CHAPTER 18. Numerical Methods in Linear Algebra

Sec. 18.1 Linear Systems: Gauss Elimination

Problem Set 18.1. Page 893

5. **System without solution.** The left side of the second equation equals minus three times the left side of the first equation. Hence for a solution to exist the right sides should be related in the same fashion; they should equal, for instance, 16 and –48 (instead of 48). Of course, for most systems with more than two equations, one cannot immediately see whether there will be solutions, but the Gauss elimination (with partial pivoting) will work in each case, giving the solution(s) or indicating that there is none.

7. **System with a unique solution. Pivoting.** Worked-out examples are given in the text. They show all the details. Review those first because there is little we can do for a better understanding and we shall have to restrict ourselves to a more detailed discussion of Table 18.1, which contains the algorithm for the Gauss elimination, and the addition of a few remarks. Consider Table 18.1. To follow the discussion, control it for Prob. 7 in terms of matrices with paper and pencil. In each case, write down all three rows of a matrix, not just one or two rows, as was done below to save some space and to avoid copying the same numbers several times. At the beginning, $k = 1$. Since $a_{11} = 0$, you must pivot. Line 2 in Table 18.1 requests to look for the absolutely greatest a_{j1}. This is a_{31}. According to the algorithm, you have to interchange Equations 1 and 3, that is, Rows 1 and 3 of the *augmented* matrix . This gives

$$\left[\begin{array}{ccc|c} 13 & -8 & 0 & 79 \\ 6 & 0 & -8 & -38 \\ 0 & 6 & 13 & 61 \end{array}\right]. \tag{A}$$

Don't forget to interchange the entries on the right side (that is, in the last column of the augmented matrix). In line 2 of Table 18.1, the phrase 'the smallest' $j \geqq k$ is necessary since there may be several entries of the same absolute value (or even of the same size), and the computer needs unique instructions what to do in each operation. To get 0 as the first entry of Row 2, subtract 6/13 times Row 1 from Row 2. The new Row 2 is

$$\left[\begin{array}{ccc|c} 0 & 3.692308 & -8 & -74.461538 \end{array}\right]. \tag{B}$$

This was $k = 1$ and $j = 2$ in lines 3 and 4 in the table.

Now comes $k = 1$ and $j = n = 3$ in line 3. The calculation is $m_{31} = a_{31}/a_{11} = 0/13 = 0$. Hence the operations in line 4 simply have no effect, they merely reproduce Row 3 of the matrix in (A). This was $k = 1$.

Now comes $k = 2$. Look at line 2 in the table. Since $6 > 3.692308$, interchange Row 2 in (B) and Row 3 in (A). This gives the matrix

$$\left[\begin{array}{ccc|c} 13 & -8 & 0 & 79 \\ 0 & 6 & 13 & 61 \\ 0 & 3.692308 & -8 & -74.461538 \end{array}\right]. \tag{C}$$

In line 3 of the table with $k = 2$ and $j = k + 1 = 3$ calculate

$$m_{32} = a_{32}/a_{22} = 3.692308/6 = 0.615385.$$

Performing the operations in line 4 of the table for $p = 3, 4$, you obtain the new Row 3

$$\left[\begin{array}{ccc|c} 0 & 0 & -16 & -112 \end{array}\right].$$

The system and its matrix have now reached triangular form, and back substitution begins with line 6 of the table,

$$x_3 = a_{34}/a_{33} = -112/(-16) = 7.$$

(Remember that in the table the right sides b_1, b_2, b_3 are denoted by a_{14}, a_{24}, a_{34}, respectively.) Line 7 of the table with $i = 2, 1$ gives

$$x_2 = \frac{1}{6}(61 - 13 \cdot 7) = -5 \qquad (i = 2)$$

and

$$x_1 = \frac{1}{13}(79 - (-8 \cdot (-5) + 0 \cdot 7) = 3 \qquad (i = 1).$$

Depending on the number of digits you use in your calculation, your values may be slightly affected by round-off.

11. System with more than one solution. Solutions exist if and only if the coefficient matrix and the augmented matrix have the same rank (see Sec. 6.5). If these matrices have equal rank $r < n$ (n the number of unknowns), there exists more than one solution and, in fact, infinitely many solutions. In this case, to one or more suitable unknowns there can be assigned arbitrary values. In the present problem, $n = 3$ and the system is nonhomogeneous. For such a system you may have $r = 3$ (a unique solution), $r = 2$ (one (suitable) unknown remains arbitrary), $r = 1$ (two (suitable) variables remain arbitrary). $r = 0$ is impossible because then the matrices would be zero matrices. In most cases you have choices which of the variables you want to leave arbitrary; the present result will show this. To avoid misunderstandings: you need not determine those ranks, but the Gauss elimination will automatically give all solutions. ***Your CAS may give only some solutions*** (for example, those obtained by equating arbitrary unknowns to zero); so be careful. Following line 2 in Table 18.1, exchange Rows 1 and 2, so that the augmented matrix is

$$\left[\begin{array}{ccc|c} -5 & 7 & 2 & -4 \\ 2 & 5 & 7 & 25 \\ 1 & 22 & 23 & 71 \end{array}\right].$$

For $k = 1$ the operations in lines 3 and 4 of the table with $j = 2$ and 3 give

$$\left[\begin{array}{ccc|c} -5 & 7 & 2 & -4 \\ 0 & 7.8 & 7.8 & 23.4 \\ 0 & 23.4 & 23.4 & 70.2 \end{array}\right] \begin{array}{l} \\ \text{Row } 2 + 0.4 \text{ Row } 1 \\ \text{Row } 3 + 0.2 \text{ Row } 1. \end{array} \qquad \text{(D)}$$

For $k = 2$ the operations in lines 3 and 4 of the table with $j = 3$ give the new Row 3 as a row of zeros,

$$\left[\begin{array}{ccc|c} 0 & 0 & 0 & 0 \end{array}\right] \text{Row } 3 - 3 \text{ Row } 2.$$

This was the elimination. Now begins the back substitution. From Row 2 in (D) you obtain

$$x_2 = \frac{1}{7.8}(23.4 - 7.8x_3) = 3 - x_3. \qquad \text{(E)}$$

With this, Row 1 in (D) gives

$$x_1 = \frac{1}{-5}(-4 - 7x_2 - 2x_3) = \frac{1}{5}(4 + 7(3 - x_3) + 2x_3) = \frac{1}{5}(25 - 5x_3) = 5 - x_3. \qquad \text{(F)}$$

You see that you have no condition on x_3; hence x_3 is arbitrary. Solving (E) for x_3, you have

$$x_3 = 3 - x_2. \qquad \text{(G)}$$

Substituting (G) into (F), you obtain

$$x_1 = 5 - (3 - x_2) = 2 + x_2. \qquad \text{(H)}$$

This shows that you can leave x_2 arbitrary; then x_1 and x_3 are uniquely determined in terms of x_2. Equations (G) and (H) give the form of the solution shown on p. A39 in Appendix 2 of the book.

Sec. 18.2 Linear Systems: LU-Factorization, Matrix Inversion

Example 1. Doolittle's method (p.895). In the calculation of the entries of **L** and **U** (or $\mathbf{L}^T$ in Cholesky's method) in the factorization **A** = **LU** with given **A** you employ the usual matrix multiplication

Row times Column.

In all three methods in this section, the point is that the calculation can proceed in an order such that you solve only one equation at a time. This is possible because you are dealing with triangular matrices, so that the sums of $n = 3$ products often reduce to sums of 2 products or even to a single product, as you will see. This will be a discussion of the steps of the calculation on p. 895 in terms of the matrix equation **A** = **LU**, written out (see the result on p, 896 at the top)

$$\mathbf{A} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}.$$

Remember that in Doolittle's method the main diagonal of L is 1, 1, 1. Also, the notation m_{jk} suggests *multiplier*, because in Doolittle's method the matrix L is the matrix of the multipliers in the Gauss elimination. Begin with Row 1 of **A**. The entry $a_{11} = 3$ is the dot product of the first row of L and the first column of U; thus,

$$3 = [1 \quad 0 \quad 0]\,[u_{11} \quad 0 \quad 0]^T = 1 \cdot u_{11},$$

where 1 is prescribed . Thus, $u_{11} = 3$. Similariy, $a_{12} = 5 = 1 \cdot u_{12} + 0 \cdot u_{22} + 0 \cdot 0 = u_{12}$; thus $u_{12} = 5$. Finally, $a_{13} = 2 = u_{13}$. This takes care of the first row of **A**. In connnection with the second row of **A** you have to consider the second row of **L**, which involves m_{21} and 1. You obtain

$$\begin{aligned} a_{21} &= 0 = m_{21} u_{12} + 0 \quad + 0 \; = m_{21} \cdot 5, && \text{hence } m_{21} = 0 \\ a_{22} &= 8 = m_{21} u_{12} + 1 \cdot u_{22} + 0 \; = u_{22}, && \text{hence } u_{22} = 8 \\ a_{23} &= 2 = m_{21} u_{13} + 1 \cdot u_{23} + 0 \; = u_{23}, && \text{hence } u_{23} = 2. \end{aligned}$$

In connection with the third row of **A** you have to consider the third row of **L**, consisting of m_{31}, m_{32}, 1. You obtain

$$\begin{aligned} a_{31} &= 6 = m_{31} u_{11} + 0 + 0 && = m_{31} \cdot 3, && \text{hence } m_{31} = 2 \\ a_{32} &= 2 = m_{31} u_{12} + m_{32} u_{22} + 0 && = 2 \cdot 5 + m_{32} \cdot 8 && \text{hence } m_{32} = -1 \\ a_{33} &= 8 = m_{31} u_{13} + m_{32} u_{23} + 1 \cdot u_{33} && = 2 \cdot 2 - 1 \cdot 2 + u_{33}, && \text{hence } u_{33} = 6. \end{aligned}$$

In (4) on p. 896 the first line concerns the first row of **A** and the second line concerns the first column of **A**; hence in that respect the order of calculation is slightly different from that in Example 1.

Problem Set 18.2. Page 899

7. **Cholesky's method.** You see that the given matrix **A** is symmetric. Its Cholesky factorization is

$$\begin{bmatrix} 9 & 6 & 12 \\ 6 & 13 & 11 \\ 12 & 11 & 26 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}.$$

This matrix **A** is positive definite. For a larger matrix this may be difficult to check, although in some cases it may be concluded from the kind of physical (or other) application. However, it is not necessary to check for definiteness because all that might happen is that you obtain a complex triangular matrix **L** and would then probably choose another method. Going through **A** row by row and applying matrix multiplication (Row times Column) as just before you calculate the following.

$$a_{11} = 9 = l_{11}^2 + 0 + 0 \quad = l_{11}^2, \quad \text{hence} \quad l_{11} = 3$$
$$a_{12} = 6 = l_{11} l_{21} + 0 + 0 \quad = 3 l_{21}, \quad \text{hence} \quad l_{21} = 2$$
$$a_{13} = 12 = l_{11} l_{31} + 0 + 0 \quad = 3 l_{31}, \quad \text{hence} \quad l_{31} = 4.$$

In the second row of **A** you have $a_{21} = a_{12}$ (symmetry!) and need only two calculations,

$$a_{22} = 13 = l_{21}^2 + l_{22}^2 + 0 \quad = 4 + l_{22}^2, \quad \text{hence} \quad l_{22} = 3$$
$$a_{23} = 11 = l_{21} l_{31} + l_{22} l_{32} + 0 \quad = 2 \cdot 4 + 3 l_{22}, \quad \text{hence} \quad l_{32} = 1.$$

In the third row of **A** you have $a_{31} = a_{13}$ and $a_{32} = a_{23}$ and need only one calculation,

$$a_{33} = 26 = l_{31}^2 + l_{32}^2 + l_{33}^2 = 16 + 1 + l_{33}^2, \quad \text{hence} \quad l_{33} = 3.$$

Now solve $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{b} = [17.4 \quad 23.6 \quad 30.8]^T$. You first use **L** and solve $\mathbf{Ly} = \mathbf{b}$, where $\mathbf{y} = [y_1 \quad y_2 \quad y_3]^T$. Since **L** is triangular, you just do back substitution as in the Gauss algorithm. Now since **L** is *lower* triangular, whereas the Gauss elimination produces an *upper* tringular matrix, begin with the first equation and obtain y_1. Then obtain y_2 and finally y_3. This simple calculation is written to the right of the corresponding equations.

$$\begin{bmatrix} 3 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 1 & 3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 17.4 \\ 23.6 \\ 30.8 \end{bmatrix} \qquad \begin{aligned} y_1 &= \tfrac{1}{3} \cdot 17.4 = 5.8 \\ y_2 &= \tfrac{1}{3}(23.6 - 2y_1) = 4 \\ y_3 &= \tfrac{1}{3}(30.8 - 4y_1 - y_2) = 1.2 \end{aligned}$$

In the second part of the procedure you solve $\mathbf{L}^T\mathbf{x} = \mathbf{y}$ for **x**. This is another back substitution. Since $\mathbf{L}^T$ is *upper* triangular, just as in the Gauss method after the elimination has been completed, the present back substitution is exactly as in the Gauss method, beginning with the last equation, which gives x_3, then using the second equation to get x_2, and finally the first equation to obtain x_1. These calculations are again written to the right of the corresponding equations.

$$\begin{bmatrix} 3 & 2 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5.8 \\ 4 \\ 1.2 \end{bmatrix} \qquad \begin{aligned} x_1 &= \tfrac{1}{3}(5.8 - 2x_2 - 4x_3) = 0.6 \\ x_2 &= \tfrac{1}{3}(4 - x_3) = 1.2 \\ x_3 &= \tfrac{1}{3} \cdot 1.2 = 0.4 \end{aligned}$$

Check the solution by substituting it into the given linear system.

17. **Matrix inversion.** The method suggested in this section is illustrated by Example 1 in Sec. 6.7, at which you may perhaps look first. The matrix in the present problem is

$$\begin{bmatrix} -2 & 4 & -1 \\ -2 & 3 & 0 \\ 7 & -12 & 2 \end{bmatrix}.$$

To find its inverse, apply the Gauss-Jordan method to the 3×6 matrix

$$\mathbf{G} = \left[\begin{array}{ccc|ccc} -2 & 4 & -1 & 1 & 0 & 0 \\ -2 & 3 & 0 & 0 & 1 & 0 \\ 7 & -12 & 2 & 0 & 0 & 1 \end{array}\right].$$

The left 3×3 submatrix is the given matrix. The right 3×3 submatrix is the 3×3 unit matrix. At the end of the process the left 3×3 submatrix will be the 3×3 unit matrix, and the right 3×3 submatrix will be the inverse of the given matrix. Leave Row 1 of G unchanged. Replace Row 2 by Row 2 – Row 1. Replace Row 3 by Row 3 + 3.5 Row 1. This gives the new matrix

$$\mathbf{H} = \left[\begin{array}{ccc|ccc} -2 & 4 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & -1 & 1 & 0 \\ 0 & 2 & -1.5 & 3.5 & 0 & 1 \end{array}\right].$$

Now leave Rows 1 and 2 of **H** unchanged. Replace Row 3 by Row 3 + 2 Row 2. The new matrix is

$$\mathbf{J} = \left[\begin{array}{ccc|ccc} -2 & 4 & -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 0.5 & 1.5 & 2 & 1 \end{array}\right].$$

This was Gauss. The given matrix is triangularized. Now comes Jordan and diagonalizes it. Multiply Row 1 by −1/2, Row 2 by −1, and Row 3 by 2. This gives the matrix

$$\mathbf{K} = \left[\begin{array}{ccc|ccc} 1 & -2 & 0.5 & -0.5 & 0 & 0 \\ 0 & 1 & -1 & 1 & -1 & 0 \\ 0 & 0 & 1 & 3 & 4 & 2 \end{array}\right].$$

Now eliminate 0.5 and −1 from the third column of **K**. Replace Row 1 by Row 1 − 0.5 Row 3. Replace Row 2 by Row 2 + Row 3. Leave Row 3 unchanged. The new matrix is

$$\mathbf{M} = \left[\begin{array}{ccc|ccc} 1 & -2 & 0 & -2 & -2 & -1 \\ 0 & 1 & 0 & 4 & 3 & 2 \\ 0 & 0 & 1 & 3 & 4 & 2 \end{array}\right].$$

Finally eliminate −2 in the second column of **M**. Replace Row 1 of **M** by Row 1 + 2 Row 2. The new matrix is

$$\mathbf{N} = \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 6 & 4 & 3 \\ 0 & 1 & 0 & 4 & 3 & 2 \\ 0 & 0 & 1 & 3 & 4 & 2 \end{array}\right].$$

The last three columns constitute the inverse of the given matrix. The following discussion may perhaps help you to a better understanding of the method. Follow the discussion with paper and pencil and an example of your own or in terms of the problem just solved. From $\mathbf{Ax} = \mathbf{b}$ you have $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, the existence of the inverse being assumed. The Gauss elimination converts the given system to $\mathbf{Ux} = \mathbf{b}^*$ with upper triangular **U** and $\mathbf{b}^*$ obtained from **b** in the calculations. $\mathbf{x} = \mathbf{U}^{-1}\mathbf{b}^*$ is then obtained by back substitution. In the Gauss-Jordan method you go on and reduce **U** to the identity matrix **I**, so that the system becomes $\mathbf{Ix} = \mathbf{b}^{**}$ with $\mathbf{b}^{**}$ obtained from $\mathbf{b}^*$ in this process. Since $\mathbf{Ix} = \mathbf{x}$, you have directly $\mathbf{x} = \mathbf{b}^{**}$, that is, $\mathbf{b}^{**}$ is the solution, and back substitution is avoided. Now comes the crucial point of this discussion. If for **b** you chose the first column of the unit matrix, call it $\mathbf{b}_1$, you are dealing with $\mathbf{Ax} = \mathbf{b}_1$, hence with $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}_1$. But by the usual matrix multiplication, $\mathbf{A}^{-1}\mathbf{b}_1$ is simply the first column of the inverse matrix because this multiplication picks from each row of $\mathbf{A}^{-1}$ the first entry. And by Gauss-Jordan, this solution **x** appears as the transform of $\mathbf{b}_1$ in this process, that is, you have actually obtained the first column of the inverse (as Column 4 of your above 3 × 6 matrix **G**). Similarly, if you choose the second column of the unit matrix, call it $\mathbf{b}_2$, Gauss-Jordan will give you as solution the second column of the inverse matrix as the transform of $\mathbf{b}_2$. And so on.

Sec. 18.3 Linear Systems: Solution by Iteration

Problem Set 18.3. Page 905

5. Gauss-Seidel iteration. This is a case in which you reorder the equations so that the large entries stand on

the main diagonal in order to obtain convergence. That is, the third equation becomes the first and is solved for x_1. The first equation becomes the second and is solved for x_2. The second equation becomes the third and is solved for x_3. With this rearrangement you can expect convergence. Indeed, **C** in (7) can be shown to have the eigenvalues 0, 0.151, and −0.061, approximately. (In verifying this, don't forget to divide the rows of the coefficient matrix of the rearranged system by 6, 9, 8, respectively.) Hence you can expect rapid convergence (see the discussion between formulas (7) and (8) in the text). In contrast, if you left the given order and solved the first equation for x_1, the second for x_2, and the third for x_3, you do not get convergence because then the eigenvalues of **C** are 0, 8.5, and −51, approximately.

7. **Effect of starting values.** The point of the problem is to show that there is surprisingly little difference between corresponding values, as the answer on p. A40 in Appendix 2 shows, although the starting values differ considerably. Hence it is hardly necessary to search extensively for "good" starting values.

13. **Convergence.** The matrix of the system is

$$\begin{bmatrix} 4 & 0 & 5 \\ 1 & 6 & 2 \\ 8 & 2 & 1 \end{bmatrix}.$$

To obtain convergence, reorder the rows as shown.

$$\begin{bmatrix} 8 & 2 & 1 \\ 1 & 6 & 2 \\ 4 & 0 & 5 \end{bmatrix}.$$

Then divide the rows by 8, 6, and 5, respectively, as required in (13) (see $a_{jj} = 1$ at the end of the formula). This gives

$$\begin{bmatrix} 1 & 1/4 & 1/8 \\ 1/6 & 1 & 1/3 \\ 4/5 & 0 & 1 \end{bmatrix}.$$

You now have to consider

$$\mathbf{B} = \mathbf{I} - \mathbf{A} = \begin{bmatrix} 0 & -1/4 & -1/8 \\ -1/6 & 0 & -1/3 \\ -4/5 & 0 & 0 \end{bmatrix}.$$

The eigenvalues are obtained as the solutions of the characteristic equation

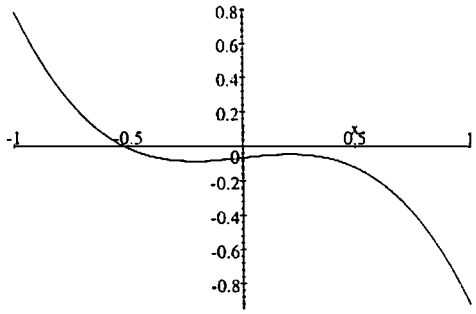
$$\det(\mathbf{B} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & -1/4 & -1/8 \\ -1/6 & -\lambda & -1/3 \\ -4/5 & 0 & -\lambda \end{vmatrix}$$

$$= -\lambda^3 + \frac{17}{120}\lambda - \frac{1}{15} = 0.$$

A plot shows that there is a real root near −0.5, but there are no further real roots because for large $|\lambda|$ the curve comes closer and closer to the curve of $-\lambda^3$. Hence the other eigenvalues must be complex conjugates. A root-finding method (Sec. 17.2) gives a more accurate value −0.5196. Division of the characteristic equation by $\lambda + 0.5196$ gives the quadratic equation

$$-\lambda^2 + 0.5196\lambda - 0.1283 = 0.$$

The roots are $0.2598 \pm 0.2466i$. Since all the roots are less than 1 in absolute value, the spectral radius is less than 1, by definition. This is necessary and sufficient for convergence (see at the end of the section).



Section 18.3. Problem 13. Curve of the characteristic polynomial

15. **Matrix norm.** This simple problem illustrates that the three norms usually tend to give similar values. The same is true with Prob. 19. Hence one often chooses the norm that is most convenient from a computational point of view. See, however, in the next section that a matrix norm often results from a choice of a vector norm, so that in that respect, one is not completely free to choose.

Sec. 18.4 Linear Systems: Ill-conditioning, Norms

Problem Set 18.4. Page 912

7. **Matrix norms and condition numbers.** You have to consider the given matrix

$$\mathbf{A} = \begin{bmatrix} 4 & 1 \\ 0 & 2 \end{bmatrix} \text{ and its inverse } \mathbf{A}^{-1} = \begin{bmatrix} 1/4 & -1/8 \\ 0 & 1/2 \end{bmatrix}.$$

Begin with the l_1-norm. You have to remember that the l_1-vector norm gives for matrices the column "sum" norm (the "..." indicating that we take sums of absolute values). This gives 4 for $\mathbf{A}$ (the first column) and 5/8 for $\mathbf{A}^{-1}$ (the second column). Hence $\kappa(\mathbf{A}) = 4(5/8) = 2.5$. Now turn to the l_∞ norm. You have to remember that this vector norm gives for matrices the row "sum" norm. This gives 5 for $\mathbf{A}$ (the first row) and 1/2 for $\mathbf{A}^{-1}$ (the second row). Hence $\kappa(\mathbf{A}) = 5(1/2) = 2.5$. This is the same value as before. (Is this the case for all triangular real 2×2 matrices? For all real 2×2 matrices? How would you start to experiment on these questions?)

15. **Ill-conditioning.** The given system is

$$4.50x_1 + 3.55x_2 = 5.20$$
$$3.55x_1 + 2.80x_2 = 4.10.$$

Its coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 4.50 & 3.55 \\ 3.55 & 2.80 \end{bmatrix}$$

consists of two almost proportional rows. Indeed, the system is very ill-conditioned. Its column sum norm is $4.50 + 3.55 = 8.05$. Its row sum norm is the same because $\mathbf{A}$ is symmetric. The inverse of $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \begin{bmatrix} -1120 & 1420 \\ 1420 & -1800 \end{bmatrix}$$

and has the column sum norm $1420 + 1800 = 3220$, which equals the row sum norm, again for reasons of symmetry. The product of the norms is the condition number

$$\kappa(A) = 8.05 \cdot 3220 = 25921.$$

This is very large and makes it plausible that the small change from $\mathbf{b}_1$ to $\mathbf{b}_2$ by 0.1 in the second component causes the solution to change from −2,4 to −144,184, a change of about one thousand times that of that component. Also, if you try to solve the system by the Gauss elimination with a small number of decimals, you obtain nonsensical results, whereas for calculations with 8 or 9 decimals the results are satisfactory.

17. **Small residuals for poor solutions.** In the present case, formula (1) with the suggested $\tilde{x} = [a \quad y]^T$ (see p. A40 in Appendix 2 of the book) is

$$\mathbf{r} = \begin{bmatrix} 5.2 \\ 4.1 \end{bmatrix} - \begin{bmatrix} 4.50 & 3.55 \\ 3.55 & 2.80 \end{bmatrix} \begin{bmatrix} a \\ y \end{bmatrix}.$$

In components this can be written

$$r_1 = 5.2 - 4.50a - 3.55y \qquad \text{(A)}$$
$$r_2 = 4.1 - 3.55a - 2.80y.$$

If you set $r_1 = 0, r_2 = 0$, you would get the exact solution because this would be the given system in Prob. 15. Following the suggestion on p. A40, choose an a, say, a large a such as $a = 100$, and solve each equation $r_1 = 0$ and $r_2 = 0$ separately, with a_1 and a_2 as given in (A). You obtain

$$5.2 - 450 - 3.55y = 0, \quad \text{solution } y = \frac{1}{3.55}(5.2 - 450) = -125.296$$
$$4.1 - 355 - 2.80y = 0, \quad \text{solution } y = \frac{1}{2.80}(4.1 - 355) = -125.321.$$

From this you see that you can expect a small residual if you set

$$x_1 = a = 100, \qquad x_2 = y = -125.3.$$

Indeed, you obtain

$$r_1 = 5.2 - 450 - 3.55(-125.3) = 0.015$$
$$r_2 = 4.1 - 355 - 2.8(-125.3) = -0.060.$$

Sec. 18.5 Method of Least Squares

Problem Set 18.5. Page 916

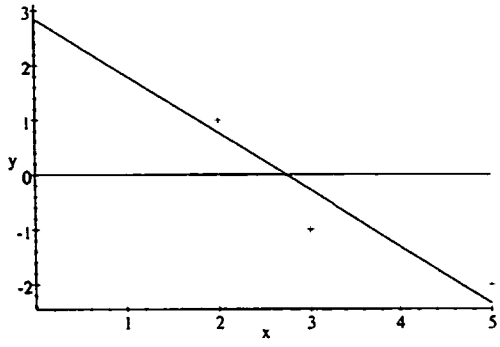
1. **Fitting by a straight line.** As in Example 1 in the text, the straight line for fitting the points by the method of least squares is obtained by solving the normal equations (4). In working with paper and pencil, it is best to first make up an orderly table of the given data and auxiliary quantities needed in (4). This may look as follows.

	x_j	y_j	x_j^2	$x_j y_j$
	0	3	0	0
	2	1	4	2
	3	−1	9	−3
	5	−2	25	−10
Sum	10	1	38	−11

Since you have $n = 4$ pairs of values, with the sums in your table the augmented matrix of (4) has the form

$$\begin{bmatrix} 4 & 10 & 1 \\ 10 & 38 & -11 \end{bmatrix}.$$

The solution is $a = 37/13 = 2.84615, b = -27/26 = -1.03846$. Hence the desired straight line is $y = 2.84615 - 1.03846x$.



Section 18.6. Problem 1. Given data and straight line fitted by least squares

11. **Fitting by a quadratic parabola.** A quadratic parabola is uniquely determined by three given points. In this problem, five points are given. You can fit a quadratic parabola by solving the normal equations (8). Arrange the data and auxiliary quantities in (8) again in a table.

	x	y	x^2	x^3	x^4	xy	x^2y
	2	0	4	8	16	0	0
	3	3	9	27	81	9	27
	5	4	25	125	625	20	100
	6	3	36	216	1296	18	108
	7	1	49	343	2401	7	49
Sum	23	11	123	719	4419	54	284

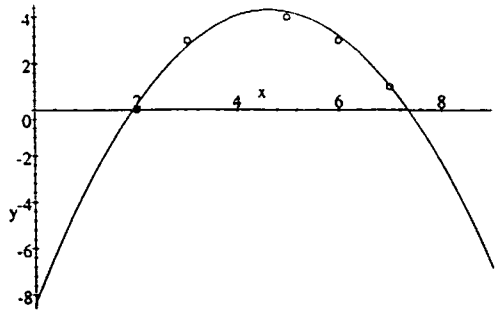
Hence the augmented matrix of the system of normal equations is

$$\begin{bmatrix} 5 & 23 & 123 & 11 \\ 23 & 123 & 719 & 54 \\ 123 & 719 & 4419 & 284 \end{bmatrix}.$$

The solution obtained, for instance, by Gauss elimination is

$$b_0 = -8.357, \quad b_1 = 5.446, \quad b_2 = -0.589.$$

Hence the desired quadratic parabola that fits the data by the least squares principle is $y = -8.357 + 5.446x - 0.589x^2$.



Section 18.6. **Problem 11.** Given points and quadratic parabola fitted by least squares

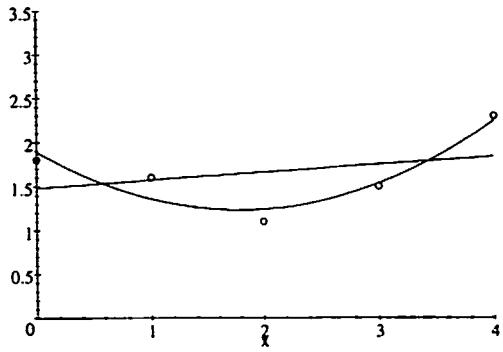
13. **Comparison of linear and quadratic fit.** The figure shows that a straight line obviously is not sufficient. The quadratic parabola gives a much better fit. It depends on the physical or other law underlying the data whether the fit by a quadratic polynomial is satisfactory and whether the remaining discrepancies can be attributed to chance variations, such as inaccuracy of measurement. Calculation shows that the augmented matrix of the normal equations for the straight line is

$$\begin{bmatrix} 5 & 10 & 8.3 \\ 10 & 30 & 17.5 \end{bmatrix}$$

and gives $y = 1.48 + 0.09x$. The augmented matrix for the quadratic polynomial is

$$\begin{bmatrix} 5 & 10 & 30 & 8.3 \\ 10 & 30 & 100 & 17.5 \\ 30 & 100 & 354 & 56.31 \end{bmatrix}$$

and gives $y = 1.896 - 0.741x + 0.208x^2$.



Section 18.6. **Problem 13.** Fit by a straight line and by a quadratic parabola

Sec. 18.7 Inclusion of Matrix Eigenvalues

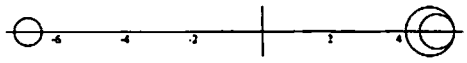
Problem Set 18.7. Page 924

1. **Gerschgorin circles.** Gerschgorin's theorem is one of the earliest theorems on the numerical determination of matrix eigenvalues. The application of the theorem to a real or complex square matrix is very simple. In the present problem the Gerschgorin disks have the centers 5.1, 4.9, and –6.8 and the radii 0.5, 0.7, and 0.4, respectively. These disks consist of two disjoint parts (see the figure). The right part is

formed by the union of two disks; hence it contains two eigenvalues. The left part, consisting of a single disk, must contain a single eigenvalue. This follows from Theorem 2. Note that the matrix is *not* skew-symmetric because its main diagonal entries are not zero. Hence you cannot apply Theorem 5 in Sec. 18.6 and conclude that its eigenvalues are pure imaginary or zero. 3D-values of the eigenvalues are -6.791 and $4.996 \pm 0.387i$. These can be determined, for instance, by sketching or plotting the curve of the characteristic polynomial, which by developing the characteristic determinant is found to be

$$-\lambda^3 + 3.2\lambda^2 + 42.75\lambda - 170.512.$$

The curve intersects the x-axis only once, near $\lambda = -7$, a value that can be improved to -6.791 by Newton's method. Then division by $\lambda + 6.791$ gives a quadratic equation whose roots are complex conjugates, as given before. You see that the left Gerschgorin disk does contain just one eigenvalue, whereas the other two eigenvalues lie in the union of the other two Gerschgorin disks with centers at 5.1 and 4.9.



Section 18.7. Problem 1. Gerschgorin circles

7. **Similarity transformation.** The matrix in Prob. 3 shows a typical situation. It may have resulted from a numerical method of diagonalization which left off-diagonal entries of various sizes but not exceeding 10^{-2} in absolute value. Gerschgorin's theorem then gives circles of radius 2×10^{-2}. These furnish upper bounds for the deviation of the eigenvalues from the main diagonal entries. This describes the starting situation for the present problem. Now in various applications, one is often interested in the eigenvalue of largest or smallest absolute value. In your matrix, the smallest eigenvalue is about 5, with a maximum possible deviation of 2×10^{-2}, as given by Gerschgorin's theorem. You now wish to decrease the size of this Gerschgorin disk as much as possible. Example 2 in the text shows how you should proceed. The entry 5 stands in the first row and column. Hence you should apply to **A** a similarity transformation involving a diagonal matrix **T** with main diagonal $a, 1, 1$, where a is as large as possible. The inverse of **T** is the diagonal matrix with main diagonal $1/a, 1, 1$. Leave a arbitrary and first determine the result of the similarity transformation (as in Example 2)

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 1/a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 5 & 0.01 & 0.01 \\ 0.01 & 8 & 0.01 \\ 0.01 & 0.01 & 9 \end{bmatrix} \begin{bmatrix} a & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 5 & 0.01/a & 0.01/a \\ 0.01a & 8 & 0.01 \\ 0.01a & 0.01 & 9 \end{bmatrix}.$$

You see that the Gerschgorin disks of the transformed matrix **B** are

Center	Radius
5	$0.02/a$
8	$0.01(a+1)$
9	$0.01(a+1)$

The last two disks must be small enough so that they do not touch or even overlap the first disk. Since $8 - 5 = 3$, the radius of the second disk after the transformation must be less than $3 - 0.02/a$, that is,

$$0.01(a+1) < 3 - 0.02/a.$$

Multiplication by $100a$ (> 0) gives $a^2 + a < 300a - 2$. If you replace the inequality sign by an equality sign, you obtain the quadratic equation $a^2 - 299a + 2 = 0$. Hence a must be less than the larger root 298.9933 of this equation, say, for convenience, $a = 298$. Then the radius of the second disk is $0.01(a+1) = 2.99$, so that the disk will not touch the first one, and neither will the third, which is farther away from the first. The first disk is substantially reduced in size, by a factor of almost 300, the radius of the reduced disk being

$$0.02/298 = 0.000067114.$$

The choice of $a = 100$ would give a reduction by a factor 100, as requested in the problem. Your systematic approach shows that you can do better.

11. Spectral radius. By definition, the spectral radius of a square matrix **A** is the absolute value of an eigenvalue of **A** that is largest in absolute value. Since every eigenvalue of **A** lies in a Gerschgorin disk, for every eigenvalue of **A** you must have (make a sketch)

$$|a_{jj}| + \sum |a_{jk}| \geqq |\lambda_j|$$

where you sum over all off-diagonal entries in Row j (and the eigenvalues of **A** are numbered suitably). By taking this for a largest $|\lambda_j|$, you accomplish two things. First, on the right you obtain the spectral radius. Second, on the left you obtain the row "sum" norm. This proves the statement.

19. Collatz's theorem. The matrix **A** has equal row sums 7; hence 7 must be an eigenvalue of **A**. The other eigenvalue of **A** is 4; it has algebraic multiplicity 2, that is, it is a double root of the characteristic equation. In the present case the characteristic equation can be solved by subtracting Row 2 of the characteristic matrix from Row 1, then Row 3 from Row 2, obtaining

$$\begin{bmatrix} 4-\lambda & -4+\lambda & 0 \\ 0 & 4-\lambda & -4+\lambda \\ 1 & 1 & 5-\lambda \end{bmatrix}.$$

Then add Column 1 to Column 2 and develop the determinant of this matrix by the first row (which now contains two zeros); this gives

$$\begin{vmatrix} 4-\lambda & 0 & 0 \\ 0 & 4-\lambda & -4+\lambda \\ 1 & 2 & 5-\lambda \end{vmatrix} = (4-\lambda)[(4-\lambda)(5-\lambda) + 2(4-\lambda)] = (4-\lambda)^2[5-\lambda+2].$$

This shows that the eigenvalues are 7 and 4, as claimed, and 4 has the algebraic multiplicity 2.

Sec. 18.8 Eigenvalues by Iteration (Power Method)

Example 1. Six vectors are listed. The first was scaled. The others were obtained by multiplication by **A** and subsequent scaling. You can use any of these vectors for obtaining a corresponding Rayleigh quotient q as an approximate value of an (unknown) eigenvalue of **A** and a corresponding error bound δ for q. Hence you have six possibilities using one of the given vectors (and many more if you want to compute further vectors). You must not use two of the given vectors because of the scaling, but just one vector, for instance, $\mathbf{x}_1$, and then its product $\mathbf{A}\mathbf{x}_1$. That is,

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.890244 \\ 0.609756 \\ 1 \end{bmatrix}, \quad \mathbf{A}\mathbf{x}_1 = \begin{bmatrix} 0.668415 \\ 0.388537 \\ 0.717805 \end{bmatrix}.$$

From these data you calculate the inner products

$$m_0 = \mathbf{x}_1^T\mathbf{x}_1 = 2.164337$$
$$m_1 = \mathbf{x}_1^T\mathbf{A}\mathbf{x}_1 = 1.549770$$
$$m_2 = (\mathbf{A}\mathbf{x}_1)^T\mathbf{A}\mathbf{x}_1 = 1.112983.$$

These now give the Rayleigh quotient q and error bound δ of q

$$q = m_1/m_0 = 0.716048$$
$$\delta = \sqrt{m_2/m_0 - q^2} = 0.038887\ .$$

q approximates the eigenvalue 0.72 of $\mathbf{A}$, so that the error of q is

$$\epsilon = 0.72 - q = 0.003952.$$

These values agree with those in the table on p. 928 of the book.

Problem Set 18.8. Page 928

1. **Power method without scaling.** Without scaling the components of the vectors successively obtained will generally keep growing (or decreasing). Since the given matrix is symmetric, you can apply Theorem 1, which yields error bounds for the approximations of λ (usually the largest eigenvalue in absolute value, but no general statements can be made). Our simple matrix has the eigenvalues 11 and 1, as can readily be computed, and the problem serves only to explain the method in a very simple case in which you can see what is going on in each step. The computation of the vectors needed gives the following results.

$\mathbf{A}$	$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$
$\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 1 \end{bmatrix}$	$\begin{bmatrix} 13 \\ 7 \end{bmatrix}$	$\begin{bmatrix} 145 \\ 73 \end{bmatrix}$	$\begin{bmatrix} 1597 \\ 799 \end{bmatrix}$

In each step of the further calculations take two adjacent vectors $\mathbf{x}$ and $\mathbf{y} = \mathbf{A}\mathbf{x}$, that is, $\mathbf{x}_0$ and $\mathbf{x}_1$, then $\mathbf{x}_1$ and $\mathbf{x}_2$, then $\mathbf{x}_2$ and $\mathbf{x}_3$. For the first of these three pairs $\mathbf{x}$, $\mathbf{y}$, namely, for $\mathbf{x}_0$ and $\mathbf{x}_1$, now compute (see Theorem 1)

$$m_0 = \mathbf{x}_0^T\mathbf{x}_0 = 1^2 + 1^2 = 2,$$
$$m_1 = \mathbf{x}_0^T\mathbf{x}_1 = 1\cdot 13 + 1\cdot 7 = 20$$
$$m_2 = \mathbf{x}_1^T\mathbf{x}_1 = 13^2 + 7^2 = 218$$
$$q = m_1/m_0 = 20/2 = 10 \qquad \text{(approximation of an eigenvalue)}$$
$$\delta^2 = m_2/m_0 - q^2 = 218/2 - 10^2 = 9$$
$$\delta = 3 \qquad \text{(error bound; see Theorem 1).}$$

Since $\mathbf{A}$ is symmetric, its eigenvalues are real. Hence conclude from Theorem 1 that an eigenvalue of $\mathbf{A}$ must lie in the closed interval

$$q - \delta = 7 \leqq \lambda \leqq q + \delta = 13.$$

It is typical that the error bound is much larger than the actual error, but the interval $q - \delta \leqq \lambda \leqq q + \delta$ is best possible, that is, for values q and δ calculated from a given symmetric matrix $\mathbf{A}$ (of any size) there is a symmetric matrix $\mathbf{B}$ for which the endpoints of that interval are eigenvalues of $\mathbf{B}$. (Another question would be how to actually find such a matrix $\mathbf{B}$ in a concrete case. Problem 10 contributes to this in a very special case.) For the next pair $\mathbf{x}_1$ and $\mathbf{x}_2$ you obtain

$$m_0 = \mathbf{x}_1^T\mathbf{x}_1 = 13^2 + 7^2 = 218 \qquad \text{(this is } m_2 \text{ of the previous step)}$$
$$m_1 = \mathbf{x}_1^T\mathbf{x}_2 = 13\cdot 145 + 7\cdot 73 = 2396$$
$$m_2 = \mathbf{x}_2^T\mathbf{x}_2 = 145^2 + 73^2 = 26354$$
$$q = m_1/m_0 = 10.99083 \qquad \text{(approximation of an eigenvalue)}$$
$$\delta^2 = m_2/m_0 - q^2 = 0.091659$$
$$\delta = 0.302752 \qquad \text{(error bound for } q\text{)}.$$

From this and Theorem 1 conclude that an eigenvalue of **A** must lie in the interval

$$q - \delta = 10.68807 \leqq \lambda \leqq q + \delta = 11.29358.$$

You see that the approximation has substantially increased in quality. The same is true for the error bound, which is smaller than in the first step by a factor 10. But, again, the error is much smaller than the bound, so that the latter does not give an indication of the size of the error. In the third step compute

$$m_0 = \mathbf{x}_2^T\mathbf{x}_2 = 26354 \qquad \text{(this is } m_2 \text{ of the previous step)}$$
$$m_1 = \mathbf{x}_2^T\mathbf{x}_3 = 145\cdot 1597 + 73\cdot 799 = 289892$$
$$m_2 = \mathbf{x}_3^T\mathbf{x}_3 = 1597^2 + 799^2 = 3188810$$
$$q = m_1/m_0 = 10.99992 \qquad \text{(approximation of an eigenvalue)}$$
$$\delta^2 = m_2/m_0 - q^2 = 0.0007589$$
$$\delta = 0.027548 \qquad \text{(error bound for } q\text{)}.$$

From this and Theorem 1 conclude that the interval

$$q - \delta = 10.97237 \leqq \lambda \leqq q + \delta = 11.02747$$

must contain an eigenvalue of **A**. Again, the error 0.00008 of q is much smaller than the error bound 0.027548.

13. Power method with scaling. The given matrix is

$$\mathbf{A} = \begin{bmatrix} 3.6 & -1.8 & 1.8 \\ -1.8 & 2.8 & -2.6 \\ 1.8 & -2.6 & 2.8 \end{bmatrix}.$$

Use the same notation as in Example 1 in the text. From $\mathbf{x}_0 = [1 \quad 1 \quad 1]^T$ calculate $\mathbf{Ax}_0$ and then scale it as indiciated in the problem, calling the resulting vector $\mathbf{x}_1$. This is the first step. In the second step calculate $\mathbf{Ax}_1$ and then scale it, calling the resulting vector $\mathbf{x}_2$. And so on. The numerical results are as follows.

$$\mathbf{x}_0 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{Ax}_0 \begin{bmatrix} 3.6 \\ -1.6 \\ 2.0 \end{bmatrix} \qquad \mathbf{x}_1 \begin{bmatrix} 1 \\ -0.444444 \\ 0.555556 \end{bmatrix} \qquad \mathbf{Ax}_1 \begin{bmatrix} 5.4 \\ -4.488889 \\ 4.511111 \end{bmatrix}$$

$$\mathbf{x}_2 \begin{bmatrix} 1 \\ -0.831276 \\ 0.835391 \end{bmatrix} \qquad \mathbf{Ax}_2 \begin{bmatrix} 6.6 \\ -6.299588 \\ 6.300412 \end{bmatrix} \qquad \mathbf{x}_3 \begin{bmatrix} 1 \\ -0.954483 \\ 0.954608 \end{bmatrix}$$

The calculation of approximations q (Rayleigh quotients) and error bounds δ proceeds similarly to the method without scaling (see the previous problem in this Manual). However, in the first step use

$$x = x_0 \quad \text{and} \quad y = Ax = Ax_0 \quad (\text{not } x_1).$$

In the second step use

$$x = x_1 \quad \text{and} \quad y = Ax = Ax_1 \quad (\text{not } x_2).$$

In the third step use

$$x = x_2 \quad \text{and} \quad y = Ax = Ax_2 \quad (\text{not } x_3).$$

The calculations that give approximations q (Rayleigh quotients) and error bounds are as follows.

m_0	$x_0^T x_0 = 3$	$x_1^T x_1 = 1.506173$	$x_2^T x_2 = 2.388897$
m_1	$x_0^T A x_0 = 4$	$x_1^T A x_1 = 9.901235$	$x_2^T A x_2 = 17.100002$
m_2	$(Ax_0)^T Ax_0 = 19.52$	$(Ax_1)^T Ax_1 = 69.660247$	$(Ax_2)^T Ax_2 = 122.94000$
$q = m_1/m_0$	1.333333	6.573770	7.158115
$\delta^2 = m_2/m_0 - q^2$	4.728889	3.035378	0.224465
δ	2.174601	1.742234	0.47377
$q - \delta$	−0.841267	4.831536	6.684337
$q + \delta$	3.507935	8.316004	7.631893

Solving the characteristic equation shows that the matrix has the eigenvalues 0.2, 1.8, and 7.2. Corresponding eigenvectors are

$$z_1 = [0 \quad 1 \quad 1]^T, \quad z_2 = [2 \quad 1 \quad -1]^T, \quad z_3 = [1 \quad -1 \quad 1]^T,$$

respectively. You see that the interval obtained in the first step includes the eigenvalues 0.2 and 1.8. Only in the second step and third step of the iteration did you obtain intervals that include the largest eigenvalue, as is usually the case from the beginning on. The reason for this interesting observation is the fact that x_0 is a linear combination of all three eigenvectors,

$$x_0 = z_1 + \frac{1}{3}(z_2 + z_3),$$

as can be easily verified, and its needs several iterations until the powers of the largest eigenvalue make the iterate x_j come close to z_3, the eigenvector corresponding to $\lambda = 7.2$. This siutation occurs quite frequently, and and one needs the more steps for obtaining satisfactory results the closer in absolute value the other eigenvalues are to the absolutely largest one. See also Prob. 11 for some further explanation.

Sec. 18.9 Tridiagonalization and QR-Factorization

Example 2. The tridiagonalized matrix is (p.936)

$$B = \begin{bmatrix} 6 & -\sqrt{18} & 0 \\ -\sqrt{18} & 7 & \sqrt{2} \\ 0 & \sqrt{2} & 6 \end{bmatrix}.$$

We use the abbreviations c_2, s_2, and t_2 for $\cos\theta_2$, $\sin\theta_2$, and $\tan\theta_2$, respectively. We multiply **B** from the left by

$$C_2 = \begin{bmatrix} c_2 & -s_2 & 0 \\ -s_2 & c_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

The purpose of this multiplication is to obtain a matrix $C_2B = [b_{jk}^{(2)}]$ for which the off-diagonal entry $b_{21}^{(2)}$ is zero. Now this entry is the inner product of Row 2 of C_2 times Column 1 of **B**, that is,

$$-s_2 \cdot 6 + c_2(-\sqrt{18}) = 0, \quad \text{thus} \quad t_2 = -\sqrt{18}/6 = -\sqrt{1/2}.$$

From this and the formulas that express cos and sin in terms of tan we obtain

$$c_2 = 1/\sqrt{1+t_2^2} = \sqrt{2/3} = 0.816496581,$$

$$s_2 = t_2/\sqrt{1+t_2^2} = -\sqrt{1/3} = -0.577350269.$$

θ_3 is determined similarly, with the purpose of obtaining $b_{32}^{(3)} = 0$ in $\mathbf{C}_3\mathbf{C}_2\mathbf{B} = [b_{jk}^{(3)}]$.

Problem Set 18.9. Page 937

1. Tridiagonalization. The given matrix

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}$$

is symmetric. Hence you can apply Householder's method for obtaining a tridiagonal matrix (which will have two zeros instead of the entries 0.22). Proceed as in Example 1 of the text. Since $\mathbf{A}$ is of size $n = 3$, you have to perform $n - 2 = 1$ step. (In Example 1 we had $n = 4$ and needed $n - 2 = 2$ steps.) Calculate the vector $\mathbf{v}_1$ from (4). Denote it simply by $\mathbf{v}$ and its components by $v_1 (= 0)$, v_2, v_3 because you do only one step. Similarly, denote S_1 in (4c) by S. Compute

$$S = \sqrt{a_{21}^2 + a_{31}^2} = \sqrt{0.02^2 + 0.22^2} = 0.2209072203.$$

If you compute, using, say, 6 digits, you may expect that instead of those two zeros in the tridiagonalized matrix you obtain entries of the order 10^{-6} or even larger in absolute value. You always have $v_1 = 0$. From (4a) you obtain the second component

$$v_2 = \sqrt{\frac{1 + a_{21}/S}{2}} = \sqrt{\frac{1 + 0.02/0.2209072203}{2}} = 0.7384225572.$$

From (4b) with $j = 3$ and sgn $a_{21} = +1$ (because a_{21} is positive) you obtain the third component

$$v_3 = a_{31}/(2v_2S) = 0.22/(2v_2S) = 0.6743382884.$$

With these values you now compute $\mathbf{P}_r$ from (2), where $r = 1, \cdots, n-2$, so that you have only $r = 1$ and can denote $\mathbf{P}_1$ simply by $\mathbf{P}$. Note well that $\mathbf{v}^T\mathbf{v}$ would be the dot product of the vector by itself (thus the square of its length), whereas $\mathbf{v}\mathbf{v}^T$ is a 3×3 matrix because of the usual matrix multiplication. You thus obtain from (2)

$$\mathbf{P} = \mathbf{I} - 2\mathbf{v}\mathbf{v}^T$$

$$= \mathbf{I} - 2\begin{bmatrix} v_1^2 & v_1v_2 & v_1v_3 \\ v_2v_1 & v_2^2 & v_2v_3 \\ v_3v_1 & v_3v_2 & v_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 - 2v_1^2 & -2v_1v_2 & -2v_1v_3 \\ -2v_2v_1 & 1 - 2v_2^2 & -2v_2v_3 \\ -2v_3v_1 & -2v_3v_2 & 1 - 2v_3^2 \end{bmatrix}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & -0.0905357460 & -0.9958932066 \\ 0 & -0.9958932064 & 0.0905357456 \end{bmatrix}.$$

Finally use $\mathbf{P}$ and its inverse $\mathbf{P}^{-1} = \mathbf{P}$ for the similarity transformation that will produce the tridiagonal matrix

$$\mathbf{B} = \mathbf{PAP} = \mathbf{P}\begin{bmatrix} 0.49 & -0.2209072204 & -10^{-10} \\ 0.02 & -0.2245286502 & -0.2607429487 \\ 0.22 & -0.4164644318 & -0.1629643431 \end{bmatrix}$$

$$= \begin{bmatrix} 0.49 & -0.2209072204 & -10^{-10} \\ -0.2209072204 & 0.4350819672 & 0.1859016396 \\ -10^{-10} & 0.1859016396 & 0.2449180330 \end{bmatrix}.$$

The point of the use of similarity transformations is that they preserve the spectrum of **A**, consisting of the eigenvalues

$$0.09, \quad 0.36, \quad 0.72,$$

which can be found, for instance, by plotting the characteristic polynomial of **A** and applying Newton's method for improving the values obtained.

9. **QR-factorization.** The purpose of this factorization is the determination of approximate values of all the eigenvalues of a given matrix. To save work, one usually begins by tridiagonalizing a given matrix, which must be symmetric. The given matrix

$$\mathbf{B}_0 = [b_{jk}] = \begin{bmatrix} 7.0 & 0.5 & 0 \\ 0.5 & 3.5 & 0.1 \\ 0 & 0.1 & -1.5 \end{bmatrix}$$

is tridiagonal. Hence QR can begin. Proceed as in Example 2 on p. 935 of the book. (See also in this Manual above.) Write c_2, s_2, t_2 for $\cos\theta_2$, $\sin\theta_2$, $\tan\theta_2$, respectively. Consider the matrix

$$\mathbf{C}_2 = \begin{bmatrix} c_2 & s_2 & 0 \\ -s_2 & c_2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

with the angle of rotation θ_2 determined so that in the product $\mathbf{W}_0 = \mathbf{C}_2\mathbf{B}_0 = [w_{jk}]$ the entry w_{21} is zero. By the usual matrix multiplication (row times column) w_{21} is the inner product of Row 2 of $\mathbf{C}_2$ times Column 1 of $\mathbf{B}_0$, that is,

$$-s_2 b_{11} + c_2 b_{21} = 0, \quad \text{hence} \quad t_2 = s_2/c_2 = b_{21}/b_{11}.$$

From this and the formulas for cos and sin in terms of tan (usually discussed in calculus) you obtain

$$c_2 = 1/\sqrt{1 + (b_{21}/b_{11})^2} = 0.9974586997 \qquad \text{(I/1)}$$

$$s_2 = \frac{b_{21}}{b_{11}}/\sqrt{1 + (b_{21}/b_{11})^2} = 0.0712470500.$$

Now calculate $\mathbf{C}_2\mathbf{B}_0$ (as in the middle of p. 936). Denote this matrix (which has no notation on p. 936) by $\mathbf{W} = [w_{jk}]$. Thus

$$\mathbf{W} = [w_{jk}] = \mathbf{C}_2\mathbf{B}_0 = \begin{bmatrix} 7.017834423 & 0.7480940249 & 0.0071247050 \\ 0 & 3.455481924 & 0.09974587000 \\ 0 & 0.1 & -1.5 \end{bmatrix}.$$

C_2 has served its purpose: instead of $b_{21} = 0.5$ you now have $w_{21} = 0$. (Instead of $w_{21} = 0$ on the computer you may get -10^{-10} or another very small entry.) Now use the abbreviations c_3, s_3, t_3 for $\cos\theta_3$, $\sin\theta_3$, $\tan\theta_3$. Consider the matrix

$$\mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & c_3 & s_3 \\ 0 & -s_3 & c_3 \end{bmatrix}$$

with the angle of rotation θ_3 such that the product matrix $\mathbf{R}_0 = [r_{jk}] = \mathbf{C}_3\mathbf{W} = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_0$ the entry r_{32} is zero. This entry is the inner product of Row 3 of $\mathbf{C}_3$ times Column 2 of W. Hence

$$-s_3 w_{22} + c_3 w_{32} = 0, \quad \text{so that} \quad t_3 = s_3/c_3 = w_{32}/w_{22} = 0.02893952340.$$

This gives for c_3 and s_3

$$c_3 = 1/\sqrt{1 + t_3^2} = 0.9995815152, \qquad s_3 = t_3/\sqrt{1 + t_3^2} = 0.0289274127. \tag{II/1}$$

Using this, you obtain

$$\mathbf{R}_0 = \mathbf{C}_3\mathbf{W} = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_0 = \begin{bmatrix} 7.017834423 & 0.7480940249 & 0.0071247050 \\ 0 & 3.456928598 & 0.0563130089 \\ 0 & 0 & -1.502257663 \end{bmatrix}.$$

(Instead of 0 you may obtain 10^{-10} or another very small term. Similarly in the further calculations.) Finally, multiply $\mathbf{R}_0$ from the right by $\mathbf{C}_2^T\mathbf{C}_3^T$. This gives

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{C}_2^T\mathbf{C}_3^T = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_0\mathbf{C}_2^T\mathbf{C}_3^T$$

$$= \begin{bmatrix} 7.053299490 & 0.2462959646 & 0 \\ 0.2462959646 & 3.448329498 & -0.0434564273 \\ 0 & -0.0434564273 & -1.501628991 \end{bmatrix}.$$

The given matrix $\mathbf{B}_0$ (and thus, also the matrix $\mathbf{B}_1$) has the eigenvalues

$$7.070049927, \quad 3.431961091, \quad -1.502011018.$$

You see that the main diagonal entries of $\mathbf{B}_1$ are approximations that are not very accurate, a fact that you could have concluded from the relatively large size of the off-diagonal entries of $\mathbf{B}_1$. In practice, one would perform further steps of the iteration until all off-diagonal elements have decreased in absolute value to less than a given bound. The answer on p. A41 in Appendix 2 gives the results of two more steps, which are obtained by the following calculations.

Step 2. The calculations are the same as before, with $\mathbf{B}_0 = [b_{jk}]$ replaced by $\mathbf{B}_1 = [b_{jk}^{(1)}]$. Hence, instead of (I/1) you now have

$$c_2 = 1/\sqrt{1 + (b_{21}^{(1)}/b_{11}^{(1)})^2} = 0.9993908803 \tag{I/2}$$

$$s_2 = (b_{21}^{(1)}/b_{11}^{(1)})/\sqrt{1 + (b_{21}^{(1)}/b_{11}^{(1)})^2} = 0.0348979852.$$

You can now write the matrix $\mathbf{C}_2$, which has the same general form as before, and calculate the product

$$\mathbf{W}_1 = [w_{jk}^{(1)}] = \mathbf{C}_2\mathbf{B}_1$$

$$= \begin{bmatrix} 7.057598419 & 0.3664856925 & -0.0015165418 \\ 0 & 3.437633820 & -0.0434299571 \\ 0 & -0.0434564273 & -1.501628991 \end{bmatrix}.$$

Now calculate the entries of $\mathbf{C}_3$ from (II/1) with $t_3 = w_{32}/w_{22}$ replaced by $t_3^{(1)} = w_{32}^{(1)}/w_{22}^{(1)}$, that is,

$$c_3 = 1/\sqrt{1 + (t_3^{(1)})^2} = 0.9999201074 \tag{II/2}$$

$$s_3 = t_3^{(1)}/\sqrt{1 + (t_3^{(1)})^2} = -0.0126403677.$$

You can now write $\mathbf{C}_3$, which has the same general form as in Step 1, and calculate

$$\mathbf{R}_1 = \mathbf{C}_3\mathbf{W}_1 = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_1$$

$$= \begin{bmatrix} 7.057598419 & 0.3664856925 & -0.0015165418 \\ 0 & 3.437908483 & -0.0244453448 \\ 0 & 0 & -1.502057993 \end{bmatrix}.$$

This gives the next result

$$\mathbf{B}_2 = [b_{jk}^{(2)}] = \mathbf{R}_1\mathbf{C}_2^T\mathbf{C}_3^T = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_1\mathbf{C}_2^T\mathbf{C}_3^T$$

$$= \begin{bmatrix} 7.066089109 & 0.1199760792 & 0 \\ 0.1199760792 & 3.4358484889 & 0.0189865653 \\ 0 & 0.0189865653 & -1.501937990 \end{bmatrix}.$$

The approximations of the eigenvalues have improved. The off-diagonal entries are smaller than in $\mathbf{B}_1$. Nevertheless, in practice the accuracy would still not be sufficient, so that one would do several more steps. Do one more step, whose result is also given on p. A41 in Appendix 2 of the book.

Step 3. The calculations are the same as in Step 2, with $\mathbf{B}_1 = [b_{jk}^{(1)}]$ replaced by $\mathbf{B}_2 = [b_{jk}^{(2)}]$. Hence calculate the entries of $\mathbf{C}_2$ from

$$c_2 = 1/\sqrt{1 + (b_{21}^{(2)}/b_{11}^{(2)})^2} = 0.9998558858 \qquad \text{(I/3)}$$

$$s_2 = (b_{21}^{(2)}/b_{11}^{(2)})/\sqrt{1 + (b_{21}^{(2)}/b_{11}^{(2)})^2} = 0.0169766878.$$

You can now write the matrix $\mathbf{C}_2$ and calculate the product

$$\mathbf{W}_2 = [w_{jk}^{(2)}] = \mathbf{C}_2\mathbf{B}_2$$

$$= \begin{bmatrix} 7.067107581 & 0.1782881229 & 0.0003223290 \\ 0 & 3.433316936 & 0.0189838291 \\ 0 & 0.0189865653 & -1.501937990 \end{bmatrix}.$$

Now calculate the entries of $\mathbf{C}_3$ from (II/2) with $t_2^{(1)}$ replaced by $t_3^{(2)} = w_{32}^{(2)}/w_{32}^{(2)}$, that is,

$$c_3 = 1/\sqrt{1 + (t_3^{(2)})^2} = 0.9999847092 \qquad \text{(II/3)}$$

$$s_3 = t_3^{(2)}/\sqrt{1 + (t_3^{(2)})^2} = 0.0055300094.$$

Write $\mathbf{C}_3$ and calculate

$$\mathbf{R}_2 = \mathbf{C}_3\mathbf{W}_2 = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_2$$

$$= \begin{bmatrix} 7.067107581 & 0.1782881229 & 0.0003223290 \\ 0 & 3.433369434 & 0.0106778076 \\ 0 & 0 & -1.502020005 \end{bmatrix}$$

and, finally,

$$\mathbf{B}_3 = \mathbf{R}_2\mathbf{C}_2^T\mathbf{C}_3^T = \mathbf{C}_3\mathbf{C}_2\mathbf{B}_2\mathbf{C}_2^T\mathbf{C}_3^T$$

$$= \begin{bmatrix} 7.069115852 & 0.0582872409 & 0 \\ 0.0582872409 & 3.432881194 & -0.0083061848 \\ 0 & -0.0083061848 & -1.501997039 \end{bmatrix}.$$

This is again an improvement over the result of Step 2.