7 Techniques for linear systems of equations

We consider the following system of n linear equations in n unknowns.

$$a_{i1}x_1 + a_{i2} + \cdots + a_{in}x_n = b_i, \;\; i = 1, 2, \ldots, n.$$

Which can be presented in matrix form as

$$\mathbf{A}\boldsymbol{x} = \mathbf{b} \tag{81}$$

where

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

is the coefficient matrix, $\mathbf{x} = (x_1, x_2, \cdots, x_n)^\top$ is the vector of unknowns and $\mathbf{b} = (b_1, b_2, \cdots, b_n)^\top$ is the known right hand side. The augmented matrix for the system is

$$\tilde{\mathbf{A}} = \left(\begin{array}{cccc|c} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{array}\right) \tag{82}$$

In this section we discuss methods of solving such systems other than using the inverse of $\mathbf{A}$ which is an expensive process. The methods are grouped into two categories: *direct methods* and *indirect methods*.

7.1 Direct Methods

Direct methods obtain the solution in a finite number of steps and return an exact solution if all computations are done in infinite precision. The idea is that a triangular system is easy to solve by applying backward substitution (in case of an upper triangular system). How can we then reduce the given system to an equivalent triangular system?

7.1.1 Gaussian Elimination

We assume that the system has a unique solution. The method reduces the system to an upper triangular system using elementary row operations.

Example 7.1 *Solve the following using Gaussian Elimination*

$$\begin{aligned} x_1 + 2x_2 &= -1 \\ 2x_1 + x_2 + x_3 &= 3 \\ -2x_1 - x_2 + x_3 &= 1 \end{aligned}$$

Solution: The augmented matrix is

$$\tilde{\mathbf{A}} = \left(\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 2 & 1 & 2 & 3 \\ -2 & -1 & 1 & 1 \end{array} \right)$$

Reducing $\tilde{\mathbf{A}}$ to upper triangular gives

$$\left(\begin{array}{ccc|c} 1 & 2 & 0 & -1 \\ 0 & 1 & \frac{-1}{3} & \frac{-5}{3} \\ 0 & 0 & 1 & 2 \end{array} \right).$$

Solving by back substitution gives: $x_3 = 2$, $x_2 = -1$, $x_1 = 1$. Or the solution is $\mathbf{x} = (1, -1, 2)^T$. The gaussin elimination method is to systematically reduce the system (82) into a triangular system through a sequence of finite steps as described below.

Let $A^{(1)}$ denote the initial augmented matrix (82). Re-arrange $A^{(1)}$ so that $a_{11} \neq 0$. If this is not possible, then $det(A) = 0$ meaning the system is singular. Otherwise we have

$$A^{(1)} = \left(\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{array} \right)$$

Define the $n-1$ multipliers

$$m_{i1}^{(1)} = -\frac{a_{i1}^{(1)}}{a_{11}^{(1)}},\ i > 1.$$

Then reduce $A^{(1)} \to A^{(2)}$ using the procedure

$$m_{i1}^{(1)} R_1 + R_i \to R_i \, , \; i > 1$$

Example 7.2 *From the previous example,*

$$m_{21} = -\frac{a_{21}}{a_{11}} = -\frac{2}{1}.$$

□

That is, by elementary row operations, we reduce all the entries below $a_{11}^{(1)}$ to zero so that we have

$$A^{(2)} = \left(\begin{array}{cccc|c} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{array} \right)$$

Next define the $n-2$ multipliers

$$m_{i2}^{(2)} = -\frac{a_{i2}^{(2)}}{a_{22}^{(2)}}, \; i > 2,$$

and then reduce $A^{(2)} \to A^{(3)}$ using the operations

$$m_{i2}^{(2)} R_2 + R_i \to R_i \, , \; i > 2.$$

This will generate

$$A^{(3)} = \left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} & b_n^{(3)} \end{array} \right)$$

Assume $a_{33} \neq 0$, obtain $A^{(4)}$ using the same procedure. Continue until you have

$$A^{(n)} = \left(\begin{array}{ccccc|c} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{3n}^{(3)} & b_3^{(3)} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn}^{(n)} & b_n^{(n)} \end{array}\right)$$

$A^{(n)}$ is upper triangular and $\det(A^{(n)}) \neq 0$.

Definition 7.1 *The numbers* $a_{11}^{(1)}, a_{22}^{(2)}, a_{33}^{(3)}, \ldots, a_{nn}^{(n)}$ *are called* pivot elements.

The system can now be solved using backsubstitution. That is,

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}$$

and

$$x_i = \frac{1}{a_{ii}^{(i)}} \left(b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij} x_j \right), i = n-1, n-2, \ldots, 1.$$

Exercise 7.1.1-1: Solve using Gaussian Elimination:

$$\begin{aligned} x_1 + x_2 + x_3 &= -1 \\ 2x_1 + x_2 + x_3 &= -8 \\ 4x_1 + 6x_2 + 8x_3 &= 14 \end{aligned}$$

Definition 7.2 (Pivoting) *The process of swapping two rows to avoid a zero pivot is called pivoting*

If the pivot element is so small compared to the elements below it in the same column, we accumulate rounding errors and in the process we may end up with a completely wrong solution. Partial pivoting is used to reduce on the amount of rounding errors when we use Gaussian Elimination.

Definition 7.3 (Partial Pivoting) *The process of swapping rows to ensure that in each column the pivot element is the largest in absolute value of all elements in that column is called partial pivoting.*

Example 7.3 Solving

$$\begin{aligned} 0.003x_1 + 59.14x_2 &= 59.17 \\ 5.291x_1 - 6.130x_2 &= 46.78 \end{aligned}$$

without pivoting using three dcimal places yields $x_1 = -10.00$, $x_2 = 1.001$, a wrong solution. Such errors that arise because a computer can only store real numbers with a finite precision are called rounding errors.

7.1.2 Triangular LU Decomposition

Given a system

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where $\mathbf{A}$ is a nonsingular matrix, we *decompose/factorise* $\mathbf{A}$ as

$$\mathbf{A} = \mathbf{L}\mathbf{U}$$

where $\mathbf{L}$ is *lower triangular* and $\mathbf{U}$ is *upper triangular*. Then the system becomes

$$\mathbf{L}\mathbf{U}\mathbf{x} = \mathbf{b} \tag{83}$$

For a unique factorisation of $\mathbf{A}$, we fix $\mathbf{L}$ to be *unit lower triangular* (with only ones on the main diagonal) and the resulting triangular decomposition is more specifically known as the Doolittle factorisation. Then, let

$$\mathbf{U}\mathbf{x} = \mathbf{y} \tag{84}$$

so that

$$\mathbf{L}\mathbf{y} = \mathbf{b}. \tag{85}$$

We solve equation (85) by forward substitution to obtain $\mathbf{y}$ and then use $\mathbf{y}$ in equation (84) to obtain $\mathbf{x}$ by backward substitution. If $\mathbf{A}$ is symmetric and positive definite, then we can find the decomposition $\mathbf{L}\mathbf{U} = \mathbf{A}$ where $\mathbf{U} = \mathbf{L}^t$. This particular triangular decomposition, that is, $\mathbf{L}\mathbf{L}^t = \mathbf{A}$ is the Choleski decomposition of $\mathbf{A}$. Further if $\mathbf{L}\mathbf{U} = \mathbf{A}$ where $\mathbf{U}$ is unit upper triangular, then the decomposition is called a Crout decomposition.

Example 7.4 Solve using triangular decomposition

$$\begin{aligned} x_1 + 2x_2 &= -1 \\ 2x_1 + x_2 + x_3 &= 3 \\ -2x_1 - x_2 + x_3 &= 1 \end{aligned}$$

Solution: We have the system

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \\ -2 & -1 & 1 \end{pmatrix} \mathbf{x} = \begin{pmatrix} -1 \\ 3 \\ 1 \end{pmatrix}$$

Let $\mathbf{A} = \mathbf{LU}$, that is

$$\begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 1 \\ -2 & -1 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}.$$

Multiplying out the right hand side gives

$$\mathbf{A} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{pmatrix}.$$

By equality of matrices, from the first row we get,

$$u_{11} = 1\,,\ u_{12} = 2\,,\ u_{13} = 0.$$

From the second row,

$$\begin{aligned} l_{21}u_{11} = 2 \qquad &\Longrightarrow l_{21} = 2, \\ l_{21}u_{12} + u_{22} = 1 &\Longrightarrow u_{22} = 1 - (2)(2) = -3, \\ l_{21}u_{13} + u_{23} = 1 &\Longrightarrow u_{23} = 1. \end{aligned}$$

From the third row,

$$\begin{aligned} l_{31}u_{11} = -2 \qquad &\Longrightarrow l_{31} = -2, \\ l_{31}u_{12} + l_{32}u_{22} = -1, &\Longrightarrow l_{32} = -1(-1+4)/3 = -1, \\ l_{32}u_{13} + l_{32}u_{23} + u_{33} = 1, &\Longrightarrow u_{33} = 2. \end{aligned}$$

Hence we have

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -2 & -1 & 1 \end{pmatrix} \text{ and } \mathbf{U} = \begin{pmatrix} 1 & 2 & 0 \\ 0 & -3 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

Solving $\mathbf{Ly} = \mathbf{b}$, that is,

$$\begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -2 & -1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 3 \\ 1 \end{pmatrix},$$

by forward substitution gives $y_1 = -1, y_2 = 5$ and $y_3 = 4$.
Then solving $\mathbf{Ux} = \mathbf{y}$, that is,

$$\begin{pmatrix} 1 & 2 & 0 \\ 0 & -3 & 1 \\ 0 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \\ 4 \end{pmatrix},$$

by back substitution gives $x_3 = 2, x_2 = -1$ and $x_1 = 1$, that is, $\mathbf{x} = (1, -1, 2)^T$.

Exercise 7.1.2-2:
Solve by LU decomposition:

$$\begin{array}{rcrcrcrcr} x_1 & + & 5x_2 & + & 4x_3 & + & 3x_4 & = & -5 \\ 2x_1 & + & 7x_2 & + & 6x_3 & + & 10x_4 & = & 7 \\ 3x_1 & + & 2x_2 & + & 8x_3 & + & 255x_4 & = & 65 \\ 2x_1 & + & x_2 & - & 2x_3 & + & 27x_4 & = & 20 \end{array}$$

7.2 Iterative methods for linear systems

Iterative methods for

$$\mathbf{Ax} = \mathbf{b} \tag{86}$$

are of the form

$$\mathbf{x}^{(k+1)} = \mathbf{Gx}^{(k)} + \mathbf{c}, k = 0, 1, 2, \ldots. \tag{87}$$

The matrix $\mathbf{G}$ is the called the *iteration matrix* and the vector $\mathbf{c}$ is called the *iteration vector*. The scheme generates a sequence of vectors $\mathbf{x}^{(k)}$, $k = 0, 1, 2, \ldots$, which if converges, it converges to the solution $\mathbf{x}$ of the system. The principal of the schemes discussed here is the same but the difference is in how the iteration matrix $\mathbf{G}$ and vector $\mathbf{c}$ are formulated from $\mathbf{A}$ and $\mathbf{b}$.

7.2.1 Jacobi Iteration Scheme

In this case we decompose $\mathbf{A}$ into

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

where $\mathbf{L}$ is the strictly lower part of $\mathbf{A}$, $\mathbf{D}$ is the diagonal of $\mathbf{A}$ and $\mathbf{U}$ is the strictly upper part of $\mathbf{A}$. Then (86) becomes

$$(\mathbf{L} + \mathbf{D} + \mathbf{U})\mathbf{x} = \mathbf{b}. \tag{88}$$

This can be rewritten as

$$(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}\mathbf{x} = \mathbf{b}$$

to give

$$\mathbf{D}\mathbf{x} = -(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{b}.$$

Hence we have

$$\mathbf{x} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x} + \mathbf{D}^{-1}\mathbf{b}.$$

We now set

$$\mathbf{x}^{(k+1)} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b}, k = 0, 1, 2, \ldots. \tag{89}$$

Equation (89) is in the form

$$\mathbf{x}^{(k+1)} = \mathbf{G}_j\mathbf{x}^{(k)} + \mathbf{c}_j$$

and is the Jacobi iteration scheme where

$$\mathbf{G}_j = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$$

is the *Jacobi iteration matrix* and

$$\mathbf{c}_j = \mathbf{D}^{-1}\mathbf{b}$$

is the *Jacobi iteration vector*.

Example 7.5 Solve using Jacobi iteration

$$\begin{array}{rcrcrcl} 10x_1 & + & x_2 & + & x_3 & = & 24 \\ -x_1 & + & 20x_2 & + & x_3 & = & 21 \\ x_1 & + & -2x_2 & + & 100x_3 & = & 300 \end{array}$$

Solution: From

$$\mathbf{A} = \begin{pmatrix} 10 & 1 & 1 \\ -1 & 20 & 1 \\ 1 & -2 & 100 \end{pmatrix} \text{ and } \mathbf{b} = \begin{pmatrix} 24 \\ 21 \\ 300 \end{pmatrix},$$

we gget

$$\mathbf{L} = \begin{pmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ 1 & -2 & 0 \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 100 \end{pmatrix}, \text{ and } \mathbf{U} = \begin{pmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

We can now form

$$\mathbf{G}_j = -\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U}) = -\begin{pmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{20} & 0 \\ 0 & 0 & \frac{1}{100} \end{pmatrix}\begin{pmatrix} 0 & 1 & 1 \\ -1 & 0 & 1 \\ 1 & -2 & 0 \end{pmatrix} = -\begin{pmatrix} 0 & \frac{1}{10} & \frac{1}{10} \\ \frac{-1}{20} & 0 & \frac{1}{20} \\ \frac{1}{100} & \frac{-2}{100} & 0 \end{pmatrix}$$

and

$$\mathbf{c}_j = \mathbf{D}^{-1}\mathbf{b} = \begin{pmatrix} \frac{1}{10} & 0 & 0 \\ 0 & \frac{1}{20} & 0 \\ 0 & 0 & \frac{1}{100} \end{pmatrix}\begin{pmatrix} 24 \\ 21 \\ 300 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix}$$

The Jacobi iterations for the problem are then given by

$$\mathbf{x}^{(k+1)} = \begin{pmatrix} 0 & -1/10 & -1/10 \\ 1/20 & 0 & -1/20 \\ -1/100 & 2/100 & 0 \end{pmatrix}\mathbf{x}^{(k)} + \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix}, k = 0, 1, 2, \ldots. \tag{90}$$

Let

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

then

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0 & -1/10 & -1/10 \\ 1/20 & 0 & -1/20 \\ -1/100 & 2/100 & 0 \end{pmatrix}\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix},$$

$$\mathbf{x}^{(2)} = \begin{pmatrix} 0 & -1/10 & -1/10 \\ 1/20 & 0 & -1/20 \\ -1/100 & 2/100 & 0 \end{pmatrix} \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix} + \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix} = \begin{pmatrix} 1.995 \\ 1.02 \\ 2.997 \end{pmatrix},$$

$$\mathbf{x}^{(3)} = \begin{pmatrix} 0 & -1/10 & -1/10 \\ 1/20 & 0 & -1/20 \\ -1/100 & 2/100 & 0 \end{pmatrix} \begin{pmatrix} 1.995 \\ 1.02 \\ 2.997 \end{pmatrix} + \begin{pmatrix} 2.4 \\ 1.05 \\ 3 \end{pmatrix} = \begin{pmatrix} 1.9983 \\ 0.9999 \\ 3.0005 \end{pmatrix}$$

Continuing this way we find that the iteration converges to:

$$\mathbf{x} = \begin{pmatrix} 2.000 \\ 1.000 \\ 3.000 \end{pmatrix}.$$

7.2.2 Gauss-Seidel Iteration

In this case, after decomposing (86) into the form (88), we write is as

$$(\mathbf{L} + \mathbf{D})\mathbf{x} = -\mathbf{U}\mathbf{x} + \mathbf{b}.$$

This is rearranged to obtain

$$\mathbf{x} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{x} + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}.$$

We set

$$\mathbf{x}^{(k+1)} = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}, k = 0, 1, 2, \ldots. \tag{91}$$

Equation (91) is in the form

$$\mathbf{x}^{(k+1)} = \mathbf{G}_g\mathbf{x}^{(k)} + \mathbf{c}_g$$

and is the Gauss-Seidel iteration scheme, where,

$$\mathbf{G}_g = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}$$

is the *Gauss-Seidel iteration matrix* and

$$\mathbf{c}_g = (\mathbf{L} + \mathbf{D})^{-1}\mathbf{b}$$

is the *Gauss-Seidel iteration vector*.

Example 7.6 Solve the previous example using Gauss Seidel iteration.

Solution:

$$\mathbf{G}_g = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U} = \begin{pmatrix} 0 & -1/10 & 1/10 \\ 0 & -1/200 & -11/200 \\ 0 & 9/10000 & -1/10000 \end{pmatrix}$$

and

$$\mathbf{c}_g = (\mathbf{L}+\mathbf{D})^{-1}\mathbf{b} = \begin{pmatrix} 2.4 \\ 1.17 \\ 1.9994 \end{pmatrix}.$$

Let

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

then

$$\mathbf{x}^{(1)} = \begin{pmatrix} 0 & -1/10 & 1/10 \\ 0 & -1/200 & -11/200 \\ 0 & 9/10000 & -1/10000 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} + \begin{pmatrix} 2.4 \\ 1.17 \\ 1.9994 \end{pmatrix} = \begin{pmatrix} 2.4 \\ 1.17 \\ 1.9994 \end{pmatrix},$$

$$\mathbf{x}^{(2)} = \begin{pmatrix} 0 & -1/10 & 1/10 \\ 0 & -1/200 & -11/200 \\ 0 & 9/10000 & -1/10000 \end{pmatrix} \begin{pmatrix} 2.4 \\ 1.17 \\ 1.9994 \end{pmatrix} + \begin{pmatrix} 2.4 \\ 1.17 \\ 1.9994 \end{pmatrix}$$

Continuing this way we find that the iteration converges to

$$\mathbf{x} = \begin{pmatrix} 2.000 \\ 1.000 \\ 3.000 \end{pmatrix}.$$

7.3 Convergence of the iteration scheme

Observe that the iterations above are actually fixed point iterative schemes. Thus, If $\boldsymbol{x}^*$ is the solution then,

$$\boldsymbol{x}^* = \mathbf{G}\boldsymbol{x}^* + \mathbf{c}. \tag{92}$$

Subtracting (92) from (87) gives

$$\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^* = \mathbf{G}(\boldsymbol{x}^{(k)} - \boldsymbol{x}^*). \tag{93}$$

Define

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x}^*$$

as the error in $\boldsymbol{x}^{(k)}$. Then from (93) we have

$$\boldsymbol{e}^{(k+1)} = \mathbf{G}\boldsymbol{e}^{(k)} = \mathbf{G}^2\boldsymbol{e}^{(k-1)} = \mathbf{G}^3\boldsymbol{e}^{(k-2)} = \cdots = \mathbf{G}^{k+1}\boldsymbol{e}^{(0)}. \tag{94}$$

where $\boldsymbol{e}^{(0)}$ is the error in the initial guess to the solution. This helps us prove the following theorem

Theorem 7.1 (Convergence if G is convergent) *The iterative scheme* (87) *converges to a limit with any arbitrary choice of the initial approximation* $\mathbf{x}^{(0)}$ *if and only if* $\mathbf{G}$ *is a convergent matrix.*

Proof:
If $\mathbf{G}$ is a convergent matrix, then $\mathbf{G}^n \to \mathbf{O}$ as $n \to \infty$, where $\mathbf{O}$ is the zero matrix. Then using this in (93) gives

$$\boldsymbol{e}^{(k+1)} = \mathbf{G}^{k+1}\boldsymbol{e}^{(0)} \to \mathbf{O}\boldsymbol{e}^{(0)} = \mathbf{0} \text{ as } k \to \infty.$$

Hence the scheme converges. On the other hand, since $\boldsymbol{e}^{(0)}$ is fixed at the begining of the scheme, then $\mathbf{G}^{k+1}\boldsymbol{e}^{(0)} \to \mathbf{0} \implies \mathbf{G}^{(k+1)} \to \mathbf{O}$ as $k \to \infty$. □

The spectral radius of the iteration matrix $\mathbf{G}$ is also fundamental in determining the convergence of the scheme (87).

Theorem 7.2 (Convergence if $\rho(\mathbf{G}) < 1$) *The iterative scheme* (87) *converges if and only if the spectral radius of* $\mathbf{G}$ *is less than* 1*, that is* $\rho(\mathbf{G}) < 1$.

Proof:
Let

$$\mathbf{G}\mathbf{V} = \mathbf{V}\boldsymbol{\Lambda}$$

be an eigen value decomposition of $\mathbf{G}$, that is, $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{G}$ and $\mathbf{V}$ is a matrix whose columns are eigenvectors of $\mathbf{G}$. For convergence, we have that

$$\boldsymbol{e}^{(k)} = \mathbf{G}^k\boldsymbol{e}^{(0)} \to \mathbf{0} \text{ as } k \to \infty.$$

This means that $\mathbf{G}^k \to \mathbf{O}$ as $k \to \infty$ since $\boldsymbol{e}^{(0)}$ is fixed by the initial guess. Since $\mathbf{G}^k$ has the same eigenvectors as $\mathbf{G}$, to be precise,

$$\mathbf{G}^k\boldsymbol{V} = \boldsymbol{V}\boldsymbol{\Lambda}^k,$$

then $\mathbf{G}^k \to \mathbf{O}$ if and only if $\boldsymbol{\Lambda} \to \mathbf{O}$. This means that $\lambda_i \to 0$ as $k \to \infty$ for all i, since $\boldsymbol{\Lambda}$ is a diagonal matrix of eigenvalues of $\mathbf{G}$, $\lambda_i, i = 1, 2, 3, \cdots, n$. That is to say, we have convergence if and only if $|\lambda_i| < 1 \ \forall\, i$. Thus we have convergence if and only if $\rho(\mathbf{G}) < 1$.

Rate of convergence

Now, suppose $\mathbf{G}$ has n eigenvectors and eigenvalues $\boldsymbol{V}_i$ and λ_i respectively, $i = 1, 2, 3, \ldots, n$. We can use the $\boldsymbol{v}_i$'s as a basis for $\boldsymbol{e}^{(0)}$ so that

$$\boldsymbol{e}^{(0)} = \sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i, \tag{95}$$

for some $\alpha_i \in \mathbb{R}, i = 1, 2, \ldots, n$. Then

$$\boldsymbol{e}^{(1)} = \mathbf{G}\boldsymbol{e}^{(0)} = \mathbf{G}\sum_{i=1}^{n} \alpha_i \boldsymbol{v}_i = \sum_{i=1}^{n} \alpha_i \mathbf{G}\boldsymbol{v}_i = \sum_{i=1}^{n} \alpha_i \lambda_i \boldsymbol{v}_i.$$

Similarly

$$\boldsymbol{e}^{(2)} = \mathbf{G}\boldsymbol{e}^{(1)} = \mathbf{G}\sum_{i=1}^{n} \alpha_i \lambda_i \boldsymbol{v}_i = \sum_{i=1}^{n} \alpha_i \lambda_i \mathbf{G}\boldsymbol{v}_i = \sum_{i=1}^{n} \alpha_i \lambda_i^2 \boldsymbol{v}_i.$$

Proceeding in a similar way gives

$$\boldsymbol{e}^{(k)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{v}_i. \tag{96}$$

This further shows that $\boldsymbol{e}^{(k)} \to \mathbf{0}$ as $k \to \infty$ if and only if $\lambda_i^k \to 0$ as $k \to \infty \ \forall\, i$. In particular, the behaviour of $\boldsymbol{e}^{(k)}$ will be dominated by the largest λ_i, that is, if

$$\rho = \max_i \ |\lambda_i|,$$

then ρ determines the rate of convergence of the scheme. The smaller the value of ρ, the faster the convergence.

Example 7.7 How many iterations does it take to reduce the initial error by a factor 10^{-3}?

Solution: Let ρ be the spectral radius of $\mathbf{G}$. Then show that

$$\boldsymbol{e}^{(k)} = \rho^k \boldsymbol{e}^{(0)} \implies \|\boldsymbol{e}^{(k)}\| = \rho^k \|'\boldsymbol{e}^{(0)}\|.$$

So after the k^{th} iteration, we want that $\rho^k = 10^{-3}$, $\implies k = \log 10^{-3}/\log \rho$

Definition 7.4 *The number* $\log \rho$ *is called the* speed of convergence.

Theorem 7.3 (Convergence due to Diagonal dominance) *If* $\mathbf{A}$ *is strictly diagonally dominant, then the sequence of solutions produced by either Jacobi or Gauss-Seidel iteration converges to the solution of* $\mathbf{Ax} = \mathbf{b}$ *for any* $\mathbf{x}^{(0)}$.

Proof: A square matrix $\mathbf{A}$ is said to be *strictly diagonally dominant* if

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| \quad \text{for all } i = 1, 2, 3, \ldots, n.$$

For Jacobi,

$$\mathbf{G} = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}),$$

then

$$\|\boldsymbol{x}\|_\infty = \frac{\|\mathbf{L} + \mathbf{U}\|_\infty}{\|\mathbf{D}\|_\infty}$$

$$= \frac{\max\limits_{1\leq i\leq n} \sum\limits_{j=1, j\neq i} |a_{ij}|}{\max\limits_{1\leq i\leq n} |a_{ii}|} < 1$$

since $\mathbf{A}$ is strictly diagonally dominant.

Exercise 7.3-1: Check that

$$A = \begin{pmatrix} 3 & -1 & 1 \\ 1 & -4 & 2 \\ -2 & -1 & 5 \end{pmatrix}$$

is strictly diagonally dominant.

Exercise 7.3-2:

(a) Consider the iteration scheme (87). Show that $\|\boldsymbol{e}^{(k+1)}\| \leq \|\mathbf{G}\|^k \|\boldsymbol{e}^{(0)}\|$.

(b) Check for the convergence of the Jacobi and Gauss-Seidel schemes for linear systems whose:

(i)

$$\mathbf{A} = \begin{pmatrix} 3 & 3 & 3 \\ 0 & 2 & 1 \\ 1 & 0 & 2 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 6 \\ 5 \\ 4 \end{pmatrix}$$

(ii)

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ 1 & 2 & 3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 6 \\ 3 \\ 3 \end{pmatrix}$$

(c) Verify that the spectral radius for Gauss-Seidel is less than that for Jacobi, that is, $\rho(\mathbf{G}_g) < \rho(\mathbf{G}_j)$. The implication of this is that the Gauss-Seidel scheme converges faster than the Jacobi scheme.