Tentamen Numerical Mathematics 2
July 9, 2010

Duration: 3 hours.
In front of the questions one finds the weights used to determine the final mark.

Problem 1

a. [2] Consider $Ax = b$ with A nonsingular. Compute the absolute and the relative condition number of this problem. How is the standard condition number of A related to the latter?

b. [2] Give an example of a problem $Ax = b$ in which the relative error in b propagates with the standard condition number of A.

c. [3] Make an LU factorization (i) without pivoting, (ii) with partial pivoting and (iii) with complete pivoting of the matrix in the following system and solve it using the respective factorizations:

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

What is the advantage of pivoting?

d. [3] Consider $Ax = b$ and A singular. How can we get an approximate solution of this equation using the pseudo-inverse of A? Indicate how the pseudo-inverse is constructed. Which property holds for the thus obtained solution x? In which cases is the pseudo-inverse useful?

Problem 2

a. [3] Let A be real symmetric, and μ, x an approximate eigenpair. Show that

$$\min_{\lambda \in \sigma(A)} |\lambda - \mu| \leq \frac{||Ax - \mu x||_2}{||x||_2}$$

b. [2] Consider the iteration

$$(A - \mu I)y_{n+1} = x_n, \quad x_{n+1} = y_{n+1}/||y_{n+1}||, \quad n = 0, 1, 2, ...$$

where x_0, A and μ are given. Where does the vector x_n converge to if all eigenvalues of A are different? And what determines its speed of convergence?

c. [3] Determine the vector that defines the Householder transformation which turns the vector $[2, 1, 1]^T$ into a vector of the form $[\alpha, 0, 0]^T$.

d. [2] Let A be a square nonsymmetric real matrix, to which we apply the QR method. Show that after each step of the QR method the eigenvalues of the result matrix are equal to those of the original matrix. How can we finally read the eigenvalues (can there be complex eigenvalues)? What is done to speedup the method?

Exam questions continue on other side

Problem 3

a. [2] Why is the Chebyshev polynomial of interest for interpolation and how is it used there?

b. [3] What is meant by Gauss Legendre and Gauss Chebyshev integration? And what if we also add the name Lobatto to it? For which degree of polynomials are they exact?

c. [2] Let $f(x)$ be a continuous function on [a,b] and $p_n(x)$ a polynomial of degree n. How can we check whether p_n is the best polynomial approximation of f on [a,b]?

d. [2] Explain how Legendre polynomials can be used to derive high-order implicit Runge-Kutta methods.

Workout Tentamen Numerical Mathematics 2
July 9, 2010

Problem 1

a. [2] Consider $Ax = b$ with A nonsingular. Compute the absolute and the relative condition number of this problem. How is the standard condition number of A related to the latter?
Discussed during course. By definition, see Definition 2.1 on page 34, the absolute condition number is $K_{abs} = \max_{\delta b} \frac{||\delta x||}{||\delta b||}$ and $K_{rel} = \max_{\delta b} \frac{||\delta x||/||x||}{||\delta b||} = K_{abs} \frac{||b||}{||x||}$, where it holds here that $A(x + \delta x) = b + \delta b$. Subtracting $Ax = b$ we find $A\delta x = \delta b$. So in this case $K_{abs} = \max_{\delta b} \frac{||A^{-1}\delta b||}{||\delta b||}| \equiv ||A^{-1}||$ and $K_{rel} = ||A^{-1}|| \frac{||b||}{||x||}$.
The standard condition number for $Ax = b$ is given by $K_A = ||A^{-1}||||A||$, (3.4) on page 60. This can be found from maximizing the current relative condition number over all possible b and since there is a unique relation ship between x and b we can formally also maximize it over all x. So $K_A = \max_b K_{rel} = \max_x K_{rel} = ||A^{-1}|| \max_x \frac{||b||}{||x||} = ||A^{-1}|| \max_x \frac{||Ax||}{||x||} = ||A^{-1}||||A||$

b. [2] Give an example of a problem $Ax = b$ in which the relative error in b propagates with the standard condition number of A.
Discussed during course and in lab session 1, see also Exercise 10 on page 123. Let A be a symmetric 2x2 matrix and $Av_i = \lambda_i v_i$ with $0 < \lambda_1 < \lambda_2$ and $||v_i||_2 = 1$. Then consider the problem $Ax = v_2$ and the perturbed one $A(x + \delta x) = v_2 + \epsilon v_1$. Clearly $x = v_2/\lambda_2$ and $\delta x = \epsilon v_1/\lambda_1$. Clearly $\frac{||\delta x||_2/||x||_2}{||\delta b||_2/||b||_2} = \lambda_2/\lambda_1$. Now here $||A||_2 = \lambda_2$ and $||A^{-1}|| = 1/\lambda_1$. So indeed the relative error in b propagates with the standard condition number.

c. [3] Make an LU factorization (i) without pivoting, (ii) with partial pivoting and (iii) with complete pivoting of the matrix in the following system and solve it using the respective factorizations:

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

What is the advantage of pivoting?
Discussed in the course, see alos Fig. 3.2 on page 89.
(i) Without pivoting the structure of the LU factorization looks like

$$\begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & u_{22} \end{bmatrix}$$

Now l_{21} contains the multiplier, i.e. the factor by which the first row of A must be multiplied to get a zero at the (2,1) position after subtraction. So $l_{21} = 2$. u_{22} follows from the elimination step $u_{22} = a_{22} - l_{21}a_{12} = -1$. Hence we get the following new formulation of the problem

$$\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

The solution follows in two steps. First we solve $Ly = b$. So $y_1 = 1$ and $y_2 = 3 - l_{21}y_1 = 3 - 2 \cdot 1 = 1$. And next we solve $Ux = y$, so $x_2 = -1$ and $x_1 = (y_1 - u_{12}x_2)/u_{11} = (1 - 3 \cdot (-1))/2 = 2$.
(ii) With partial pivoting we have to look for the maximum in the first column. This

appears to be the second element. We want the maximum value in the column at the top position. So we interchange the first and second column. With this the problem turns into.

$$\begin{bmatrix} 4 & 5 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Now the LU will be of the form

$$\begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 0 & u_{22} \end{bmatrix}$$

Here $l_{21} = 1/2$ and $u_{22} = a_{22} - l_{21}a_{12} = 3 - \frac{1}{2} \cdot 5 = \frac{1}{2}$. Hence we get the following new formulation of the problem

$$\begin{bmatrix} 1 & 0 \\ \frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

The solution process is as before and yields of course the same solution.
(iii) With complete pivoting we have to look for the maximum (in abs. value) in the whole matrix, which is 5. Now we are going to interchange first the two rows, which gives us the system from the partial pivoting case. Next we have to interchange the columns and the associated unknowns. This yields the following system.

$$\begin{bmatrix} 5 & 4 \\ 3 & 2 \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Now the LU will be of the form

$$\begin{bmatrix} 1 & 0 \\ l_{21} & 1 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 0 & u_{22} \end{bmatrix}$$

Here $l_{21} = 3/5$ and $u_{22} = a_{22} - l_{21}a_{12} = 2 - \frac{3}{5} \cdot 4 = -\frac{2}{5}$. Hence we get the following new formulation of the problem

$$\begin{bmatrix} 1 & 0 \\ \frac{3}{5} & 1 \end{bmatrix} \begin{bmatrix} 5 & 4 \\ 0 & \frac{3}{5} \end{bmatrix} \begin{bmatrix} x_2 \\ x_1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$$

Solving gives again the same solution as before.
The purpose of pivoting is to prevent the propagation of round-off errors.
A few additional remarks:
In partial pivoting all multipliers are less than one in absolute value, which precludes that the equation that will be adapted is not overwhelmed by the first row. With complete pivoting this is also the case but now also row wise in U the maximum is on the diagonal. One could even write $U = D\tilde{U}$ where D is the diagonal from U and hence $\tilde{U}$ has diagonal one. In this case all elements of $\tilde{U}$ are less equal to one in absolute value. Hence this means that in complete pivoting we also try to limit the propagation of round-off errors during the back substitution.

d. [3] Consider $Ax = b$ and A singular. How can we get an approximate solution of this equation using the pseudo-inverse of A? Indicate how the pseudo-inverse is constructed. Which property holds for the thus obtained solution x? In which cases is the pseudo-inverse useful?
Treated in the course, many similarities with lab session exercise 1c, see also page 116

in the book.
For the pseudo inverse we use the singular value decomposition (SVD). Hence there exist unitary matrices U and V such that $U^*AV = F$ where F is a diagonal matrix with on the diagonal the singular values. These are the square root of the eigenvalues of A^*A. One defines a pseudo inverse of F by a matrix which has the size of the transform of F and on the diagonal the reciprocals of the singular values. However if a singular value is less than a user specified tolerance we set this reciprocal to zero. The pseudo inverse is indicated by $F^\dagger$ and this defines also the pseudo inverse of A: $A^\dagger = V^*F^\dagger U$. The solution of the above system is then given by $x = A^\dagger b$.
If A is singular then in principle one can add any multiple of the singular vector to x. The pseudo inverse is such that we get the solution which is shortest in 2-norm.
The pseudo inverse is in particular useful for ill-posed systems (see 1b above). By a suitable choice of the tolerance one can filter out the influence of almost zero singular values, which can dramatically increase the influence of round off errors. Of course what is small is relative. We mean small singular values with respect to the biggest one. The biggest one will be proportional to the norm of A.

Problem 2

a. [3] Let A be real symmetric, and μ, x an approximate eigenpair. Show that

$$\min_{\lambda \in \sigma(A)} |\lambda - \mu| \leq \frac{||Ax - \mu x||_2}{||x||_2}$$

Treated during the course, it is Theorem 5.5 (page 190) from the book.

b. [2] Consider the iteration

$$(A - \mu I)y_{n+1} = x_n, \quad x_{n+1} = y_{n+1}/||y_{n+1}||, \quad n = 0, 1, 2, ...$$

where x_0, A and μ are given. Where does the vector x_n converge to if all eigenvalues of A are different? And what determines its speed of convergence?
Treated during the course (see Section 5.3.2 on page 195), convergence studied during lab session 2. This is inverse iteration or the inverse power method. This is the Power method with matrix $(A - \mu I)^{-1}$. The method will converge to the eigenvector corresponding to the biggest eigenvalue of this matrix. This matrix has the same eigenvectors as A and if λ is an eigenvalue of A then $1/(\lambda - \mu)$ is an eigenvalue of this matrix. So the biggest eigenvalue of this matrix is obtained for the eigenvalue of A which is closest to μ. We know that for the power method the convergence is determined by ratio of the one but largest eigenvalue divided by the largest eigenvalue. So if we call the eigenvalue closest to μ λ^* and the one but closest eigenvalue $\lambda^\dagger$ then the speed of convergence is given by $|\lambda^* - \mu|/|\lambda^\dagger - \mu|$.

c. [3] Determine the vector that defines the Householder transformation which turns the vector $[2, 1, 1]^T$ into a vector of the form $[\alpha, 0, 0]^T$.
Treated during the course, see also Section 5.6.1 (page 204) from the book. The Householder matrix is a mirroring operator defined by $H = I - \frac{2}{||v||^2} vv^T$. The vector v is the normal on the mirroring plane. The Householder matrix is also an orthogonal matrix so the length of the original vector is the same as the target vector. Hence $\alpha = \pm\sqrt{6}$. The normal of the mirroring plane is then simply the vector $v = [2, 1, 1] - \pm\sqrt{6}[1, 0, 0]$. So $v = [1 - \sqrt{6}, 1, 1]$ will do the job.

d. [2] Let A be a square nonsymmetric real matrix, to which we apply the QR method. Show that after each step of the QR method the eigenvalues of the result matrix are equal to those of the original matrix. How can we finally read the eigenvalues (can there be complex eigenvalues)? What is done to speedup the method?
Treated during the course and in labsession 2, see also Section 5.4, first part of 5.5 and 5.7.1, also see the sheets with results on Nestor where another example is given. The iteration is as follows $QR = A_n,\ A_{n+1} = RQ$ for $n = 0, 1, 2, ...$ with $A_0 = A$. From this we have that $A_{n+1} = Q^T A_n Q$ which is a similarity transformation. Hence the eigenvalues of A_{n+1} are the same as those from A_n and hence from A. If we start from a real matrix all iterates A_n stay real and we will converge to the real Schur form. This has 1x1 and 2x2 blocks on the diagonal and a lower triangular part that is zero. The 1x1 blocks on the diagonal give the real eigenvalues of the matrix and the eigenvalues of the 2x2 blocks give the complex eigenvalues.
This method is speeded up by using shifts as follows $QR = A_n - s_n I,\ A_{n+1} = RQ + s_n I$. The eigenvalue which is closest to this shift is converging fastest in the current step and will occur at the last position of A_{n+1}. By picking the last element of the matrix as shift during the whole iteration, we get a very fast convergence of the eigenvalue which was closest to the first shift (i.e. the value of the last element of A). Once convergence of this last element is obtained we continue the process on a deflated matrix, i.e. we consider the matrix where the last row and column are omitted)

Problem 3

a. [2] Why is the Chebyshev polynomial of interest for interpolation and how is it used there?
Discussed during course and also in lab session 3, see (8.7),(8.6), and minimax property on page 428. The interpolation error is a product of a polynomial, which has as zeros the interpolation points and a coefficient one in front of the term with highest degree, and a derivative of the interpolated function at some point in the interpolation interval divided by a factorial. We can have control over the magnitude of the polynomial part of the error by choosing the interpolation points appropriately. The best would be if it is up to a factor a Chebyshev polynomial, since $\max_{x\in[-1,1]} |T_n(x)/a_n^T| \leq \max_{x\in[-1,1]} |p_n(x)/a_n^p|$ for any polynomial of degree n. So in order to minimize the polynomial part we have to take the zeros of the Chebyshev polynomial of the same degree and shift them to the interval we interpolate on, and use these as interpolation points.
An alternative answer could be to mention formula (10.25) and the formula on top of page 337, which shows that Chebyshev interpolation leads us very close to the best behavior we can ever have.

b. [3] What is meant by Gauss Legendre and Gauss Chebyshev integration? And what if we also add the name Lobatto to it? For which degree of polynomials are they exact?
Treated during course (see notes on orthogonal polynomials on Nestor), see also Section 10.2 from the book, and lab session 3.
For Gauss integration we like to find the $n + 1$ interpolation points such that

$$\int_a^b w(x) f(x) dx$$

is exact for a polynomial of as high degree as possible. For the case $w(x) \equiv 1$ this occurs if we take as interpolation points the zeros of the Legendre orthogonal

polynomial of degree $n+1$ shifted to the interval [a,b]. It can be shown that the polynomial that can still be integrated exactly is of degree $2n+1$. We use Chebyshev polynomials if $w(x)$ is $1/\sqrt{1-x^2}$ shifted to the interval [a,b]. The degree of exactness is then the same as with Legendre polynomials. For Lobatto rules the ends of the interval are taken as interpolation points. This means that we have to take the zeros from a linear combination of three orthogonal polynomials which is also zero at the end points of the integration interval. Hence we have that the attainable degree is $2n-1$.

c. [2] Let $f(x)$ be a continuous function on [a,b] and $p_n(x)$ a polynomial of degree n. How can we check whether p_n is the best polynomial approximation of f on [a,b]?
Treated during lecture, see Section 10.8 and lab session 3.
We can look at the error. If it satisfies the equioscillation theorem of Chebyshev then it is the polynomial of best approximation.

d. [2] Explain how Legendre polynomials can be used to derive high-order implicit Runge-Kutta methods.
This is not part of the material this year. But for the interested reader see Section 11.8.3. Runge-Kutta methods are used for the integration of initial value problems $y' = f(t,y)$. Now we can also write this ODE as $y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f(s,y(s))ds$. This has be shown before in Numerical Mathematics 1 and is the basis the explicit and implicit midpoint methods, the trapezium method and Adams-Bashfort and Adams-Moulton methods. Now we can apply Gauss Legendre methods to the integral. In Section .. a number of these methods are listed. In fact implicit midpoint can be viewed as a Gauss Legendre method with 1 interpolation point and the trapezium rule as a Gauss Legendre Lobatto rule. Forward and backward Euler are Gauss-Legendre-Radau methods. Gauss Legendre methods give high-order of accuracy for few interpolation points (which means few stages in Runge-Kutta method). However they are strongly implicit. If we have a system of ODEs of order N then we need to solve an implicit system of order sN where s is the number of stages. This makes them less popular for the time being.