}^n$, and $y \in \mathbb{R}^n$ be two vectors with

$$x = \begin{pmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{pmatrix}, \quad y = \begin{pmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{pmatrix}.$$

$$|\xi_i| \geq 1 \quad |\eta_i| \geq 1$$

Consider the evaluation of the two inner products

$$\mu = x^T x$$

$$\gamma = x^T y$$

Which of the two inner products would you expect to be less sensitive to the perturbations caused by the finite precision of IEEE floating point arithmetic?

Solution:

The computation of $x^T x$ is expected to be the more reliable since it is the sum of positive numbers whose condition number is therefore 1. $x^T y$ may have terms of opposite signs and therefore will not be perfectly conditioned and could be ill-conditioned. Therefore $x^T x$ in general would be expected to be less sensitive.

4.b

(15 points)

Use the notation from the first part of the problem and assume the following lemma is true.

Lemma 4.1. *The computed inner product satisfies the following error bounds:*

$$fl(x^T y) = x^T(y + \Delta y) = (x + \Delta x)^T y, \quad |\Delta x| \leq \omega_n |x|, \quad |\Delta y| \leq \omega_n |y|$$

$$|x^T y - fl(x^T y)| \leq \omega_n \sum_{i=1}^{n} |\xi_i \eta_i| = \omega_n |x|^T |y|$$

$$\omega_n = \frac{nu}{1 - nu}$$

where u is unit roundoff.

Prove the following backward error lemma:

Lemma 4.2. *If $A \in \mathbb{R}^{n \times n}$ and $x \in \mathbb{R}^n$ then the matrix vector product $\hat{y} \in \mathbb{R}^n$ computed in finite precision satisfies*

$$\hat{y} = (A + \Delta A)x, \quad |\Delta A| \leq \omega_n |A|, \quad \omega_n = \frac{nu}{1 - nu}$$

Solution:

The desired lemma is a backward error result. The first line of the first lemma is the assumed backward error result for an inner product. Since each element of the computed matrix vector product is an inner product (even if it is computed as an accumulated set of vector triads) we can choose to cast the rounding error back to perturbations on the rows of A and keep x the same for all computed elements. This yields the desired matrix vector result directly.

Problem 5

(30 points)

5.a

(15 points)

Consider the matrix

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix}$$

Determine or bound the condition number for inversion, i.e., solving a system of linear equations, for the matrix A.

Solution:

We have an LU factorization for A:

$$A = \begin{pmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & -1 & 1 & 0 \\ -1 & -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 4 \\ 0 & 0 & 0 & 8 \end{pmatrix} = LU$$

Due to the structure of L, L^{-1} follows from its first column which solves $Lx = e_1$ and U^{-1} is easily determined by solving $Ux_i = e_i$:

$$L^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 1 & 1 & 0 \\ 4 & 2 & 1 & 1 \end{pmatrix}$$

$$U^{-1} = \begin{pmatrix} 1 & 0 & 0 & -1/8 \\ 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/2 \\ 0 & 0 & 0 & 1/8 \end{pmatrix}$$

We have

$$\kappa = \|A\|\|A^{-1}\| = \|A\|\|U^{-1}L^{-1}\| \le \|A\|\|U^{-1}\|\|L^{-1}\|$$

So we can use the 1 or ∞ matrix norm and their submulticative pproperty to get an easily computable bound:

$$\|A\|_\infty = 4$$
$$\|U^{-1}\|_\infty = 3/2$$
$$\|L^{-1}\|_\infty = 8$$

$$\kappa = \|A\|\|A^{-1}\| \le 48$$

which is a pessimistic bound.

However, we can also evaluate A^{-1} easily to get

$$A^{-1} = \begin{pmatrix} 1/2 & -1/4 & -1/8 & -1/8 \\ 0 & 1/2 & -1/4 & -1/4 \\ 0 & 0 & 1/2 & -1/2 \\ 1/2 & 1/4 & 1/8 & 1/8 \end{pmatrix}$$

$$\|A\|_\infty = 4$$
$$\|A^{-1}\|_\infty = 1$$

$$\kappa_\infty = 4$$

The condition number based on the 2 norm is not simple to compute in this context. It is $\kappa_2 \approx 1.8$.

One could get an estimate of it however by using the known relationships between the various matrix norms and the matrix 2-norm.

We could also generate a lower bound as mentioned in class by taking a particular perturbation to A and looking at the two solutions that result for a particular righthand side vector b.

5.b

(15 points)

Suppose that

$$Ax \neq b \quad \text{and} \quad A = A^T$$

i.e., the matrix A is symmetric. Let $r = b - Ax$.

Show that if $r^T x \neq 0$ then there exists a backward error $E \in \mathbb{R}^{n \times n}$ such that

$$(A + E)x = b$$

where E is a symmetric rank-1 matrix, i.e.,

$$E = \sigma v v^T, \qquad v \in \mathbb{R}^n, \quad \sigma \in \mathbb{R}.$$

Solution:

To solve this problem we need only choose v and σ so that

$E = \sigma v v^T$ and $(A + E)x = b$. This solution adds a bit more reasoning to show that this E is certainly not the only backward error.

First we show that there are many rank-1 matrices that may not be symmetric that are backward errors. We have

$$\begin{gathered}(A + E)x = b \\ Ex = b - Ax = r\end{gathered}$$

So consider a rank-1 $E = rv^T$. We then have

$$\forall v \in \mathbb{R}^n \quad \text{such that} \quad v^T x = 1 \rightarrow Ex = (rv^T x) = r$$

and E is a rank-1 and possibly nonsymmetric backward error. We can also see how this could be generalized to an E that is not rank-1.

To solve the problem, we must also enforce symmetry so we can assume

$$E = \sigma v v^T$$

For any backward error we must have

$$Ex = r$$

Inserting the assumed form of E yields

$$Ex = \sigma v v^T x = r$$

and we see that we must take v colinear with r. Therefore, since $r^T x \neq 0$, we can solve the problem with

$$\sigma = \frac{1}{r^T x}$$
$$v = r$$

$$Ex = \sigma v v^T x$$
$$\frac{1}{r^T x} r r^T x$$
$$= \frac{r^T x}{r^T x} r = r$$

Therefore there exists a symmetric rank-1 backward error E.