APPENDIX O

NUMERICAL SOLUTION OF LINEAR EQUATIONS

- From B. Carnahan, H.A. Luther and J.O. Wilkes,
  Applied Numerical Methods
  John Wiley, 1969.
  (now out of print)

- Discussos:
  - Gauss elimination
  - Jacobi iteration
  - Gauss - Seidel iteration

CHAPTER 5

Systems of Equations

5.1 Introduction

This chapter is concerned with methods for solving the following system of n simultaneous equations in the n unknowns $x_1, x_2, \ldots, x_n$:

$$\begin{aligned} f_1(x_1,x_2,\ldots,x_n) &= 0,\\ f_2(x_1,x_2,\ldots,x_n) &= 0,\\ \vdots \qquad & \quad \vdots\\ f_n(x_1,x_2,\ldots,x_n) &= 0. \end{aligned} \tag{5.1}$$

The general case, in which the functions $f_1, f_2, \ldots, f_n$ do not admit of any particular simplification, is treated in Sections 5.8 and 5.9. However, if these functions are linear in the x's, (5.1) can be rewritten as:

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + \cdots + b_{1n}x_n &= u_1,\\ b_{21}x_1 + b_{22}x_2 + \cdots + b_{2n}x_n &= u_2,\\ \vdots \qquad\qquad\qquad & \quad \vdots\\ b_{n1}x_1 + b_{n2}x_2 + \cdots + b_{nn}x_n &= u_n. \end{aligned} \tag{5.2}$$

More concisely, we have

$$\mathbf{Bx} = \mathbf{u}, \tag{5.3}$$

in which $\mathbf{B}$ is the matrix of coefficients, $\mathbf{u} = [u_1,u_2,\ldots,u_n]^t$ is the right-hand side vector, and $\mathbf{x} = [x_1,x_2,\ldots,x_n]^t$ is the solution vector. Assuming negligible computational round-off error, *direct* methods for solving (5.2) exactly, in a finite number of operations, are discussed in Sections 5.3, 5.4, and 5.5. These direct techniques are useful when the number of equations involved is not too large (typically of the order of 40 or fewer equations). *Iterative* methods for solving (5.2) approximately are described in Sections 5.6 and 5.7. These iterative techniques are more appropriate when dealing with a large number of simultaneous equations (typically of the order of 100 equations or more), which will often possess certain other special characteristics.

5.2 Elementary Transformations of Matrices

Before studying systems of equations, it is useful to consider the three types of *elementary matrices*:

1. An elementary matrix of the *first kind* is an $n \times n$ diagonal matrix $\mathbf{Q}$, formed by taking the identity matrix $\mathbf{I}$ and replacing the ith diagonal element with a nonzero constant q. For example, with $n = 4$ and $i = 3$,

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & q & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that $\det \mathbf{Q} = q$, and that the inverse matrix $\mathbf{Q}^{-1} = \operatorname{diag}(1, 1, 1/q, 1)$ is again like $\mathbf{I}$, this time with $1/q$ in the ith diagonal position.

2. An elementary matrix of the *second kind* is an $n \times n$ matrix $\mathbf{R}$, formed by interchanging any two rows i and j of $\mathbf{I}$. For example, with $n = 4$, $i = 1$, and $j = 3$,

$$\mathbf{R} = \begin{bmatrix} 0 & 0 & 1 & 0\\ 0 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that $\det \mathbf{R} = -1$, and that $\mathbf{R}$ is self-inverse, that is, $\mathbf{RR} = \mathbf{I}$.

3. An elementary matrix of the *third kind* is an $n \times n$ matrix $\mathbf{S}$, formed by inserting a nonzero constant s into the i,j $(i \neq j)$ element of $\mathbf{I}$. (This may also be construed as taking $\mathbf{I}$ and adding a multiple s of each element in row j to the corresponding element in row i.) For example, with $n = 4$, $i = 3$, and $j = 1$,

$$\mathbf{S} = \begin{bmatrix} 1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ s & 0 & 1 & 0\\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that $\det \mathbf{S} = 1$.

Premultiplication of an arbitrary $n \times p$ matrix $\mathbf{A}$ by one of these elementary matrices produces an *elementary transformation* of $\mathbf{A}$, also termed an *elementary row operation*, on $\mathbf{A}$. As examples, we form the products $\mathbf{QA}$, $\mathbf{RA}$, and $\mathbf{SA}$, with $n = 3$, $i = 2$, $j = 3$, and $p = 4$.

1.
$$\mathbf{QA} = \begin{bmatrix} 1 & 0 & 0\\ 0 & q & 0\\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14}\\ a_{21} & a_{22} & a_{23} & a_{24}\\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14}\\ qa_{21} & qa_{22} & qa_{23} & qa_{24}\\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}.$$

2.
$$\mathbf{RA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{21} & a_{22} & a_{23} & a_{24} \end{bmatrix}.$$

3.
$$\mathbf{SA} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & s \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} + sa_{31} & a_{22} + sa_{32} & a_{23} + sa_{33} & a_{24} + sa_{34} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}.$$

It is apparent that premultiplication by the elementary matrices produces the following transformations of **A**:

1. **QA**: Multiplication of all elements of one row by a scalar.
2. **RA**: Interchange of two rows.
3. **SA**: Addition of a scalar multiple of elements of one row to the corresponding elements of another row.

Observe that in each case the original elementary matrix can be formed from the identity matrix **I** by manipulating it exactly as we wish to have **A** manipulated.

Postmultiplication of an arbitrary $p \times n$ matrix **A** by one of the elementary matrices is called an *elementary column operation.* The three types of operations produce the following results:

1. **AQ**: Multiplication of all elements of one column by a scalar.
2. **AR**: Interchange of two columns.
3. **AS**: Addition of a scalar multiple of elements of one column to the corresponding elements of another column.

If **A** is any matrix and **T** is the matrix resulting from elementary row or column operations on **A**, **T** and **A** are termed *equivalent matrices.* For the examples given above, if **A** is a *square* matrix,

$$\begin{aligned} \det(\mathbf{QA}) &= \det \mathbf{Q} \times \det \mathbf{A} = q \det \mathbf{A}; \\ \det(\mathbf{RA}) &= \det \mathbf{R} \times \det \mathbf{A} = -\det \mathbf{A}; \\ \det(\mathbf{SA}) &= \det \mathbf{S} \times \det \mathbf{A} = \det \mathbf{A}. \end{aligned}$$

Thus, multiplication of all the elements of one row of a square matrix by a scalar also multiplies the determinant of the matrix by that scalar. Interchange of two rows changes the sign of the determinant (but not its magnitude), and addition of a scalar multiple of elements of one row to the corresponding elements of another row has no effect on the determinant.

Clearly, the product of elementary matrices is nonsingular, for each component has an inverse. It is also true that every nonsingular matrix can be written as a product of elementary matrices.

5.3 Gaussian Elimination

The *direct* methods of solving equations (5.2) are based on manipulations using the techniques expressed by the elementary matrices of Section 5.2. We now describe one such method, known as *Gaussian elimination.* Consider a general system of three linear equations:

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + b_{13}x_3 &= u_1, \\ b_{21}x_1 + b_{22}x_2 + b_{23}x_3 &= u_2, \\ b_{31}x_1 + b_{32}x_2 + b_{33}x_3 &= u_3. \end{aligned} \tag{5.4}$$

As a first step, replace the second equation by the result of adding to it the first equation multiplied by $-b_{21}/b_{11}$. Similarly, replace the third equation by the result of adding to it the first equation multiplied by $-b_{31}/b_{11}$. The result is the system

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + b_{13}x_3 &= u_1, \\ b'_{22}x_2 + b'_{23}x_3 &= u'_2, \\ b'_{32}x_2 + b'_{33}x_3 &= u'_3. \end{aligned} \tag{5.5}$$

in which the b' and u' are the new coefficients resulting from the above manipulations. Now multiply the second equation of (5.5) by $-b'_{32}/b'_{22}$, and add the result to the third equation of (5.5). The result is the triangular system

$$\begin{aligned} b_{11}x_1 + b_{12}x_2 + b_{13}x_3 &= u_1, \\ b'_{22}x_2 + b'_{23}x_3 &= u'_2, \\ b''_{33}x_3 &= u''_3, \end{aligned} \tag{5.6}$$

in which b''_{33} and u''_3 result from the arithmetic operations. The system (5.6) is readily solved by the process of *back-substitution,* in which x_3 is obtained from the last equation; this allows x_2 to be obtained from the second equation, and then x_1 can be found from the first equation.

The above method seems primitive at a first glance, but by the time it has been made suitable for implementation by automatic machines, it furnishes a powerful tool not only for solving equations (5.2), but also for finding the inverse of the related matrix of coefficients **B**, the determinant of **B**, the adjoint of **B**, etc.

Insofar as reaching (5.6) is concerned, all can be explained in terms of elementary matrices of the third kind. Note that matrices alone suffice, the presence of x_1, x_2, and x_3 being superfluous. Define an *augmented matrix* **C** consisting of the original coefficient matrix **B** with the right-hand side vector **u** appended to it. That is,

$$\mathbf{C} = [\mathbf{B} \mid \mathbf{u}] = \begin{bmatrix} b_{11} & b_{12} & b_{13} & u_1 \\ b_{21} & b_{22} & b_{23} & u_2 \\ b_{31} & b_{32} & b_{33} & u_3 \end{bmatrix},$$

in which the broken line denotes matrix partitioning.

Also define three elementary matrices of the third kind:

$$\mathbf{S}_1 = \begin{bmatrix} 1 & 0 & 0 \\ -\dfrac{b_{21}}{b_{11}} & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad \mathbf{S}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -\dfrac{b_{31}}{b_{11}} & 0 & 1 \end{bmatrix},$$

$$\mathbf{S}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\dfrac{b'_{32}}{b'_{22}} & 1 \end{bmatrix}.$$

The operations producing (5.6) from (5.4) can then be expressed as

$$\mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{C} = \begin{bmatrix} b_{11} & b_{12} & b_{13} & u_1 \\ 0 & b'_{22} & b'_{23} & u'_2 \\ 0 & 0 & b''_{33} & u''_3 \end{bmatrix}.$$

The back-substitution is expressed in terms of premultiplication by elementary matrices of the first and third kinds. Let $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ denote the three matrices of the first kind which are needed. For example,

$$\mathbf{Q}_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \dfrac{1}{b''_{33}} \end{bmatrix}.$$

Then, with three more matrices of the third kind, which we call $\mathbf{S}_4$, $\mathbf{S}_5$, and $\mathbf{S}_6$, the complete sequence of operations results in

$$\mathbf{Q}_3\mathbf{S}_6\mathbf{S}_5\mathbf{Q}_2\mathbf{S}_4\mathbf{Q}_1\mathbf{S}_3\mathbf{S}_2\mathbf{S}_1\mathbf{C} = \begin{bmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \end{bmatrix}.$$

Let $\mathbf{E}$ denote the product of these nine elementary matrices. Then $\mathbf{EC} = \mathbf{E}[\mathbf{B} \,\vdots\, \mathbf{u}] = [\mathbf{I} \,\vdots\, \mathbf{x}]$, whence $\mathbf{EB} = \mathbf{I}$ and $\mathbf{E} = \mathbf{B}^{-1}$. Hence, as a byproduct of solving equations such as (5.4) by elimination, we see that proper planning can produce $\mathbf{B}^{-1}$. Clearly, we need not solve equations at all if only the inverse is needed, for in that event the column $\mathbf{u}$ is superfluous.

Since $\mathbf{EB} = \mathbf{I}$, $\det(\mathbf{E}) \det(\mathbf{B}) = \det(\mathbf{I}) = 1$. From Section 5.2, the determinant of an $\mathbf{S}$ or third-kind elementary matrix is unity, whereas the determinant of a $\mathbf{Q}$ or first-kind matrix equals the value of that diagonal element which is usually not unity. Hence $\det(\mathbf{E}) = \det(\mathbf{Q}_3) \times \det(\mathbf{Q}_2) \det(\mathbf{Q}_1)$. That is, $\det(\mathbf{E})$ is the product of the diagonal elements (such as $1/b''_{33}$) of the matrices $\mathbf{Q}_1$, $\mathbf{Q}_2$, and $\mathbf{Q}_3$ used in the elimination process. This means that $\det(\mathbf{B})$ is the product of their reciprocals.

The above arithmetic operations can be separated into two types: (*a*) *normalization* steps in which the diagonal elements are converted to unity, and (*b*) *reduction* steps in which the off-diagonal elements are converted to zero. Note that by augmenting the coefficient matrix with several right-hand side vectors, we can solve several sets of simultaneous equations, each having the same coefficient matrix, at little extra computational cost.

Example. Consider the system of equations

$$\begin{aligned} 2x_1 - 7x_2 + 4x_3 &= 9, \\ x_1 + 9x_2 - 6x_3 &= 1, \\ -3x_1 + 8x_2 + 5x_3 &= 6, \end{aligned}$$

for which the solution is $x_1 = 4, x_2 = 1$, and $x_3 = 2$. The augmented matrix $[\mathbf{B} \,\vdots\, \mathbf{u} \,\vdots\, \mathbf{I}]$ will be formed, and the Gaussian elimination procedure just described will be carried out, except that the normalization steps will be introduced in a somewhat different order. Starting with the matrix,

$$\begin{bmatrix} 2 & -7 & 4 & 9 & 1 & 0 & 0 \\ 1 & 9 & -6 & 1 & 0 & 1 & 0 \\ -3 & 8 & 5 & 6 & 0 & 0 & 1 \end{bmatrix},$$

we multiply the top row by 1/2, add -1 times the new first row to the second row, and 3 times the new first row to the third row. The result is

$$\begin{bmatrix} 1 & -\dfrac{7}{2} & 2 & \dfrac{9}{2} & \dfrac{1}{2} & 0 & 0 \\ 0 & \dfrac{25}{2} & -8 & -\dfrac{7}{2} & -\dfrac{1}{2} & 1 & 0 \\ 0 & -\dfrac{5}{2} & 11 & \dfrac{39}{2} & \dfrac{3}{2} & 0 & 1 \end{bmatrix}. \tag{5.7}$$

This is equivalent to having formed the equations

$$\begin{aligned} x_1 - \frac{7}{2}x_2 + 2x_3 &= \frac{9}{2}, \\ \frac{25}{2}x_2 - 8x_3 &= -\frac{7}{2}, \\ -\frac{5}{2}x_2 + 11x_3 &= \frac{39}{2}. \end{aligned}$$

Note that the operations performed are equivalent to the matrix multiplication,

$$\begin{bmatrix} \dfrac{1}{2} & 0 & 0 \\ -\dfrac{1}{2} & 1 & 0 \\ \dfrac{3}{2} & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -7 & 4 & 9 & 1 & 0 & 0 \\ 1 & 9 & -6 & 1 & 0 & 1 & 0 \\ -3 & 8 & 5 & 6 & 0 & 0 & 1 \end{bmatrix}$$

which yields as a result

$$\begin{bmatrix} 1 & -\dfrac{7}{2} & 2 & \dfrac{9}{2} & \dfrac{1}{2} & 0 & 0 \\ 0 & \dfrac{25}{2} & -8 & -\dfrac{7}{2} & -\dfrac{1}{2} & 1 & 0 \\ 0 & -\dfrac{5}{2} & 11 & \dfrac{39}{2} & \dfrac{3}{2} & 0 & 1 \end{bmatrix}.$$

Returning to (5.7), multiply the second row by 2/25, and then add 5/2 times the new second row to the third row. The result is:

$$\begin{bmatrix} 1 & -\frac{7}{2} & 2 & \frac{9}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{16}{25} & -\frac{7}{25} & -\frac{1}{25} & \frac{2}{25} & 0 \\ 0 & 0 & \frac{47}{5} & \frac{94}{5} & \frac{7}{5} & \frac{1}{5} & 1 \end{bmatrix}.$$

The *forward course* has now been completed and, corresponding to (5.6), we may write

$$\begin{aligned} x_1 - \frac{7}{2}x_2 + 2x_3 &= \frac{9}{2}, \\ x_2 - \frac{16}{25}x_3 &= -\frac{7}{25}, \\ \frac{47}{5}x_3 &= \frac{94}{5}. \end{aligned}$$

To carry out the back-substitution, start by multiplying the last row by 5/47. Then multiply the new last row by 16/25 and add to the second row. Multiply this same last row by -2 and add to the first row. The result is

$$\begin{bmatrix} 1 & -\frac{7}{2} & 0 & \frac{1}{2} & \frac{19}{94} & -\frac{2}{47} & -\frac{10}{47} \\ 0 & 1 & 0 & 1 & \frac{13}{235} & \frac{22}{235} & \frac{16}{235} \\ 0 & 0 & 1 & 2 & \frac{7}{47} & \frac{1}{47} & \frac{5}{47} \end{bmatrix}.$$

Finally, multiply the second row by 7/2 and add to the first. The result is

$$\begin{bmatrix} 1 & 0 & 0 & 4 & \frac{93}{235} & \frac{67}{235} & \frac{6}{235} \\ 0 & 1 & 0 & 1 & \frac{13}{235} & \frac{22}{235} & \frac{16}{235} \\ 0 & 0 & 1 & 2 & \frac{7}{47} & \frac{1}{47} & \frac{5}{47} \end{bmatrix}.$$

This means, of course, that $x_1 = 4$, $x_2 = 1$, $x_3 = 2$ and the inverse of the matrix of coefficients is

$$\begin{bmatrix} \frac{93}{235} & \frac{67}{235} & \frac{6}{235} \\ \frac{13}{235} & \frac{22}{235} & \frac{16}{235} \\ \frac{7}{47} & \frac{1}{47} & \frac{5}{47} \end{bmatrix}.$$

The determinant of the coefficient matrix **B** equals the product of the reciprocals of the diagonal elements appearing in the Q-type matrices involved in the above transformation. Inspection shows that the relevant diagonal elements are simply the multiplying factors used in the normalization steps, so that

$$\det \mathbf{B} = \left[\frac{1}{2} \times \frac{2}{25} \times \frac{5}{47}\right]^{-1} = 235.$$

5.4 Gauss-Jordan Elimination

A variation that accomplishes the effect of back-substitution simultaneously with the reduction of the subdiagonal elements will now be illustrated, again for the system,

$$\begin{aligned} 2x_1 - 7x_2 + 4x_3 &= 9, \\ x_1 + 9x_2 - 6x_3 &= 1, \\ -3x_1 + 8x_2 + 5x_3 &= 6. \end{aligned}$$

Suppose that $\mathbf{B}^{-1}$ is required and form the augmented matrix $[\mathbf{B} \mid \mathbf{u} \mid \mathbf{I}]$:

$$\begin{bmatrix} 2 & -7 & 4 & 9 & 1 & 0 & 0 \\ 1 & 9 & -6 & 1 & 0 & 1 & 0 \\ -3 & 8 & 5 & 6 & 0 & 0 & 1 \end{bmatrix}.$$

As before, normalize the first row by dividing by the *pivot* element 2; then reduce the remaining elements of the first column to zero by subtracting the new first row from the second row, and also by subtracting -3 times the new first row from the third row. The result is

$$\begin{bmatrix} 1 & -\frac{7}{2} & 2 & \frac{9}{2} & \frac{1}{2} & 0 & 0 \\ 0 & \frac{25}{2} & -8 & -\frac{7}{2} & -\frac{1}{2} & 1 & 0 \\ 0 & -\frac{5}{2} & 11 & \frac{39}{2} & \frac{3}{2} & 0 & 1 \end{bmatrix}.$$

Next, normalize the second row by dividing by the pivot element 25/2; then reduce the remaining elements of the second column to zero by subtracting $-(7/2)$ times the new second row from the first row, and $-(5/2)$ times the new second row from the third row. Note that the reduction process now involves both the subdiagonal *and* superdiagonal elements. The result is

$$\begin{bmatrix} 1 & 0 & -\frac{6}{25} & \frac{88}{25} & \frac{9}{25} & \frac{7}{25} & 0 \\ 0 & 1 & -\frac{16}{25} & -\frac{7}{25} & -\frac{1}{25} & \frac{2}{25} & 0 \\ 0 & 0 & \frac{47}{5} & \frac{94}{5} & \frac{7}{5} & \frac{1}{5} & 1 \end{bmatrix}.$$

Finally, normalize the last row by dividing by the pivot element 47/5; then reduce the remaining elements of the third column to zero by subtracting $-(6/25)$ and $-(16/25)$

times the new third row from the first and second rows, respectively. The resulting matrix is $[\mathbf{I}\,|\,\mathbf{x}\,|\,\mathbf{B}^{-1}]$, where $\mathbf{x}$ is the solution vector, and $\mathbf{B}^{-1}$ is the inverse of the original matrix of coefficients:

$$\begin{bmatrix} 1 & 0 & 0 & 4 & \dfrac{93}{235} & \dfrac{67}{235} & \dfrac{6}{235} \\ 0 & 1 & 0 & 1 & \dfrac{13}{235} & \dfrac{22}{235} & \dfrac{16}{235} \\ 0 & 0 & 1 & 2 & \dfrac{7}{47} & \dfrac{1}{47} & \dfrac{5}{47} \end{bmatrix}.$$

The determinant of the original coefficient matrix is again the product of the pivot elements and thus equals $2 \times 25/2 \times 47/5$ or 235.

We conclude this section by developing an algorithm for the above procedure, which is called *Gauss-Jordan elimination*. Let the starting array be the $n \times (n + m)$ augmented matrix $\mathbf{A}$, consisting of an $n \times n$ coefficient matrix with m appended columns:

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} & a_{1,n+1} & a_{1,n+2} & \cdots & a_{1,n+m} \\ a_{21} & a_{22} & \cdots & a_{2n} & a_{2,n+1} & a_{2,n+2} & \cdots & a_{2,n+m} \\ \vdots & & & \vdots & \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & a_{n,n+1} & a_{n,n+2} & \cdots & a_{n,n+m} \end{bmatrix}.$$

Let $k = 1, 2, \ldots, n$ be the pivot counter, so that a_{kk} is the pivot element for the kth pass of the reduction. It is understood that the values of the elements of $\mathbf{A}$ will be modified during computation. The algorithm is

$$\left.\begin{array}{ll} \textit{Normalization} & \\ & a_{kj} \leftarrow \dfrac{a_{kj}}{a_{kk}}, \quad j = n+m, n+m-1, \ldots, k \\ \textit{Reduction} & \\ & \left.\begin{array}{l} a_{ij} \leftarrow a_{ij} - a_{ik}a_{kj}, \\ j = n+m, n+m-1, \ldots, k \end{array}\right\} \begin{array}{l} i = 1, 2, \ldots, n \\ (i \neq k) \end{array} \end{array}\right\} k = 1, 2, \ldots, n. \tag{5.8}$$

Note (*a*) Since no nonzero elements appear to the left of a_{kk} in the kth row at the beginning of the kth pass, it is unnecessary to normalize a_{kj} for $j < k$; (*b*) In order to avoid premature modification of elements in the pivot column, the column counter j is always decremented from its highest value $(n + m)$ until the pivot column is reached.

Thus far, elementary matrices of the second kind have not been used; neither has mention been made of the fact that at some stage, say the first, a potential divisor or pivot, such as b_{11}, may be zero. In this event, we can think of interchanging rows, which is expressible, of course, in terms of elementary row operations of the second kind. A related problem is that of maintaining sufficient accuracy during intermediate calculations in order to achieve specified accuracy in the final results. This might be expected for a nearly singular system; it can also happen when the magnitude of one of the pivot elements is relatively small. Consider, for instance, the system

$$\begin{aligned} 0.0003\,x_1 + 3.0000\,x_2 &= 2.0001, \\ 1.0000\,x_1 + 1.0000\,x_2 &= 1.0000, \end{aligned}$$

which has the exact solution $x_1 = 1/3, x_2 = 2/3$. If the equations are solved using pivots on the matrix diagonal, as indicated in the previous examples, there results

$$\begin{aligned} 1.0000\,x_1 + 10000\,x_2 &= 6667, \\ x_2 &= 6666/9999. \end{aligned}$$

If x_2 from the second equation is taken to be 0.6667, then from the first equation $x_1 = 0.0000$; for $x_2 = 0.66667$, $x_1 = 0.30000$; for $x_2 = 0.666667$, $x_1 = 0.330000$, etc. The solution depends highly on the number of figures retained. If the equations are solved in reverse order, that is, by interchanging the two rows and proceeding as before, then x_2 is found to be $1.9998/2.9997 \doteq 0.66667$ while $x_1 = 0.33333$. This example indicates the advisability of choosing as the pivot the coefficient of largest absolute value in a column, rather than merely the first in line. The handling of the situation is developed in more detail in Example 5.2.

on to page 300
(omitting some examples)

5.5 A Finite Form of the Method of Kaczmarz

We consider the solution of (5.2) under the assumption that $\mathbf{B}$ is nonsingular. The procedure consists in first converting the system of (5.2) into an equivalent system,

$$\mathbf{A}\mathbf{x} = \mathbf{v}, \qquad \mathbf{A}^* = \mathbf{A}^{-1}, \tag{5.9}$$

such that (5.9) and (5.2) have the same solution vector $\mathbf{x}$. Then, using an arbitrary initial vector, $\mathbf{r}_0$, we define

$$\mathbf{r}_j = \mathbf{r}_{j-1} - [(\boldsymbol{\alpha}_j, \mathbf{r}_{j-1}) - v_j]\boldsymbol{\alpha}_j, \qquad 1 \leqslant j \leqslant n, \tag{5.10}$$

where $\mathrm{A} = (a_{ij})$ and $\boldsymbol{\alpha}_j = [\bar{a}_{j1}, \bar{a}_{j2}, \ldots, \bar{a}_{jn}]^t$. Then $\mathbf{A}\mathbf{r}_n = \mathbf{v}$ and

$$\mathbf{B}\mathbf{r}_n = \mathbf{u}. \tag{5.11}$$

We show that (5.10) defines a vector $\mathbf{r}_n$ such that $\mathbf{A}\mathbf{r}_n = \mathbf{v}$. Notice first that the equations of (5.9) may be written

$$(\boldsymbol{\alpha}_i, \mathbf{x}) = v_i, \qquad 1 \leqslant i \leqslant n.$$

Notice next that, multiplying (5.10) on the left by $\boldsymbol{\alpha}_j$,

$$(\boldsymbol{\alpha}_j, \mathbf{r}_j) = v_j,$$

so that the jth equation of (5.9) is satisfied by $\mathbf{r}_j$. However,

$$(\boldsymbol{\alpha}_i, \mathbf{r}_j) = (\boldsymbol{\alpha}_i, \mathbf{r}_{j-1}), \qquad i \neq j,$$

since $(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) = 0$ for $i \neq j$. Thus, if $i < j$, $(\boldsymbol{\alpha}_i, \mathbf{r}_j) = (\boldsymbol{\alpha}_i, \mathbf{r}_{j-1}) = \cdots = (\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i) = v_i$. We see inductively that $\mathbf{A}\mathbf{r}_n = \mathbf{v}$, and the unique solution of (5.9) has been found.

Turn now to the solution of (5.2). Let $\mathbf{B}^* = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots, \boldsymbol{\beta}_n]$ where $\boldsymbol{\beta}_i = [\bar{b}_{i1}, \bar{b}_{i2}, \ldots, \bar{b}_{in}]^t$. A system equivalent to (5.2) and having the properties of (5.9) is built in orthodox manner from the linearly independent vectors $\boldsymbol{\beta}_i$ by using the *Gram-Schmidt* orthogonalization procedure, as follows. Let

$$\boldsymbol{\gamma}_1 = \boldsymbol{\beta}_1;$$

$$\boldsymbol{\alpha}_i = \frac{\boldsymbol{\gamma}_i}{\sqrt{(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_i)}}, \qquad 1 \leqslant i \leqslant n; \tag{5.12}$$

$$\boldsymbol{\gamma}_j = \boldsymbol{\beta}_j - \sum_{i=1}^{j-1} (\boldsymbol{\alpha}_i, \boldsymbol{\beta}_j)\boldsymbol{\alpha}_i, \qquad 2 \leqslant j \leqslant n.$$

Then it is readily found that $(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i) = 1$, while $(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_j) = 0$ if $i \neq j$ is shown to be true inductively. Thus, if $\mathbf{A}^* = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \ldots, \boldsymbol{\alpha}_n]$, then $\mathbf{A}^* = \mathbf{A}^{-1}$. This is verified by direct multiplication. Finally, let

$$v_1 = \frac{u_1}{\sqrt{(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_1)}},$$

$$v_j = \frac{1}{\sqrt{(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_j)}} \left\{ u_j - \sum_{i=1}^{j-1} (\boldsymbol{\beta}_j, \boldsymbol{\alpha}_i) v_i \right\}. \tag{5.13}$$

With these definitions, a solution of (5.9) is a solution of (5.2). For, if $(\boldsymbol{\alpha}_j, \mathbf{x}) - v_j = 0$, $1 \leqslant j \leqslant n$, then by (5.12),

$$(\boldsymbol{\beta}_j, \mathbf{x}) - \sum_{i=1}^{j-1} (\boldsymbol{\beta}_j, \boldsymbol{\alpha}_i)(\boldsymbol{\alpha}_i, \mathbf{x}) - (\boldsymbol{\gamma}_j, \mathbf{x}) = 0$$

or

$$(\boldsymbol{\beta}_j, \mathbf{x}) - \sum_{i=1}^{j-1} (\boldsymbol{\beta}_j, \boldsymbol{\alpha}_i) v_i - \sqrt{(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_j)}\, v_j = 0.$$

Then, by (5.13),

$$(\boldsymbol{\beta}_j, \mathbf{x}) = u_j.$$

In application, all can be accomplished by using the array $[\mathbf{B} \mid \mathbf{u}]$ and forming in the same locations the array $[\mathbf{A} \mid \mathbf{v}]$. If it is desired to vary the vector $\mathbf{u}$ after $\mathbf{A}$ has been built, it will be necessary to record the n numbers $\sqrt{(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_i)}$ and the $(n^2 - n)/2$ numbers $(\boldsymbol{\beta}_j, \boldsymbol{\alpha}_i)$, $1 \leqslant i < j$, $2 \leqslant j \leqslant \mathrm{n}$. The building of the matrix $[\mathbf{A} \mid \mathbf{v}]$ from $[\mathbf{B} \mid \mathbf{u}]$ can be visualized best by writing the conjugates of relations (5.12) [but not of (5.13)]. Then observe that the first row of $[\mathbf{A} \mid \mathbf{v}]$ is formed from the first row of $[\mathbf{B} \mid \mathbf{u}]$, the second row of $[\mathbf{A} \mid \mathbf{v}]$ from the second row of $[\mathbf{B} \mid \mathbf{u}]$ and the just established row of $[\mathbf{A} \mid \mathbf{v}]$, etc. Each operation involved can be viewed as tantamount to premultiplication by an elementary matrix of the first or third kind. Thus there exists a nonsingular matrix $\boldsymbol{\phi}$ such that

$$\boldsymbol{\phi}[\mathbf{B} \mid \mathbf{u}] = [\mathbf{A} \mid \mathbf{v}].$$

It will be seen that

$$\det(\boldsymbol{\phi}) = \prod_{i=1}^{n} \frac{1}{\sqrt{(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_i)}}. \tag{5.14}$$

This knowledge can be useful in case the matrix $\mathbf{B}$ is *ill-conditioned*, that is, has rows or columns so nearly dependent on each other that rounding or truncation errors can cause the calculated determinantal value to deviate markedly from its true value. Now recall that $|\det(\mathbf{A})| = 1$ to realize that (5.14) can accomplish the purpose cited. Note also that the sequences $\{\boldsymbol{\alpha}_j\}$, $\{v_j\}$, and $\{\mathbf{r}_j\}$ can progress together, so that the method can properly be called an n-step method.

After orthogonalization, it is also possible to find the solution vector $\mathbf{x}$ as $\mathbf{A}^*\mathbf{v}$.

Example. As a simple illustration of the Kaczmarz method, we consider the following problem, also discussed in Sections 5.6 and 5.7:

$$\begin{aligned} 4x_1 + 2x_2 + x_3 &= 11, \\ -x_1 + 2x_2 \phantom{{}+ 4x_3} &= 3, \\ 2x_1 + x_2 + 4x_3 &= 16. \end{aligned}$$

The matrix $[\mathbf{B} \mid \mathbf{u}]$ is

$$\begin{bmatrix} 4 & 2 & 1 & 11 \\ -1 & 2 & 0 & 3 \\ 2 & 1 & 4 & 16 \end{bmatrix}.$$

The first row of $[\mathbf{A}|\mathbf{v}]$ is that of $[\mathbf{B}|\mathbf{u}]$ divided by $\sqrt{4^2+2^2+1^2}$ or $\left[\frac{4}{\sqrt{21}},\frac{2}{\sqrt{21}},\frac{1}{\sqrt{21}},\frac{11}{\sqrt{21}}\right]$. Prior to normalizing, the second row is

$$[-1,2,0,3]-0\left[\frac{4}{\sqrt{21}},\frac{2}{\sqrt{21}},\frac{1}{\sqrt{21}},\frac{11}{\sqrt{21}}\right]=[-1,2,0,3].$$

The normalizing factor is $\sqrt{1^2+2^2}$, so that the second row of $[\mathbf{A}|\mathbf{v}]$ is $\left[\frac{-1}{\sqrt{5}},\frac{2}{\sqrt{5}},0,\frac{3}{\sqrt{5}}\right]$. Prior to normalizing, the third row is

$$[2,1,4,16]-\left(\frac{14}{\sqrt{21}}\right)\left[\frac{4}{\sqrt{21}},\frac{2}{\sqrt{21}},\frac{1}{\sqrt{21}},\frac{11}{\sqrt{21}}\right]$$
$$+0\left[\frac{-1}{\sqrt{5}},\frac{2}{\sqrt{5}},0,\frac{3}{\sqrt{5}}\right]\quad\text{or}\quad\left[-\frac{2}{3},-\frac{1}{3},\frac{10}{3},\frac{26}{3}\right].$$

Thus for $[\mathbf{A}|\mathbf{v}]$, we have

$$\begin{bmatrix}\frac{4}{\sqrt{21}} & \frac{2}{\sqrt{21}} & \frac{1}{\sqrt{21}} & \frac{11}{\sqrt{21}}\\ -\frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} & 0 & \frac{3}{\sqrt{5}}\\ -\frac{2}{\sqrt{105}} & -\frac{1}{\sqrt{105}} & \frac{10}{\sqrt{105}} & \frac{26}{\sqrt{105}}\end{bmatrix}.$$

Using $\mathbf{r}_0=[1,1,1]^t$, we find

$$\mathbf{r}_1=[1,1,1]^t-\left(-\frac{4}{\sqrt{21}}\right)\left[\frac{4}{\sqrt{21}},\frac{2}{\sqrt{21}},\frac{1}{\sqrt{21}}\right]^t$$
$$=\left[\frac{37}{21},\frac{29}{21},\frac{25}{21}\right]^t.$$

Then

$$\mathbf{r}_2=\left[\frac{37}{21},\frac{29}{21},\frac{25}{21}\right]^t-\left(\frac{-2}{\sqrt{5}}\right)\left[\frac{-1}{\sqrt{5}},\frac{2}{\sqrt{5}},0\right]^t$$
$$=\left[\frac{143}{105},\frac{229}{105},\frac{25}{21}\right]^t.$$

There results

$$\mathbf{r}_3=\left[\frac{143}{105},\frac{229}{105},\frac{25}{21}\right]^t-\left(-\frac{19}{\sqrt{105}}\right)\left[\frac{-2}{\sqrt{105}},\frac{-1}{\sqrt{105}},\frac{10}{\sqrt{105}}\right]^t$$
$$=[1,2,3]^t.$$

5.6 Jacobi Iterative Method

Consider again the solution of the linear system $\mathbf{Bx}=\mathbf{u}$:

$$\begin{aligned}b_{11}x_1+b_{12}x_2+\cdots+b_{1n}x_n&=u_1,\\ b_{21}x_1+b_{22}x_2+\cdots+b_{2n}x_n&=u_2,\\ \vdots\qquad\qquad\qquad&\;\;\vdots\\ b_{n1}x_1+b_{n2}x_2+\cdots+b_{nn}x_n&=u_n.\end{aligned}\tag{5.2}$$

We now formulate the *Jacobi iterative method* for *approximating* the solution of (5.2). The degree of approximation, however, can normally be improved by expending more computational effort, that is, by performing an increased number of iterations.

First, solve for the x_i, giving:

$$\begin{aligned}x_1&=(u_1-b_{12}x_2-b_{13}x_3-\cdots-b_{1n}x_n)/b_{11},\\ x_2&=(u_2-b_{21}x_1-b_{23}x_3-\cdots-b_{2n}x_n)/b_{22},\\ \vdots&\qquad\qquad\qquad\qquad\qquad\vdots\\ x_n&=(u_n-b_{n1}x_1-b_{n2}x_2-\cdots-b_{n,n-1}x_{n-1})/b_{nn}.\end{aligned}\tag{5.15}$$

The system (5.15) can be written more concisely as

$$x_i=\frac{\left(u_i-\sum\limits_{\substack{j=1\\ j\neq i}}^{n}b_{ij}x_j\right)}{b_{ii}},\quad i=1,2,\ldots,n.\tag{5.16}$$

Note that the above rearrangement is predicated on $b_{ii}\neq 0$. Usually, we try to reorder the equations and the unknowns so that *diagonal dominance* is obtained, that is, so that each diagonal element b_{ii} is larger, in absolute value, than the magnitudes of other entries in row i and column i. In this connection, also see relations (5.21).

Next, make starting guesses for the x's and insert these values into the right-hand sides of (5.15). The resulting new approximations for the x's are resubstituted into the right-hand sides of (5.15), and the process is repeated. Hopefully, the x's thus computed will show little further change after several such iterations have been made.

Example. Consider the equations

$$\begin{aligned}4x_1+2x_2+x_3&=11,\\ -x_1+2x_2\qquad&=\;\;3,\\ 2x_1+x_2+4x_3&=16,\end{aligned}$$

which have the solution vector $\mathbf{x}=[1,2,3]^t$, that is, $x_1=1$, $x_2=2$, and $x_3=3$. Rewrite the equations as

$$x_1=\frac{11}{4}-\frac{1}{2}x_2-\frac{1}{4}x_3,$$
$$x_2=\frac{3}{2}+\frac{1}{2}x_1,$$
$$x_3=4-\frac{1}{2}x_1-\frac{1}{4}x_2,$$

and arbitrarily choose a starting vector $\mathbf{x}_0=[1,1,1]^t$ in which the subscript denotes the *zero*th stage of iteration. Using a second subscript to denote the iteration number, the first iteration gives

$$x_{11}=\frac{11}{4}-\frac{1}{2}\times 1-\frac{1}{4}\times 1=2,$$
$$x_{21}=\frac{3}{2}+\frac{1}{2}\times 1=2,$$
$$x_{31}=4-\frac{1}{2}\times 1-\frac{1}{4}\times 1=\frac{13}{4}.$$

That is, $\mathbf{x}_1 = [2, 2, 13/4]^t$. Similarly, the next four iterations yield

$$\mathbf{x}_2 = \left[\frac{15}{16}, \frac{5}{2}, \frac{5}{2}\right]^t,$$

$$\mathbf{x}_3 = \left[\frac{7}{8}, \frac{63}{32}, \frac{93}{32}\right]^t,$$

$$\mathbf{x}_4 = \left[\frac{133}{128}, \frac{31}{16}, \frac{393}{128}\right]^t,$$

$$\mathbf{x}_5 = \left[\frac{519}{512}, \frac{517}{256}, \frac{767}{256}\right]^t.$$

The approximation computed at the fifth iteration is roughly within 1% of the exact solution. The accuracy could be improved by performing more iterations. Observe that a whole new solution vector is computed before it is used in the next iteration.

In order to establish a criterion for the convergence of the Jacobi method, regard the rearranged equations (5.15) as the system

$$\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{v}, \tag{5.17}$$

in which

$$\mathbf{A} = -\begin{bmatrix} 0 & \dfrac{b_{12}}{b_{11}} & \dfrac{b_{13}}{b_{11}} & \cdots & \dfrac{b_{1n}}{b_{11}} \\ \dfrac{b_{21}}{b_{22}} & 0 & \dfrac{b_{23}}{b_{22}} & \cdots & \dfrac{b_{2n}}{b_{22}} \\ \vdots & & & & \vdots \\ \dfrac{b_{n1}}{b_{nn}} & \dfrac{b_{n2}}{b_{nn}} & \cdots & \dfrac{b_{n,n-1}}{b_{nn}} & 0 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} \dfrac{u_1}{b_{11}} \\ \dfrac{u_2}{b_{22}} \\ \vdots \\ \dfrac{u_n}{b_{nn}} \end{bmatrix}. \tag{5.18}$$

If the starting vector $\mathbf{x}_0$ is near the solution vector $\mathbf{x}$, convergence will be faster. In any event, define

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{v}, \tag{5.19}$$

in which the subscript k is the iteration number. This means that

$$\mathbf{x}_k = \mathbf{A}^k\mathbf{x}_0 + [\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^{k-1}]\mathbf{v}.$$

From this, we see that convergence normally requires that

$$\lim_{k\to\infty} \mathbf{A}^k = \mathbf{0}. \tag{5.20}$$

From (4.23), it is also a necessary and sufficient condition that

$$\lim_{k\to\infty} [\mathbf{I} + \mathbf{A} + \mathbf{A}^2 + \cdots + \mathbf{A}^k] = (\mathbf{I} - \mathbf{A})^{-1}.$$

Thus, when (5.20) is satisfied, $\mathbf{x} = \lim_{k\to\infty} \mathbf{x}_k$ exists and $\mathbf{x} = \mathbf{0} + (\mathbf{I} - \mathbf{A})^{-1}\mathbf{v}$; that is $(\mathbf{I} - \mathbf{A})\mathbf{x} = \mathbf{v}$ or $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{v}$.

Thus, convergence hinges on the truth of (5.20). From page 222, (5.20) is true if and only if all eigenvalues of $\mathbf{A}$ are in modulus less than unity. For this to be so, from (4.25), (4.26), and the subsequent development, we have the sufficient conditions

$$\left.\begin{aligned} &\sum_{i=1}^{n} |a_{ij}| \leqslant \mu < 1, \qquad 1 \leqslant j \leqslant n, \\ \text{or} \\ &\sum_{j=1}^{n} |a_{ij}| \leqslant \mu < 1, \qquad 1 \leqslant i \leqslant n, \\ \text{or} \\ &\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \leqslant \mu < 1. \end{aligned}\right\} \tag{5.21}$$

By using (5.18), these sufficiency conditions can also be translated into an equivalent set of conditions on the elements of the original coefficient matrix $\mathbf{B}$. For example, the second condition of (5.21) becomes

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |b_{ij}| < |b_{ii}|, \qquad 1 \leqslant i \leqslant n. \tag{5.22}$$

If, as frequently occurs, matrix $\mathbf{B}$ is *irreducible* (that is, a matrix of the form $\begin{bmatrix} \mathbf{B}_1 & \mathbf{O} \\ \mathbf{B}_2 & \mathbf{B}_3 \end{bmatrix}$, where $\mathbf{B}_1$ is square and $\mathbf{O}$ is the null matrix, cannot be found by permuting rows and columns of $\mathbf{B}$), the sufficiency condition can be relaxed (for example, see Ralston and Wilf [1]) to

$$\sum_{\substack{j=1\\ j\neq i}}^{n} |b_{ij}| \leqslant |b_{ii}|, \qquad 1 \leqslant i \leqslant n, \tag{5.23}$$

with strict inequality holding for at least one value of i.

5.7 Gauss-Seidel Iterative Method

The linear system considered is again that of (5.2) rephrased in the form (5.15) or (5.17). In the iterations, however, the newly-computed components of the $\mathbf{x}$ vector are always used in the right-hand sides as soon as they are obtained. This contrasts with the Jacobi method, in which the new components are not used until all n components have been found.

Example. The Gauss-Seidel method is applied to the short example considered under the Jacobi method. The form used is

$$x_1 = \frac{11}{4} - \frac{1}{2}x_2 - \frac{1}{4}x_3,$$

$$x_2 = \frac{3}{2} + \frac{1}{2}x_1,$$

$$x_3 = 4 - \frac{1}{2}x_1 - \frac{1}{4}x_2,$$

with the understanding that the most recently available x's are always used in the right-hand sides. Again $\mathbf{x}_0$ is chosen as $[1, 1, 1]^t$. The first iteration gives

$$x_{11} = \frac{11}{4} - \frac{1}{2} \times 1 - \frac{1}{4} \times 1 = 2,$$

$$x_{21} = \frac{3}{2} + \frac{1}{2} \times 2 = \frac{5}{2},$$

$$x_{31} = 4 - \frac{1}{2} \times 2 - \frac{1}{4} \times \frac{5}{2} = \frac{19}{8}.$$

That is,

$$\mathbf{x}_1 = \left[2, \frac{5}{2}, \frac{19}{8}\right]^t.$$

Similarly, the next two iterations yield

$$\mathbf{x}_2 = \left[\frac{29}{32}, \frac{125}{64}, \frac{783}{256}\right]^t,$$

$$\mathbf{x}_3 = \left[\frac{1033}{1024}, \frac{4095}{2048}, \frac{24541}{8192}\right]^t.$$

Observe that in this example the rate of convergence is much faster than that in the Jacobi method.

In order to investigate the conditions for the convergence of the Gauss-Seidel method, we first phrase the iteration in terms of the individual components. Let x_{ik} denote the kth approximation to the ith component of the solution vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]^t$. Let $[x_{10}, x_{20}, \ldots, x_{n0}]^t$ be an arbitrary initial approximation (though, as with the Jacobi method, if a good estimate is known, it should be used for efficiency). Let $\mathbf{A}$ and $\mathbf{v}$ be the same as given in (5.18), and define

$$x_{ik} = \sum_{j=1}^{i-1} a_{ij} x_{jk} + \sum_{j=i+1}^{n} a_{ij} x_{j,k-1} + v_i, \qquad (5.24)$$

for $1 \leqslant i \leqslant n$ and $1 \leqslant k$. When $i = 1$, $\sum_{j=1}^{i-1} a_{ij} x_{jk}$ is interpreted as zero, and when $i = n$, $\sum_{j=i+1}^{n} a_{ij} x_{j,k-1}$ is likewise interpreted as zero.

Write $\mathbf{A} = \mathbf{A_L} + \mathbf{A_R}$ where

$$\mathbf{A_L} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 0 \\ a_{21} & 0 & \cdots & 0 & 0 \\ \vdots & & & \vdots & \\ a_{n1} & a_{n2} & \cdots & a_{n,n-1} & 0 \end{bmatrix},$$

$$\mathbf{A_R} = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Thus $\mathbf{A_L}$ is a strictly lower-triangular matrix whose subdiagonal entries are the elements of $\mathbf{A}$ in their natural positions. A similar description applies to $\mathbf{A_R}$. It is seen that, if $\mathbf{x}_k = [x_{1k}, x_{2k}, \ldots, x_{nk}]^t$,

$$\mathbf{x}_k = \mathbf{A_L}\mathbf{x}_k + \mathbf{A_R}\mathbf{x}_{k-1} + \mathbf{v}.$$

This can be paraphrased as

$$\mathbf{x}_k = (\mathbf{I} - \mathbf{A_L})^{-1}\mathbf{A_R}\mathbf{x}_{k-1} + (\mathbf{I} - \mathbf{A_L})^{-1}\mathbf{v}, \qquad (5.25)$$

which is then of the Jacobi form. This means that a necessary and sufficient condition for the convergence of (5.24) is that the eigenvalues of $(\mathbf{I} - \mathbf{A_L})^{-1}\mathbf{A_R}$ be less than unity in modulus. The eigenvalues of $(\mathbf{I} - \mathbf{A_L})^{-1}\mathbf{A_R}$ are found by solving $\det((\mathbf{I} - \mathbf{A_L})^{-1}\mathbf{A_R} - \lambda\mathbf{I}) = 0$, or $\det([\mathbf{I} - \mathbf{A_L}]^{-1} \times [\mathbf{A_R} - \lambda(\mathbf{I} - \mathbf{A_L})]) = 0$, or $\det(\mathbf{A_R} - \lambda\mathbf{I} + \lambda\mathbf{A_L}) = 0$. Thus the Gauss-Seidel process converges if and only if the zeros of the determinant of

$$\begin{bmatrix} -\lambda & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21}\lambda & -\lambda & a_{23} & \cdots & a_{2n} \\ a_{31}\lambda & a_{32}\lambda & -\lambda & \cdots & a_{3n} \\ \vdots & & & & \vdots \\ a_{n1}\lambda & a_{n2}\lambda & a_{n3}\lambda & \cdots & -\lambda \end{bmatrix} \qquad (5.26)$$

are less than one in absolute value:

Since $a_{ii} = 0$, $1 \leqslant i \leqslant n$, while $a_{ij} = -b_{ij}/b_{ii}$ for $i \neq j$, the determinant of (5.26) has the same zeros as the determinant of

$$\begin{bmatrix} b_{11}\lambda & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21}\lambda & b_{22}\lambda & b_{23} & \cdots & b_{2n} \\ b_{31}\lambda & b_{32}\lambda & b_{33}\lambda & \cdots & b_{3n} \\ \vdots & & & & \vdots \\ b_{n1}\lambda & b_{n2}\lambda & b_{n3}\lambda & \cdots & b_{nn}\lambda \end{bmatrix}. \qquad (5.27)$$

It develops that conditions analogous to the first two of (5.21) prove sufficient to guarantee convergence, namely,

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \left|\frac{b_{ij}}{b_{ii}}\right| \leqslant \mu < 1 \quad \text{or} \quad \sum_{\substack{j=1 \\ j \neq i}}^{n} \left|\frac{b_{ji}}{b_{jj}}\right| \leqslant \mu < 1, \quad 1 \leqslant i \leqslant n. \qquad (5.28)$$

The first of these may be demonstrated as follows. We have already seen in (4.14) that since

$$|b_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |b_{ij}|,$$

$\mathbf{B}$ is nonsingular. Thus a solution vector $\mathbf{x}$ exists such that $\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{v}$, whence

$$x_i = \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j + v_i,$$

in which $a_{ij} = -b_{ij}/b_{ii}$. Subtracting this from (5.24) yields

$$|x_{ik} - x_i| \leqslant \sum_{j=1}^{i-1} |a_{ij}|\,|x_{jk} - x_j| + \sum_{j=i+1}^{n} |a_{ij}|\,|x_{j,k-1} - x_j|. \qquad (5.29)$$

Let e_k denote the maximum of the numbers $|x_{ik} - x_i|$ as i varies. Then

$$|x_{1k} - x_1| \leqslant \sum_{j=2}^{n} |a_{1j}|\, e_{k-1} \leqslant \mu e_{k-1} < e_{k-1}.$$

Substituting this in (5.29) yields

$$|x_{2k} - x_2| \leqslant |a_{21}|\, e_{k-1} + \sum_{j=3}^{n} |a_{2j}|\, e_{k-1} \leqslant \mu e_{k-1}.$$

Continuing as indicated gives $|x_{ik} - x_i| \leqslant \mu e_{k-1}, 1 \leqslant i \leqslant n$. This means, of course, that $|x_{ik} - x_i| \leqslant \mu^k e_0$, whence, since $0 < \mu < 1$, $\lim_{k\to\infty} x_{ik} = x_i$.

More interesting still than the sufficiency conditions of (5.28) is the fact that convergence always takes place if the matrix **B** of (5.3) is positive definite. To demonstrate this, let $\mathbf{B} = \mathbf{D} + \mathbf{L} + \bar{\mathbf{L}}^t$ where $\mathbf{D} = \bar{\mathbf{D}}$ is the matrix $\text{diag}(b_{11}, b_{22}, \ldots, b_{nn})$, and **L** is the strictly lower-triangular matrix formed from the elements of **B** below the diagonal. Starting from (5.25), it is seen that a necessary and sufficient condition for convergence is that all eigenvalues of $(\mathbf{I} - \mathbf{A}_L)^{-1}\mathbf{A}_R$ be of modulus less than unity. But $\mathbf{A}_L = -\mathbf{D}^{-1}\mathbf{L}$ and $\mathbf{A}_R = -\mathbf{D}^{-1}\mathbf{L}^*$. Thus $(\mathbf{I} - \mathbf{A}_L)^{-1}\mathbf{A}_R = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^*$. The eigenvalues of this matrix, except for sign, are those of $(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^*$, which we consider instead. Let λ_i be an eigenvalue of this matrix, and let $\mathbf{w}_i$ be the corresponding eigenvector. Since **B** is positive definite,

$$(\mathbf{w}_i,\mathbf{B}\mathbf{w}_i) = (\mathbf{w}_i,\mathbf{D}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}^*\mathbf{w}_i) > 0. \quad (5.30)$$

But $(\mathbf{D} + \mathbf{L})^{-1}\mathbf{L}^*\mathbf{w}_i = \lambda_i\mathbf{w}_i$, so that $\mathbf{L}^*\mathbf{w}_i = \lambda_i\mathbf{D}\mathbf{w}_i + \lambda_i\mathbf{L}\mathbf{w}_i$; then

$$(\mathbf{w}_i,\mathbf{L}^*\mathbf{w}_i) = \lambda_i[(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}\mathbf{w}_i)]. \quad (5.31)$$

Taking the conjugate of both sides, $(\mathbf{L}^*\mathbf{w}_i,\mathbf{w}_i) = (\mathbf{w}_i,\mathbf{L}\mathbf{w}_i) = \bar{\lambda}_i[(\mathbf{D}\mathbf{w}_i,\mathbf{w}_i) + (\mathbf{L}\mathbf{w}_i,\mathbf{w}_i)]$, or

$$(\mathbf{w}_i,\mathbf{L}\mathbf{w}_i) = \bar{\lambda}_i[(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}^*\mathbf{w}_i)]. \quad (5.32)$$

Combining (5.31) and (5.32) gives

$$(1 - \lambda_i\bar{\lambda}_i)(\mathbf{w}_i,\mathbf{L}^*\mathbf{w}_i) = (\lambda_i + \lambda_i\bar{\lambda}_i)(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i),$$

$$(1 - \lambda_i\bar{\lambda}_i)(\mathbf{w}_i,\mathbf{L}\mathbf{w}_i) = (\bar{\lambda}_i + \lambda_i\bar{\lambda}_i)(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i).$$

Substituting in $(1 - \lambda_i\bar{\lambda}_i)[(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}\mathbf{w}_i) + (\mathbf{w}_i,\mathbf{L}^*\mathbf{w}_i)]$ yields

$$(1 + \lambda_i)(1 + \bar{\lambda}_i)(\mathbf{w}_i,\mathbf{D}\mathbf{w}_i).$$

Since **D** is itself positive definite, this expression is positive, Then, by (5.30), $(1 - \lambda_i\bar{\lambda}_i) > 0$ or $|\lambda_i| < 1$.

Thus, sufficiency has been shown. It is also possible to prove that if the matrix **B** is Hermitian and all diagonal elements are positive, then convergence requires that **B** be positive definite.

The solution of systems of equations by iterative procedures such as the Jacobi and Gauss-Seidel methods is sometimes termed *relaxation* (the errors in the initial estimate of the solution vector are decreased or relaxed as calculation continues). The Gauss-Seidel and related methods are used extensively in the solution of large systems of linear equations, generated as the result of the finite-difference approximation of partial differential equations (see Chapter 7).