TEKNISKA HÖGSKOLAN I LINKÖPING
Matematiska institutionen
Beräkningsmatematik/Fredrik Berntsson

Exam TANA15 Numerical Linear Algebra, Y4, Mat4

Datum: 25:e Mars, 2025.

Hjälpmedel:

1. Föreläsningsanteckningar utskrivna från kurshemsidan utan egna anteckningar.
2. Räknedosa i fickformat, med nollställt minne och utan instruktionsbok.

Examinator: Fredrik Berntsson

Maximalt antal poäng: 25 poäng. För godkänt krävs 10 poäng.

Jourhavandelärare Fredrik Berntsson (telefon 013 28 28 60)

Good luck!

(4p) **1:** Do the following

a) Let $x \in \mathbb{R}^n$. Prove the inequality $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$.

b) Let $\|\cdot\|$ be a vector norm. Clearly explain what it means for a matrix norm to be *induced* from a vector norm.

c) Prove that $\|I\| = 1$ and $\|A\|\|A^{-1}\| \geq 1$ for all matrix norms induced by a vector norm.

d) Show that the triangle inequality $\|A + B\| \leq \|A\| + \|B\|$ holds for any matrix norm which is induced from a vector norm.

(4p) **2:** Suppose we implement matrix-vector multiplication by a loop:

```
y=zeros(n,1);
for i=1:n
  for j=1:n
    y(i)=y(i)+A(i,j)*x(j);
  end
end
```

on a machine where matrices are stored by column in main memory.

a) Suppose one memory block corresponds exactly to the size of one column `A(:,j)` or the vectors x and y. Further assume that only a couple of memory blocks fit in Cache memory. Clearly explain why the above code is inefficient. Also check the ratio between the number of memory blocks loaded into Cache memory and the number of floating point operations needed.

b) Propose an alternative implementation of matrix-vector multiply and clearly explain why it is better.

(4p) **3:** Suppose A is an $m \times n$, $m > n$, matrix and the linear system $Ax = b$ doesn't have an exact solution. The *Total least squares* solution x satisfies $(A + E)x = b + r$, where $[E, r]$ is given by

$$\min \|[E, r]\|_2 \text{ such that } (A + E)x = b + r.$$

Do the following:

a) Show that the *Total least squares* problem always has a solution.

b) Use the singular value decomposition to derive the solution to the problem. Note that it may not always be possible to find the Total least squares solution using the singular value decomposition and in the case it fails you should give a clear criteria that shows if the formula worked or not.

(4p) **4:** **a)** Let $a = (a_1, a_2, \dots, a_n)^T$ be a column vector. What is the singular value decomposition of a considered as a $n \times 1$ matrix? Similarily what is the singular value decomposition of a^T?

b) Show that if $A \in \mathbb{R}^{m \times n}$ has rank n, then $\|A(A^T A)^{-1} A^T\|_2 = 1$.

(4p) **5:** Do the following:

a) Clearly demonstrate how a bidiagonal reduction $A = UBV^T$ can be computed using Householder reflections. You have to specify which elements of the matrix are used to create each reflection. It is enough to consider the 4×4 case.

b) Give the definition of the singular values of an $m \times n$, $m > n$, matrix A. Also suppose we have all the eigenvalues $\{\lambda_i\}$ of $B^T B$, where $A = UBV^T$ is the bidiagonal reduction. Clearly show how to obtain the singular values of A in terms of the eigenvalues of B. What are the dimensions of the matrices B and $B^T B$?

(5p) **6:** **a)** Show that any matrix $A \in \mathbb{R}^{n \times n}$ can be factorized as $A = QTQ^H$, where Q is unitary and T upper triangular. This is called the *Schur decomposition.*

b) A matrix B is called non-defective if it has a full set of eigenvectors, i.e. the decomposition $B = XDX^{-1}$ exists. Use the Shur decomposition to prove that if A is defective then for any $\varepsilon > 0$ there is a non-defective matrix B such that $\|A - B\|_2 \leq \varepsilon$.

Remark From **b)** we conclude that if a matrix is supposed to be defective and we compute a numerical approximation it is likely that the matrix turns out to be non-defective due to round-off errors.

1: For **a)** we demonstrate the first inequality by

$$\|x\|_\infty^2 = \max_{1\leq i\leq n} |x_i|^2 \leq \sum_{i=1}^{n} |x_i|^2 = \|x\|_2^2.$$

Also, since $|x_i| \leq \|x\|_\infty$, we have

$$\|x\|_2^2 = \sum_{i=1}^{n} |x_i|^2 \leq \sum_{i=1}^{n} \|x\|_\infty^2 = n\|x\|_\infty.$$

For **b)** the matrix norm is defined as

$$\|A\| = \max_{x\neq 0} \frac{\|Ax\|}{\|x\|},$$

where both Ax and x are vectors and the vector norm is used.

For **c)** we use the definition of the matrix norm, and since $Ix = x$ we have

$$\|I\| = \max_{x\neq 0} \frac{\|Ix\|}{\|x\|} = \max_{x\neq 0} \frac{\|x\|}{\|x\|} = 1, \text{ so } 1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\|.$$

For **d)** we use the definition of the induced norm and find that

$$\|A+B\| = \max_{x\neq 0} \frac{\|(A+B)x\|}{\|x\|},$$

and since the triangle inequality holds for the vector norm we obtain

$$\max_{x\neq 0} \frac{\|Ax+Bx\|}{\|x\|} \leq \max_{x\neq 0} \frac{\|Ax\|+\|Bx\|}{\|x\|} \leq \max_{x\neq 0} \frac{\|Ax\|}{\|x\|} + \max_{x\neq 0} \frac{\|Bx\|}{\|x\|} = \|A\| + \|B\|.$$

2: **a)**First during the inner loop $y(i)$ and x can be kept in Cache memory. But the elements $A(i,j)$, for $j = 1,\ldots,n$, all belong to different blocks. Thus a new block needs to be loaded for each multiply `A(i,j)*x(j)`. So the ratio memory loads to multiplies is $1-1$.

b) To fix the issue is is enough to change the order of the loops. So the inner loop copmputes `y(i)=y(i)+A(i,j)*x(j)`, for $i = 1,\ldots,n$. Now the column $A(:,j)$ can be loaded into Cache and n multiplications can be performed until the next vector load is needed.

3: For **a)** we simply observe that the equation $(A+E)x = b+r$ is satisfied, for any x, if $E = -A$ and $r = -b$. The minimum is also bounded from below (by 0). Thus there is some E, r that gives the minimum.

For **b)** we can assume that the agumented matrix $[A, b]$ has full rank since otherwise the minimum would be zero and the linear system $Ax = b$ have a solution. We then compute the singular value decomposition $[A, b] = U\Sigma V^T$ of the $m \times (n+1)$ matrix. The smallest perturbation $[E, r]$ that makes the matrix $[A+E, b+r]$ rank deficient is given by the last singular component $[E, r] = -\sigma_{n+1} u_{n+1} v_{n+1}^T$. There is an x such that $(A+E)x = (b+r)$ if $[A+E, b+r](x, -1)^T = 0$, i.e. $(x, -1)^T$ belongs to the null space of $[A+E, b+r]$. By the construction above the null space is exactly v_{n+1}. So we just take the last singular vector and multiply by a constant so that the last component becomes 1. Thus $x = v_{n+1}(1:n)/v_{n+1}(n+1)$. This is the total least squares solution.

This obviously fails if $v_{n+1}(n+1) = 0$. In that case we have to figure out something else to find the total least squares solution.

4: For **a)** we note that a is $n \times 1$ and the dimensions of the factors are $U \in \mathbb{R}^{n\times n}$, $\Sigma \in \mathbb{R}^{n\times 1}$ and $V \in \mathbb{R}^{1\times 1}$. The decomposition is

$$a = U\Sigma V^T = \begin{pmatrix} \dfrac{a}{\|a\|_2} & A_2 \end{pmatrix} \begin{pmatrix} \|a\|_2 \\ 0 \end{pmatrix} (1),$$

where $A_2 \in \mathbb{R}^{n\times n-1}$ has columns that are orthogonal to a. To obtain the SVD of a^T we simply use $A^T = V\Sigma^T U^T$.

For **b)** we first compute $(A^TA)^{-1} = (V\Sigma^T U^T U \Sigma V^T)^{-1} = V(\Sigma^T\Sigma)^{-1}V^T$. Here $\Sigma^T\Sigma = \mathrm{diag}(\sigma_i^2) \in \mathbb{R}^{n\times n}$. Thus $A(A^TA)^{-1}A^T = U\Sigma V^T V(\Sigma^T\Sigma)^{-1}V^T V\Sigma^T U^T = U\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T U^T$. Since U is orthogonal $\|A(A^TA)^{-1}A\|_2 = \|\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T\|_2$. Evaluate the product of the diagonal matrices to obtain

$$\Sigma(\Sigma^T\Sigma)^{-1}\Sigma^T = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad I \in \mathbb{R}^{n\times n}.$$

The norm is the largest diagonal entry, i.e. 1.

5: For **a)** we illustrate the algorithm as follows: First we use a reflection H_1 applied from the left. The reflection is selected so the elements $A(2:4, 1)$ are set to zero. Second we apply a reflection H_2 from the right to zero out the elements $\widetilde{A}(1, 3:4)$. We get

$$H_1 \begin{pmatrix} x & x & x & x \\ x & x & x & x \\ x & x & x & x \\ x & x & x & x \end{pmatrix} . \begin{pmatrix} + & + & + & + \\ 0 & + & + & + \\ 0 & + & + & + \\ 0 & + & + & + \end{pmatrix} H_2^T = \begin{pmatrix} x & + & 0 & 0 \\ 0 & + & + & + \\ 0 & + & + & + \\ 0 & + & + & + \end{pmatrix}.$$

Now we continue with reflections H_3 and H_4 that zero out $A(3:4, 2)$ and $A(2, 4)$. We get

$$H_3 \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & x & x \\ 0 & x & x & x \\ 0 & x & x & x \end{pmatrix} H_4^T = \begin{pmatrix} x & x & 0 & 0 \\ 0 & + & + & + \\ 0 & 0 & + & + \\ 0 & 0 & + & + \end{pmatrix} H_4^T = \begin{pmatrix} x & x & 0 & 0 \\ 0 & x & + & 0 \\ 0 & 0 & + & + \\ 0 & 0 & + & + \end{pmatrix}.$$

Finally we apply one reflection H_5 from the left to zero out the element $A(4,3)$. We get

$$H_5\begin{pmatrix} x & x & 0 & 0\\ 0 & x & x & 0\\ 0 & 0 & x & x\\ 0 & 0 & x & x \end{pmatrix} = \begin{pmatrix} x & x & 0 & 0\\ 0 & x & x & 0\\ 0 & 0 & + & +\\ 0 & 0 & 0 & + \end{pmatrix},$$

which is bidiagonal.

For **b)** there easiest way to define the singular values is to say that the singular value decomposition is $A = U\Sigma V^T$, where U and V are orthogonal matrices and Σ is diagonal. The singular values σ_k are the diagonal elements of Σ provided that U and V are chosen so that the diagonal elements are positive and sorted in descending order. The dimension of B^TB is $n \times n$ and the dimension of BB^T is $m \times m$. If $A = UBV^T$ then $A^TA = UB^TBU^T$ so the eigenvalues of B^TB are the same as those of A^TA. Also suppose $A = \bar{U}\Sigma\bar{V}^T$ is the singular value decomposition of A. Then $A^TA = \bar{V}\Sigma^T\Sigma\bar{V}^T$. So the eigenvalues of A^TA are $\lambda_i = \sigma_i^2$, where σ_i are the singular values of A. Thus $\sigma_i = \sqrt{\lambda_i}$, $i = 1, 2, \ldots, n$. We are just missing $m-n$ zero singular values to get the correct dimension.

6: For **a)** we pick an eigenpair (λ, x). If we compute the full QR decomposition of $x \in \mathbb{R}^{n\times 1}$ we obtain an orthogonal matrix suxch that $Q = (x, Q_2)$, where $Q_2^H x = 0$. This is assuming that $\|x_1\|_2 = 1$. We find that

$$Q^HAQ = (x,Q_2)^TA(x,Q_2) = (x,Q_2)^H(Ax,AQ_2) = (x,Q_2)^H(\lambda x, AQ_2) =$$
$$\begin{pmatrix} \lambda x^Hx & x^HAQ_2\\ \lambda Q_2^Hx & Q_2^HAQ_2 \end{pmatrix} = \begin{pmatrix} \lambda & w^H\\ 0 & B \end{pmatrix},$$

where we have the correct structure. This is the first step of finding the Hessenberg decomposition. Now we make the induction argument that the Hessenberg decomposition exists for dimension $n-1$ and find $B = Q_1H_1Q_1^H$. We then have

$$Q^HAQ = \begin{pmatrix} \lambda & w^H\\ 0 & Q_1H_1Q_1^H \end{pmatrix} = \begin{pmatrix} 1 & 0\\ 0 & Q_1 \end{pmatrix} = \begin{pmatrix} \lambda & w^H\\ 0 & H_1 \end{pmatrix}\begin{pmatrix} 1 & 0\\ 0 & Q_1 \end{pmatrix}^H.$$

For **b)** we simply note that $A^H = (QTQ^H)^H = QT^HQ^H$. For symmetric matrices, i.e. A real and $A^T = A$, we thus get $A^T = A^H = QT^HQ^H = A = QTQ^H$. Thus $T^H = T$ which means that T is a diagonal since we already knew that T is upper triangular. Also the diagonal elements satisfy $(T)_{ii} = (\bar{T})_{ii}$ which means the elements on the diagonal are real. Since the diagonal elements of T are also the eigenvalues of A this shows that the eigenvalues are real.

For **c)** we assume that A is defective and compute its Shur decomposition $A = QTQ^H$. For A to be defective it has to have at least one eigenvalue λ_1 with an algebraic multiplicity $\gamma_1(\lambda_1)$ strictly larger than the geometric multiplicity $\gamma_2(\lambda_1)$. Thus, if all diagonal elements of T were different then the matrix A would be non-defective. Thus we pick a diagonal matrix $D = \mathrm{diag}(\epsilon_1, \ldots, \epsilon_n)$ so that $T + D$ has unique diagonal elements. Then $B = Q(T+D)Q^H$ is non-defective and $\|A-B\|_2 = \|D\|_2 \leq \max|\epsilon_i| = \epsilon$.