Numerical Methods - MA 204

Syllabus:

Interpolation by polynomials, divided differences, error of the interpolating polynomial, piecewise linear and cubic spline interpolation,
Numerical Integration, Composite rule, error formulae,
Solution of a system of linear equations, implementation of Gaussian elimination, and Gauss-Seidel Methods, partial pivoting, row echelon form, LU factorization Cholesky's method, ill-conditioning, norms.
Solution of Non-linear equations, bisection and secant methods.
Newton's method, rate of convergence, solution of a system of nonlinear equations, numerical solution of ordinary differential equations, euler and Runge-Kutta methods, multi-step methods, predictor-corrector methods, order of convergence, finite difference methods, numerical solutions of elliptic, parabolic and hyperbolic partial differential equations.
Eigen-value problem, power method, QR method, Gershgorin's theorem.
Exposure to software packages like IMSL subroutines, MATLAB

Suggested Readings:

1. S.D. Conte and Carle de Boor, Elementary Numerical Analysis An Algorithmic Approach (3rd Edition), McGraw-Hill, 1980.

2. Carl E. Froberg, Introduction to Numerical Analysis (2nd Edition), Addison-Wesley, 1981.

3. E. Kreyszig, Advanced Engineering Mathematics (8th Edition), John Wiley and Sons, 1999.

4. D. Watkinson, Fundamentals of Matrix Computations, Wiley- Interscience (2nd edition), 2002.

5. M.K. jain, S.R.K. Iyengar, R.K. Jain, Numerical Methods (6th Edition), New Age International (P) Ltd, 2012.

These are some **partial, rough** notes prepared for MA-204 students at IIT Indore. ***Mistakes and typos*** are bound to be there. Students are suggested to read them carefully and are encouraged to send their comments and suggestions.

Ashisha Kumar
Discipline of Mathematics
I.I.T. Indore.

Contents

1 Introduction

There are many mathematical problems, which are formulated while solving problems from other sciences. Some of these problems can be solved by methods which we have learned in the course of Calculus, Differential Equations and Linear algebra etc. And we are happy to get exact solutions or some information which is showing the analytic behavior of the solution. But in most of the practical purpose we can not get exact solutions by direct computations (may be because of round off errors) or the solutions in known compact forms may not exist or even the solution in compact forms is not sufficient and we want some numeric value as the solution. In this course our aim is to solve such mathematical problems using numerical methods.

2 Interpolation by polynomials

Suppose we want to know about the value of a function (or some of its derivative) at a particular point. But only information we know about the function is that we know its value (or the value of its derivatives) at certain other points. It might be some arbitrary continuous function. And we want to approximate the function by some polynomial, which agrees with the known data about the function. The reason of approximating the given functions by polynomial is that we can compute the value of a polynomial at a point just by using basic operations (addition subtraction and multiplication) of computer.

In this chapter our aim is to find an approximating polynomial to a function, whose value or the value of its certain derivatives is known at some points. These points are called nodes. To ease the computations we always try to find the minimal degree polynomial, which satisfies the given data.

Problem 2.1. Find a polynomial, which coincide with $|x|$ at the points $-1, 0$ and 1.

It means we need to find a polynomial P, which should coincide with the function $|x|$ at the points $-1, 0$ and 1. So we must have

$$P(-1) = |-1| = 1, P(0) = |0| = 0, P(1) = |1| = 1 \tag{2.1}$$

Now since there are three distinct points, we will try to find a two degree polynomial, because the degree of freedom of two degree polynomial is three. (The vector space of at most two degree polynomial is three dimensional and $\{1, x, x^2\}$ is a basis for this vector space.)

Let $P(x) = a + bx + cx^2$ be a polynomial, which passes through the points $(-1, 1), (0, 0), (1, 1)$. Then we must have

$$\begin{aligned} a + b(0) + c(0) &= 0 \\ a + b(-1) + c(-1)^2 &= 1 \\ a + b(1) + c(1)^2 &= 1 \end{aligned} \tag{2.2}$$

Remark 2.1. Here we need to solve a system of three linear equations in three unknowns. If we would have approximated it by a polynomial of degree less than two, then the system of three linear equations in two unknowns might not have any solution. And similarly if we would have approximated it by a four degree polynomial, then this system of linear equations might have infinite number of solutions. But because of increased degree of polynomial, computations will be difficult.

From (2.2), we must have

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} \tag{2.3}$$

Since the $\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}\right) = -2$, and the $Adj \begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & -1 \\ 2 & -1 & -1 \end{bmatrix}$. The inverse of the matrix is given by

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & 1 & 1 \end{bmatrix}^{-1} = -\frac{1}{2} \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & -1 \\ 2 & -1 & -1 \end{bmatrix}. \tag{2.4}$$

And hence

$$\begin{bmatrix} a \\ b \\ c \end{bmatrix} = -\frac{1}{2} \begin{bmatrix} -2 & 0 & 0 \\ 0 & 1 & -1 \\ 2 & -1 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{2.5}$$

Thus the two degree polynomial which coincide with $|x|$ at $(-1, 0, 1)$ is given by x^2.

Remark 2.2. In general there are large number of data points with numerically rich data.

- We need to compute the inverse of a large matrix.
- The solution of the system of linear equations may be far from that of exact solution due to round off errors.

Suppose we know the value of a continuous function at two points x_0, x_1, that is, we know $f(x_0), f(x_1)$.

Question 2.1. Can we find a polynomial, which satisfies $P(x_0) = f(x_0)$, $P(x_1) = f(x_1)$?

And the answer is very easy that we learned in our 10th standard. We find the equation of a line passing through $(x_0, f(x_0))$ and $(x_1, f(x_1))$, because the equation of line is a polynomial of degree one. I am sure that you know many versions of the formula of finding the equation of a line passing through two points. But we try to find this in some different way.

$$P(x) = ax + b; f(x_0) = ax_0 + b; f(x_1) = ax_1 + b;$$

or

$$P(x)(-1) + ax + (1)b = 0; f(x_0)(-1) + ax_0 + (1)b = 0; f(x_1)(-1) + ax_1 + (1)b = 0;$$

Now if you recall your matrix representations of system of linear equations from Linear Algebra course, we can write above system of equations as

$$\begin{bmatrix} P(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{bmatrix} \begin{bmatrix} -1 \\ a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{2.6}$$

Now you know that if for any particular value of x say x' the 3×3 matrix $\begin{bmatrix} P(x') & x' & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{bmatrix}$ is invertible, then by multiplying the inverse of this matrix on both sides of the equation (2.6), we get a contradiction that the matrix $\begin{bmatrix} -1 \\ a \\ b \end{bmatrix}$ is a null matrix. This implies that the matrix $\begin{bmatrix} P(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{bmatrix}$ is a singular matrix for every value of x. And hence, the determinant of this matrix is zero for all values of x. Then by simplifying the determinant $\begin{vmatrix} P(x) & x & 1 \\ f(x_0) & x_0 & 1 \\ f(x_1) & x_1 & 1 \end{vmatrix}$ with respect to first column, we get

$$P(x)(x_0 - x_1) + f(x_0)(x_1 - x) + f(x_1)(x - x_0) = 0.$$

Rewriting it we get

$$P(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1). \tag{2.7}$$

Observation 2.1. Now if we want to find a polynomial $P(x)$, which agrees with a function $f(x)$ at three distinct points x_0, x_1, x_2, that is, $P(x)$ satisfies $P(x_0) = f(x_0)$, $P(x_1) = f(x_1)$ and $P(x_2) = f(x_2)$, and if we consider the two degree polynomial $P(x) = a + bx + cx^2 + dx^3$ and proceed similar to above case then we will come to a conclusion that the determinant of the matrix $\begin{bmatrix} P(x) & x^2 & x & 1 \\ f(x_0) & x_0^2 & x_0 & 1 \\ f(x_1) & x_1^2 & x_1 & 1 \\ f(x_2) & x_2^2 & x_2 & 1 \end{bmatrix}$ is equal to zero for all values of x in the interval. By simplification with respect to first column we get.

$$P(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f(x_0) + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f(x_1) + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f(x_2). \tag{2.8}$$

Now if we write

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}, \\ L_1(x) &= \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}, \\ L_2(x) &= \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}, \end{aligned}$$

then we observe the following properties of these functions.

- $L_0 + L_1 + L_2 = 1$,
- $L_j(x_i) = \delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$,
- The degree of each L_j is two, and
- $P(x) = L_0(x)f(x_0) + L_1(x)f(x_1) + L_2(x)f(x_2)$.

The polynomials $L_j's$ are called Lagrange's fundamental polynomial. And the corresponding polynomial $P(x)$ is called Lagrange's polynomial.

Problem 2.2. Let $[a, b]$ is a given interval and f is some unknown continuous function defined on $[a, b]$, if we know the the value of the function at $n+1$ distinct points $a = x_0 < x_1 < \ldots < x_n = b$, that is we know $f(x_i)$ at $i = 0, \ldots n$. Can we find a polynomial

$$P(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n, \tag{2.9}$$

such that $P(x_i) = f(x_i)$ for $i = 0, \ldots n$?

2.1. Lagrange Polynomial

Using the similar procedure as in the above two examples J.L. Lagrange found $P(x)$, but we usually denote it $L(x)$ to give respect to his name.

$$L(x) = \sum_{j=0}^{n} \left(\prod_{i=0, i\neq j}^{n} \frac{(x-x_i)}{(x_j-x_i)} \right) f(x_j) = \sum_{j=0}^{n} L_j(x) f(x_j), \tag{2.10}$$

where $L_j(x)$ is j_{th} fundamental polynomial of degree n and is given by

$$L_j(x) = \prod_{i=0, i\neq j}^{n} \frac{(x-x_i)}{(x_j-x_i)}. \tag{2.11}$$

These L_j's satisfy the following properties.

- If we write $\omega(x) = \prod_{j=0}^{n}(x - x_j)$, then Lagrange's fundamental polynomial can also be expressed as
$$L_j(x) = \frac{\omega(x)}{(x-x_j)\omega'(x_j)}. \tag{2.12}$$
- $\sum_{j=0}^{n} L_j(x) = 1$,
- $L_j(x_k) = \delta_{jk} = \begin{cases} 1 & \text{if} \quad j = k \\ 0 & \text{if} \quad j \neq k \end{cases}$,

- the degree of each L_j is n, and
- $L(x) = \sum_{j=0}^{n} L_j(x) f(x_j)$. And
- the degree of $L(x)$ is at most n.

Remark 2.3. There are examples of data in which $L(x)$ is of degree strictly less than n.

Remark 2.4. We can use the the following steps to find Lagrange polynomial for the given data.

a) First we compute Lagrange's fundamental polynomial L_j for each j.

b) Multiply each $L_j(x)$ with the corresponding functional value $f(x_j)$.

c) Then sum these multiplications obtained in step b) over all $j's$.

Remark 2.5. If we draw the graphs of both $f(x)$ and $P(x)$, then it is clear from the conditions that these two graphs intersect each other at least at $n+1$ distinct points namely $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$. In other words $P(x)$ interpolates $f(x)$ at $n+1$ points.

We can also think of finding a polynomial P such that certain derivatives of P coincide with the same order derivatives of the given function f at some points in the interval.

2.2. Interpolating Polynomial

Let f be a continuous function defined on some interval. We say that a polynomial P is an interpolating polynomial for f, if m^{th} derivative of P coincides with m^{th} derivative of f at certain points of the interval for some nonnegative integer m. The case $m = 0$ infer that P coincide with f at some points in the interval.

Example 2.1. The polynomial $P(x) = x - f(x_0)$ is an interpolating polynomial for some function $f(x)$, because $P(x)$ agrees with $f(x)$ at $x = x_0$.

Example 2.2. The Lagrange's polynomial is also an interpolating polynomial, because it matches with function at $n+1$ distinct points in the interval. And can give some approximating value for the function at any other point in the interval with some error.

Example 2.3. If $f(x) = \cos x$, then the polynomial $P(x) = 1$ not only agrees with $f(x)$ at $x = 0$, but also the first derivative of $P(x)$ agrees with the first derivatives of $f(x)$ at $x = 0$. Thus $P(x) = 1$ is an interpolating polynomial for $\cos x$.

But the third derivative of $f(x)$ does not agree with the third derivative of $P(x)$.

Example 2.4. If $f(x) = \cos x$, then the polynomial $P(x) = 1 - \frac{x^2}{2}$ not only agrees with $f(x)$ at $x = 0$, but also the first and second derivatives of $P(x)$ agree with the first and second derivatives of $f(x)$ respectively at $x = 0$. Thus $1 - \frac{x^2}{2}$ is also an interpolating polynomial for $\cos x$.

Example 2.5. If $f(x) = \sin x$, then the polynomial $P(x) = x$ not only agrees with $f(x)$ at $x = 0$, but also the first and second derivative of $P(x)$ agrees with the respective derivatives of $f(x)$ at $x = 0$. Thus $P(x) = x$ is an interpolating polynomial for $\sin x$.

But the third order derivative of $P(x) = x$ does not coincide with the third order derivative of the function $\sin x$! If we consider $P(x) = x - \frac{x^3}{3!}$, then $P(x)$ agrees with $\sin x$ at $x = 0$ along with its first three derivatives.

Does it remind you something? (Think about it.) Let us see one more example then your guess will be *Pakka*!

Example 2.6. If $f(x) = e^x$, then the polynomial

- $P_0(x) = 1$ is an interpolating polynomial for e^x, because $P_0(x) = 1$ agrees with the function at $x = 0$.
- $P_1(x) = 1 + x$ is an interpolating polynomial for e^x, because $P_1(x)$ and its first derivative agree with the function e^x at $x = 0$.
- $P_2(x) = 1 + x + \frac{x^2}{2}$ is an interpolating polynomial for e^x, because $P_2(x)$ and its first and second order derivatives agree with the function e^x at $x = 0$.
- $P_3(x) = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!}$ is an interpolating polynomial for e^x, because $P_3(x)$ and its first three derivatives agree with the function e^x and derivatives respectively at $x = 0$.

Example 2.7. The Taylor's polynomial of degree n of the function $f(x)$ around the point x_0 is given by

$$P_n(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{1}{2!}(x - x_0)^2 f''(x_0) + \ldots + \frac{1}{n!}(x - x_0)^n f^n(x_0). \tag{2.13}$$

Then P_n agrees with $f(x)$ at $x = x_0$ along with its first n derivatives. Therefore, P_n is also an interpolating polynomial for the function $f(x)$.

Remark 2.6. In all above examples, to find the interpolating polynomial $P(x)$, we do not need the full information about the function. We only need the the specific data about the function. In Example (2.5), we can also pose the problem in a different way, without saying any thing about *sine* function as follows.
(i) Find a polynomial $P(x)$, which satisfy the following $P(0) = 0$, $P'(0) = 1$, $P''(0) = 0$, and $P'''(0) = -1$, or
(ii) Find a polynomial $P(x)$, which agree with the function $f(x)$ along the data $f(0) = 0$, $f'(0) = 1$, $f''(0) = 0$, and $f'''(0) = -1$.

Note that this $f(x)$ need not be $\sin x$, for example it might be $x - \frac{x^3}{3!} + x^{10}$.

Example 2.8. The polynomial

$$P(x) = \frac{(x - x_1)}{(x_0 - x_1)} f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)} f(x_1)$$

is an interpolating polynomial for $f(x)$, because $P(x)$ coincide with $f(x)$ at the points x_0, x_1.

On the other hand if we want to find an interpolating polynomial for the data $f(x_0) = f(x_1) = c$, then above formula yields $P(x)$ to be a constant polynomial c. But the polynomial $P(x) = (x - x_1)(x - x_2) + c$ also satisfies the data. It means that the interpolating polynomial for a given data is not necessarily unique.

Question 2.2. Since the interpolating polynomial is not unique, the question arises that by which interpolating polynomial we should approximate the function.

In above example there are two interpolating polynomial for the given data, one is a constant polynomial and other is of degree two. Note that the calculations with the constant polynomial will be easier than that of two degree polynomial.

Remark 2.7. In general we need to deal with a large and numerically rich data, so to ease the computations it will be better to approximate by an interpolating polynomial of minimal degree.

Remark 2.8. In problem (2.1) we have interpolated a data with three distinct points by a two degree polynomial and by construction it is clear that this is the only polynomial of at most degree two, which can interpolate the given data. If we change the functional value as constant 1 at each of the above three points $(-1, 0, 1)$, and follow the same procedure as above, we get the interpolating polynomial as constant polynomial 1, which will be again unique by construction.

Question 2.3. If we are given the functional value at $(n+1)$ distinct points, then by the answer of Problem (2.2) there exists a Lagrange's polynomial of at most degree n, which interpolates the given data. Now the question is that wether there are any other interpolating polynomials (for the same data) of degree at most n. Or in other words, wether the Lagrange's polynomial is unique interpolating polynomial of degree at most n for given data of $n+1$ distinct points.

Answer to this is ***Yes***.

Theorem 2.1. *(**Uniqueness of Interpolating Polynomial**) If we know the functional value of a real valued function f at $n+1$ distinct points $x_0, x_1, \ldots, x_n$, then there exists exactly one polynomial of degree at most n, which interpolates f at $x_0, x_1, \ldots, x_n$.*

Clearly Lagrange's polynomial (2.10) is of degree at most n and interpolates the given data. Only thing we need to show is that it is unique interpolating polynomial of degree at most n. We will prove it in two ways.

Proof. First, without loss of generality we consider that $x_0 < x_1 < \ldots < x_{n-1} < x_n$. Let $P(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$ be a polynomial, which satisfies the following data.

$$P(x_0) = f(x_0), P(x_1) = f(x_1), \ldots, P(x_n) = f(x_n). \tag{2.14}$$

Then we must have

$$\begin{aligned}
a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n &= f(x_0) \\
a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n &= f(x_1) \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n &= f(x_n)
\end{aligned} \tag{2.15}$$

Or in matrix form,

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_n \end{bmatrix} = \begin{bmatrix} f(0) \\ f(x_1) \\ . \\ . \\ . \\ f(x_n) \end{bmatrix} \tag{2.16}$$

This a system of $(n+1)$ linear equations in $(n+1)$ unknowns $a_0, a_1, \ldots, a_n$. This system will have a unique solution if determinant of following matrix (known as Vandermonde determinant) is non zero.

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix} \neq 0 \tag{2.17}$$

But we know that $\Delta = \prod_{0 \le i < j \le n}(x_i - x_j) \neq 0$. So all the coefficients of $P(x)$ are uniquely determined and hence $P(x)$ itself. □

For the second proof we need the following lemma.

Lemma 2.2. *Let f be a real valued function on real line. Further if f is n times differentiable function and has $(n+1)$ zeros, then n_{th} derivative of the function f^n has at least one real zero.*

Above lemma can be proved by repeated application of Rolle's Theorem. We need the following corollary to this lemma for our second proof.

Corollary 2.3. *A non zero polynomial of degree n cannot have $n+1$ distinct zeros.*

Because if polynomial $P(x) \neq 0$ of degree n has $n+1$ zeros, then its by above lemma its n_{th} derivative must have at least one zero. But the n_{th} derivative of a polynomial of degree n is the constant function $n!a_n$, where a_n is the coefficient of x^n in $P(x)$ and this constant function can not have any zero unless a_n itself is zero. This gives the contradiction to the degree of the polynomial $P(x)$.

Proof. This proof is by contradiction, so if we consider any other polynomial $Q(x) \neq P(x)$ of degree at most n such that $Q(x)$ also interpolates the same data, that is, we must have

$$Q(x_0) = f(x_0), Q(x_1) = f(x_1), \ldots, Q(x_n) = f(x_n). \tag{2.18}$$

From (2.14) and (2.18), it is clear that if the polynomial $R(x) = P(x) - Q(x) \neq 0$ must have $n+1$ zeros at $x_0, x_1, \ldots, x_n$. This gives a contradiction because $R(x)$ is of at most degree n, and hence it cannot have $n+1$ distinct zeros unless $R(x)$ itself is a zero polynomial. □

Remark 2.9. The main difficulty in dealing with Lagrange's polynomial is that if we want to increase the data even with a single point only, we need to compute it from the very beginning.

2.3. Newton's Divided Difference Interpolating Polynomial

Newton suggested to write the interpolating polynomial in the following form for a given data of $n+1$ distinct points $x_0, x_1, \ldots, x_n$.

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0)(x - x_1)\ldots(x - x_{n-1}). \tag{2.19}$$

Since $P(x)$ coincide with $f(x)$ at the points $x_0, x_1, \ldots, x_n$, we must have

$$\begin{aligned}
&f(x_0) = a_0;\\
&f(x_1) = a_0 + (x_1 - x_0)a_1;\\
&f(x_2) = a_0 + (x_2 - x_0)a_1 + (x_2 - x_0)(x_2 - x_1)a_2;\\
&\ldots = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&\ldots = \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&f(x_n) = a_0 + (x_n - x_0)a_1 + (x_n - x_0)(x_n - x_1)a_2 + \ldots + (x_n - x_0)\ldots(x_n - x_{n-1})a_n.
\end{aligned} \tag{2.20}$$

It is clear from the above system of equations that $a_0 = f(x_0)$ and to compute a_1, we plug in the value of a_0 in the second equation and similarly to compute a_k we can substitute the values of $a_0, a_1, \ldots, a_{k-1}$ obtained by solving the first k equations in $(k+1)_{th}$ equation. It is important to note that we only need to consider first $k+1$ equations to compute a_k. Thus the computation of a_k depends only on the points $x_0, \ldots, x_k$. So we can rename a_k as some new function of the points $\{x_0, x_1, x_2, \ldots, x_k\}$. To show dependency of a_k on the function f also we denote a_k by $f[x_0, x_1, x_2, \ldots, x_k]$. We call $f[x_0, x_1, x_2, \ldots, x_k]$ as k_{th} divided difference of the function f relative to points $x_0, x_1, \ldots, x_k$. We will justify the name divided difference in further discussion. Now we can use the values of $a_0, \ldots, a_k$ in the $(k+2)_{th}$ equation to compute a_{k+1} and so on. Thus we can compute all the values of $a_0, a_1, \ldots, a_n$ and hence $P(x)$. Since this $P(x)$ is of degree at most n and also interpolates the data at $n+1$ distinct points say at $x_0, x_1, x_2, \ldots, x_n$, then by Theorem (of uniqueness of interpolating polynomial) 2.1, we infer that $P(x)$ is nothing but Lagrange's polynomial written in different form. But then $a_n = f[x_0, x_1, x_2, \ldots, x_n]$, which is the

coefficient x^n in Newton's polynomial, must also be the coefficient of x^n in Lagrange's polynomial as well. And hence, equating the coefficient of x^n from (2.19) and (2.10), we get

$$a_n = f[x_0, x_1, x_2, \ldots, x_n] = \sum_{j=0}^{n} \frac{f(x_j)}{\prod_{i=0, i\neq j}^{n}(x_j - x_i)}. \tag{2.21}$$

Now it is worth to note that even if we rearrange the order of data points, the interpolating polynomial will not change and hence a_n, the coefficient of x^n in interpolating polynomial, will remain the same. This fact is also justified by the expression on R.H.S. of (2.21). Thus we can say that the n_{th} divided difference $f[x_0, x_1, x_2, \ldots, x_n]$ is dependent on the set of node points $\{x_0, x_1, x_2, \ldots, x_n\}$ but independent of the order of node points.

In the following we aim to justify the name divided difference. It is easy to see from (2.20) that

$$a_1 = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}.$$

Here, a_1 is obtained by dividing the difference of functional values by the difference of points and hence it is known as Newton's first divided difference of $f(x)$ relative to x_0, x_1 and denoted by $f[x_0, x_1]$. In general Newton's first divided difference of $f(x)$ relative to x_j, x_k, $(j \neq k)$ is defined as

$$f[x_j, x_k] = \frac{f(x_j) - f(x_k)}{(x_j - x_k)}. \tag{2.22}$$

It is clear from the definition that $f[x_j, x_k]$ is independent of the order of j and k and depends only on the set $\{j, k\}$. Using these expressions for a_0 and a_1, we can compute a_2 from first three equations of (2.20).

$$a_2 = f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}. \tag{2.23}$$

Clearly, $f[x_0, x_1, x_2]$ is independent of order of x_0, x_1, x_2. In fact one can write $f[x_0, x_1, x_2]$ as the fraction of the difference of certain first divided differences by difference of certain points as following.

Exercise 2.1. Show that

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{(x_2 - x_0)} = \frac{f[x_1, x_2] - f[x_2, x_0]}{(x_1 - x_0)} = \frac{f[x_0, x_2] - f[x_0, x_1]}{(x_2 - x_1)}. \tag{2.24}$$

Hence $f[x_0, x_1, x_2]$ is known as Newton's second divided difference of $f(x)$ relative to $\{x_0, x_1, x_2\}$. In general we define Newton's second divided difference of $f(x)$ relative to $\{x_i, x_j, x_k\}$ as

$$f[x_i, x_j, x_k] = \frac{f[x_j, x_k] - f[x_i, x_j]}{(x_k - x_i)}. \tag{2.25}$$

Moreover we can proceed in the similar way and can see that $f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k]$, the Newton's k_{th} divided difference of $f(x)$ relative to $\{x_0, x_1, \ldots, x_{k-1}, x_k\}$, can be written as the fraction of difference of certain $(k-1)_{th}$ divided difference by the difference of certain point values as follows.

$$f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k] = \frac{f[x_1, x_2, \ldots, x_{k-1}, x_k] - f[x_0, x_1, x_2, \ldots, x_{k-1}]}{(x_k - x_0)}. \tag{2.26}$$

It can be checked that

$$f[x_0, x_1, x_2, \ldots, x_{k-1}, x_k] = \sum_{i=0}^{k} \frac{f(x_i)}{\prod_{j\neq i=0}^{k}(x_i - x_j)}, \quad k = 3, \ldots, n. \tag{2.27}$$

In general we define Newton's k_{th} divided difference of $f(x)$ relative to $\{x_{i_0}, x_{i_1}, \ldots, x_{i_{k-1}}, x_{i_k}\}$ recursively as follows

$$f[x_{i_0}, x_{i_1}, x_{i_2}, \ldots, x_{i_{k-1}}, x_{i_k}] = \frac{f[x_{i_1}, x_{i_2}, \ldots, x_{i_{k-1}}, x_{i_k}] - f[x_{i_0}, x_{i_1}, x_{i_2}, \ldots, x_{i_{k-1}}]}{(x_{i_k} - x_{i_0})}. \tag{2.28}$$

Now we can rewrite the Newton's interpolating polynomial by plugging the the expressions for a_k's as Newton's divided difference in (2.19).

$$\begin{aligned} P(x) &= f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots \\ &\quad \ldots + f[x_0, x_1, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-1}), \\ &= f(x_0) + \sum_{i=1}^{n}(x - x_0)(x - x_1)\ldots(x - x_{i-1})f[x_0, x_1, \ldots, x_i] \end{aligned} \tag{2.29}$$

Thus to find the polynomial $P(x)$ completely, we need to find $f[x_0, x_1]$, $f[x_0, x_1, x_2]$, ..., $f[x_0, x_1, \ldots, x_n]$, that is, the divided difference of f of all order from 1 to n. One can directly calculate $f[x_0, x_1]$, but to find $f[x_0, x_1, x_2]$ we also need to calculate first difference of f relative to points x_1, x_2, that is, $f[x_1, x_2]$. And further to find $f[x_0, x_1, x_2, x_3]$, we need to find $f[x_1, x_2, x_3]$, that is, we need an extra computation of $f[x_2, x_3]$. If we proceed in this way, we need to compute the following table

$$\begin{bmatrix} x & f & 1_{st}f[x_i, x_{i+1}] & 2_{nd}f[x_i, x_{i+1}, x_{i+2}] & \ldots & n_{th}f[x_0, \ldots, x_n] \\ x_0 & f(x_0) & f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{(x_1-x_0)} & f[x_0, x_1, x_2] = \frac{f[x_1,x_2]-f[x_0,x_1]}{(x_2-x_0)} & \ldots & f[x_0, \ldots, x_n] \\ x_1 & f(x_1) & f[x_1, x_2] = \frac{f(x_2)-f(x_1)}{(x_2-x_1)} & f[x_1, x_2, x_3] = \frac{f[x_2,x_3]-f[x_1,x_2]}{(x_3-x_1)} & \ldots & \ldots \\ x_2 & f(x_2) & f[x_2, x_3] = \frac{f(x_3)-f(x_2)}{(x_3-x_2)} & f[x_2, x_3, x_4] = \frac{f[x_3,x_4]-f[x_2,x_3]}{(x_4-x_2)} & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ x_{n-1} & f(x_{n-1}) & f[x_{n-1}, x_n] & \ldots & \ldots & \ldots \\ x_n & f(x_n) & \ldots & \ldots & \ldots & \ldots \end{bmatrix}.$$

This table will be of upper triangular form and the coefficients $a_0 = f(x_0)$, $a_= f[x_0, x_1]$, $a_2 = f[x_0, x_1, x_2], \ldots, a_n = f[x_0, x_1, \ldots, x_n]$ are the entries of the first row. So by directly substituting the values of the coefficients in (2.19), we can obtain Newton's divided difference interpolating polynomial.

Problem 2.3. Find Newton's interpolating polynomial for the following data.

$$f(0) = 3, f(1) = 5, f(3) = 33, f(4) = 67, f(5) = 133.$$

First we form the divided difference table for the above data.

$$\begin{array}{cccccc} x & f & 1_{st} & 2_{nd} & 3_{rd} & 4_{th} \\ 0 & 3 & \frac{3-3}{(1-0)} = 0 & \frac{12-0}{(3-0)} = 4 & \frac{8-4}{(4-0)} = 1 & \frac{1-1}{(5-0)} = 0 \\ 1 & 3 & \frac{27-3}{(3-1)} = 12 & \frac{36-12}{(4-1)} = 8 & \frac{12-8}{(5-1)} = 1 & \ldots \\ 3 & 27 & \frac{63-27}{(4-3)} = 36 & \frac{60-36}{(5-3)} = 12 & \ldots & \ldots \\ 4 & 63 & \frac{123-63}{(5-4)} = 60 & \ldots & \ldots & \ldots \\ 5 & 123 & \ldots & \ldots & \ldots & \ldots \end{array}. \tag{2.30}$$

Now we substitute the corresponding values divided differences in formula for Newton's interpolating polynomial of degree four.

$$\begin{aligned} &P(x) = f(0) + f[0,1](x-0) + f[0,1,3](x-0)(x-1) + \\ &f[0,1,3,4](x-0)(x-1)(x-3) + f[0,1,3,4,5](x-0)(x-1)(x-3)(x-4) \\ &P(x) = 3 + 0(x-0) + 4(x-0)(x-1) + 1(x-0)(x-1)(x-3) + 0(x-0)(x-1)(x-3)(x-4) \\ &P(x) = 3 + 4x(x-1) + x(x-1)(x-3) \end{aligned}$$

2.4. Divided difference at variable point (Functional value in terms of divided difference.)

In the following, we will try to find the divided difference of the function at some arbitrary point of the interval. And we will see that the functional value at some given point can be expressed in terms of divided difference at that point.

$$f[x, x_0] = \frac{f(x_0) - f(x)}{(x_0 - x)}.$$
$$\implies \quad f(x) = f(x_0) + (x - x_0) f[x, x_0]. \tag{2.31}$$

Further,

$$f[x, x_0, x_1] = \frac{f[x_0, x_1] - f[x, x_0]}{(x_1 - x)}.$$
$$\implies \quad f[x, x_0] = f[x_0, x_1] + (x - x_1) f[x, x_0, x_1]. \tag{2.32}$$

Substituting (2.32) in (2.31), we get

$$f(x) = f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x, x_0, x_1]. \tag{2.33}$$

Continuing this procedure up to n^{th} divided difference, we get inductively that

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] + \dots \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x, x_0, x_1, \dots, x_{n-1}]. \end{aligned} \tag{2.34}$$

But

$$f[x, x_0, x_1, \dots, x_n] = \frac{f[x_0, x_1, x_2, \dots, x_n] - f[x, x_0, x_1, \dots, x_{n-1}]}{(x_n - x)}$$
$$\implies \quad f[x, x_0, x_1, \dots, x_{n-1}] = f[x_0, x_1, x_2, \dots, x_n] + (x - x_n) f[x, x_0, x_1, \dots, x_n].$$

Thus from (2.34), we get

$$\begin{aligned} f(x) &= f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2] \\ &\quad + \dots + (x - x_0)(x - x_1) \dots (x - x_{n-1}) f[x_0, x_1, x_2, \dots, x_n] \\ &\quad + (x - x_0)(x - x_1) \dots (x - x_n) f[x, x_0, x_1, \dots, x_n]. \end{aligned} \tag{2.35}$$

Using the expression for Newton's polynomial form (2.29), we get from (2.35)

$$f(x) = P(x) + (x - x_0)(x - x_1) \dots (x - x_n) f[x, x_0, x_1, \dots, x_n]. \tag{2.36}$$

2.5. Properties of divided difference

As we saw that first divided difference relative to points x_0, x_1 is defined as $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{(x_1 - x_0)}$. But this expression reminds us (***mean value theorem***) that if we also assume the differentiability of the function, our first divided difference $f[x_0, x_1]$ is nothing but the derivative of the function f at some point in the smallest open interval containing both x_0, x_1.

Question 2.4. Can we say that second divided difference is related to second derivative of the function?
If yes, then how?

If we assume that the function is two times differentiable and $x_0 < x_1 < x_2$, then we write

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{(x_2 - x_0)} = \frac{f'(\xi_1) - f'(\xi_2)}{(x_2 - x_0)} = \frac{(\xi_1 - \xi_0)}{(x_2 - x_0)} \frac{f'(\xi_1) - f'(\xi_2)}{(\xi_1 - \xi_0)} = \frac{(\xi_1 - \xi_0)}{(x_2 - x_0)} f''(\zeta), \tag{2.37}$$

where $\xi_0 \in (x_0, x_1), \xi_1 \in (x_1, x_2)$ and $\zeta \in (\xi_0, \xi_1)$.

Observation 2.2. Here we note that if ξ_0, ξ_1 are the middle points of the intervals (x_0, x_1) and (x_1, x_2) respectively, then $f[x_0, x_1, x_2] = \frac{1}{2}f''(\zeta)$.

We will return back to this and prove a relation between divided difference and derivative of the function in the Remark 2.12 of Theorem 2.4. Now we see some other properties of divided differences.

Using Exercise 2.1, one can see that Newton's second divided difference is independent of the order of node points and depends only on the set of all node points. If we consider the function $\omega(x) = \prod_{j=0}^{n}(x - x_j)$, then we can see that

$$f[x_0, x_1, x_2, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\omega'(x_i)}. \tag{2.38}$$

We know that $f[x, x_0] = \frac{f(x_0)-f(x)}{(x_0-x)}$. This shows that $f[x, x_0]$ is a continuous function of x for all $(x_0 \neq)x \in [a, b]$, possibly undefined at x_0. But if we consider f to be differentiable function at x_0, then we can define

$$f[x_0, x_0] = \lim_{x \to x_0} f[x, x_0] = \lim_{x \to x_0} \frac{f(x_0) - f(x)}{(x_0 - x)} = \frac{d}{dx} f|_{x=x_0} = f'(x_0). \tag{2.39}$$

Further, if we assume that f is once differentiable in whole interval (a, b), then it is clear by mean value theorem that $f[x_0, x] = \frac{f(x)-f(x_0)}{(x-x_0)} = f'(\tilde{x})$ for some $\tilde{x} \in (\min\{x_0, x\}, \max\{x_0, x\})$. Moreover,

$$\begin{aligned}
\frac{d}{dx} f[x_0, x_1, \ldots, x_n, x] &= \lim_{h \to 0} \frac{f[x_0, x_1, \ldots, x_n, x+h] - f[x_0, x_1, \ldots, x_n, x]}{h}, \\
&= \lim_{h \to 0} \frac{f[x_0, x_1, \ldots, x_n, x+h] - f[x, x_0, x_1, \ldots, x_n]}{(x+h) - x}, \\
&= \lim_{h \to 0} f[x, x_0, x_1, \ldots, x_n, x+h], \\
&= f[x, x_0, x_1, \ldots, x_n, x], \\
&= f[x_0, x_1, \ldots, x_n, x, x].
\end{aligned} \tag{2.40}$$

Problem 2.4. Use induction to prove that

$$\frac{d^k}{dx^k} f[x_0, x_1, \ldots, x_n, x] = k! f[x_0, x_1, \ldots, x_n, \overbrace{x, x, \ldots, x}^{(k+1) \text{ times}}]. \tag{2.41}$$

Now we further assume f to be second differentiable function and define

$$\begin{aligned}
f[x_0, x_0, x_0] &= \lim_{x_1 \to x_0, x_2 \to x_0} f[x_0, x_1, x_2], \\
&= \lim_{x_1 - x_0 \to 0, x_2 - x_0 \to 0} f[x_0, x_1, x_2], \\
&= \lim_{h \to 0, k \to 0} f[x_0, x_0 + h, x_0 + k], \\
&= \lim_{h \to 0, k \to 0} \frac{f[x_0 + h, x_0 + k] - f[x_0, x_0 + h]}{(x_0 + k) - x_0}, \\
&= \lim_{k \to 0} \frac{f[x_0, x_0 + k] - f[x_0, x_0]}{k}, \\
&= \lim_{k \to 0} \frac{\frac{f(x_0+k)-f(x_0)}{k} - f'(x_0)}{k}, \\
&= \lim_{k \to 0} \frac{f(x_0 + k) - f(x_0) - k f'(x_0)}{k^2}.
\end{aligned}$$

Using L'Hospital's Rule, we have now

$$\begin{aligned} f[x_0, x_0, x_0] &= \lim_{k\to 0} \frac{f(x_0+k) - f(x_0) - kf'(x_0)}{k^2} \\ &= \lim_{k\to 0} \frac{f'(x_0+k) - f'(x_0)}{2k} \\ &= \frac{1}{2} f''(x_0). \end{aligned} \tag{2.42}$$

Remark 2.10. Let f be n times differentiable, then in the similar manner one can define $f[x_0, x_0, \ldots, x_0] = \lim_{h_i \to 0} f[x_0, x_0 + h_1, \ldots, x_0 + h_n]$ and using induction one can obtain that

$$f[\overbrace{x_0, x_0, \ldots, x_0}^{(n+1) \text{ times}}] = \frac{1}{n!} f^n(x_0). \tag{2.43}$$

2.6. Error in Interpolation

Till now we approximated a continuous real valued function f defined on interval $[a, b]$ by the interpolating polynomials P_n at node points $x_0 = a, x_1, \ldots, x_n = b$. But interpolating polynomials P_n does not necessarily match with f at some arbitrary point of the interval other than node points. And we expect some difference between $P_n(\tilde{x})$ and $f(\tilde{x})$ at some point $\tilde{x} \in [a, b], \tilde{x} \neq x_i$. This difference is known as error in interpolation. Since by Theorem 2.1 there is unique interpolation polynomial P_n of degree at most n, we can define error function as

$$E_n f(x) = f(x) - P_n(x). \tag{2.44}$$

Remark 2.11. We note that Newton's divided difference interpolating polynomial $P_n(x)$ given by (2.29) interpolates the function $f(x)$ at $n+1$ distinct points and the degree of $P_n(x)$ is at most n, but Lagrange's polynomial $L(x)$ given by (2.11) also interpolates the same data so by Theorem 2.1 both $L(x)$ and $P_n(x)$ are the same polynomial written in two different ways. Therefore, for a given continuous function f the error function defined by (2.44) depends only on the set of node points. From expression (2.36) it is clear that by interpolating the function $f(x)$ by a Newton's divided difference interpolating polynomial $P_n(x)$, the error $E_n f(x)$ can be expressed a

$$E_n f(x) = f(x) - P_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n) f[x, x_0, x_1, \ldots, x_n]. \tag{2.45}$$

Question 2.5. Can we estimate this error?

Theorem 2.4. *Let f be $n+1$ times differentiable function defined on interval $[a, b]$. Let P_n be the unique interpolating polynomial of degree at most n, interpolating f at $n+1$ distinct points $x_0, x_1, x_2, \ldots, x_n$. Then the error in interpolation at some point $x_i \neq x \in [a, b]$ is given by*

$$E_n f(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \frac{f^{n+1}(\xi)}{(n+1)!}. \tag{2.46}$$

Where ξ is some point in the interval $(\min\{x, x_0, x_1 \ldots, x_n\}, \max\{x, x_0, x_1 \ldots, x_n\})$ and depends upon x.

Proof. Let x be some point in the interval other than node points. Let $\phi(t)$ be a function defined on interval $[a, b]$ by

$$\phi(t) = f(t) - P_n(t) - (t - x_0)(t - x_1) \ldots (t - x_n) K, \tag{2.47}$$

where K is a constant determined by the equation $\phi(x) = 0$, that is,

$$f(x) - P_n(x) + (x - x_0)(x - x_1) \ldots (x - x_n) K = 0. \tag{2.48}$$

It is clear that $\phi(x_i) = 0$ for all $i = 0, 1, 2, \ldots, n$. Thus function ϕ has $n+2$ zeros say $\{x, x_0, x_1, \ldots, x_n\}$. Hence its $(n+1)_{th}$ derivative has at least one zero at some point ξ in the interval $(\min\{x, x_0, x_1 \ldots, x_n\}, \max\{x, x_0, x_1 \ldots, x_n\})$, that is, $\phi^{n+1}(\xi) = 0$. Therefore,

$$0 = \phi^{n+1}(\xi) = f^{n+1}(\xi) - 0 - (n+1)!K. \tag{2.49}$$

This implies that $K = \frac{1}{n+1!} f^{n+1}(\xi)$. And hence from (2.47) and (2.48),

$$E_n f(x) = f(x) - P_n(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \frac{1}{(n+1)!} f^{n+1}(\xi). \tag{2.50}$$

□

Remark 2.12. Equating (2.45) and (2.46), we get

$$f[x, x_0, x_1, \ldots, x_n] = \frac{1}{n+1!} f^{n+1}(\xi). \tag{2.51}$$

Where $\xi \in (\min\{x, x_0, x_1 \ldots, x_n\}, \max\{x, x_0, x_1 \ldots, x_n\})$.

Corollary 2.5. *If $P_m(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_m x^m$ is of degree m, then m_{th} divided difference of P_m is a_m and $(m+1)_{th}$ divided difference is zero.*

Proof. For $f(x) = P_m(x)$, we have from (2.51) that $P_m[x, x_0, x_1, \ldots, x_n] = \frac{1}{n+1!} P_m^{n+1}(\xi)$. Thus $P_m[x, x_0, x_1, \ldots, x_m] = \frac{1}{m+1!} P_m^{m+1}(\xi) = 0$, or $P_m[x_0, x_1, \ldots, x_m, x_{m+1}] = 0$. Similarly $P_m[x, x_0, x_1, \ldots, x_{m-1}] = \frac{1}{m!} P_m^m(\xi) = a_m$, or $P_m[x_0, x_1, \ldots, x_{m-1}, x_m] = a_m$. □

2.7. Bound on the Error in Interpolation

We note from (2.46) that the error in interpolation by n degree polynomial is given by

$$E_n f(x) = (x - x_0)(x - x_1) \ldots (x - x_n) \frac{f^{n+1}(\xi)}{(n+1)!} = \frac{f^{n+1}(\xi)}{(n+1)!} \omega(x). \tag{2.52}$$

If we denote by

$$M_{n+1} = \max_{x \in [a,b]} |f^{n+1}(x)|, \tag{2.53}$$

then the error is bounded by

$$|E_n f(x)| \leq \frac{M_{n+1}}{(n+1)!} |(x - x_0)(x - x_1) \ldots (x - x_n)|. \tag{2.54}$$

If we some how know the bound for $|(x - x_0)(x - x_1) \ldots (x - x_n)|$ on $[a, b]$ or the $\max_{x \in [a,b]} |(x - x_0)(x - x_1) \ldots (x - x_n)|$, we can find the uniform bound for the error function. Certainly, the maxima and minima is attained by the continuous function $\omega(x) = (x - x_0)(x - x_1) \ldots (x - x_n)$ on some points of the closed interval $[a, b]$. And we can bound $E_n f(x)$ by $\frac{M_{n+1}}{(n+1)!} \max\{|minima|, |maxima|\}$. Since points of maxima or minima are also the points of local maxima and minima, so if we know all the points of local minima and maxima, that is, the critical points, we only need to compute the modulus of functional values at these critical points and find the maximum of those values. But we know that the critical points for some differentiable function are given by the roots of the derivative of the function. In our case critical points are given by the roots of the derivative of the polynomial $\omega(x)$, that is, by the solution of the equation $\omega'(x) = 0$. But to find these roots might not be an easy task if there is a large number of node points.

Problem 2.5. Find the bound on the error by interpolating a polynomial of five degree $x^5 - x^3 + 1$ by Lagrange's interpolating polynomial of degree two at the points $0, 2, 3$.

Solution. We will find the solution in following steps.

- Since the number of node points are two, we need to interpolate by a polynomial of degree three.
- Since we are looking to interpolate the function with the given node points, so the interval of our concern is the smallest interval containing all the nodes. In our case this is $[0, 3]$.
- Since the error function in our case is given by $E(x) = (x-0)(x-2)(x-3)\frac{1}{(2+1)!}f^{2+1}(\xi)$, to find the bound on this it is enough to find the bound on $(x-0)(x-2)(x-3)$ and $f'''(x)$ on the interval $[0, 3]$.
- We have $f'''(x) = 60x^2 - 6$. Since the derivative of $60x^2 - 6$ is non-negative on the interval $[0, 3]$, f''' is non increasing function. Thus the maximum of $|f'''(x)|$ on the interval is possible at the end points of interval. So clearly this maximum is 534.
- Further we also need to compute the maximum of $|x^3 - 5x^2 + 6x|$ on $[0, 3]$. This can be attained either at the end point of the interval or at some local maxima or minima. But at end points, $\omega(x) = x^3 - 5x^2 + 6x$ is zero, so we only need to consider the critical points of $\omega(x)$ in the interval $[0, 3]$. These critical points can be obtained by solving $3x^2 - 10x + 6 = 0$. So points of extremum are $x_e = (5 \pm \sqrt{7})/3 = 2.54, .78$. And the local extremum values will be given by $\omega(x_e) = -.63, .21$. thus the maximum value of $|x^3 - 5x^2 + 6x|$ on $[0, 3]$ is .63.
- Thus,

$$\begin{aligned} |E(x)| &= |(x-0)(x-2)(x-3)\frac{1}{(2+1)!}f^{2+1}(\xi)| \\ &\leq |(x-0)(x-2)(x-3)|\ \frac{1}{(3)!}\ |f^3(\xi)| \\ &\leq .63 \times .16 \times 534 = 53.83 \end{aligned}$$

Remark 2.13. Thus we see in above problem that error in interpolation might be significantly large. But suppose if we have freedom to choose the position of node points, (not the number of nodes), wether we can control the error.

Question 2.6. More precisely, suppose we are given a smooth function f. And we aim to interpolate f by $n+1$ node points. Further suppose that the choice of node points is in our hand. Can we choose these $n+1$ nodes in such a way that the maximum error on the smallest interval containing all the nodes is controlled?

Suppose we assume that the node points are equally spaced, that is, $a = x_0, < x_1 = x_0 + h, < x_2 = x_0 + 2h, \ldots < x_n = x_0 + nh = b$. And we know some how that $|f^{n+1}(x)| \leq M_{n+1}$. Then it is clear from from (2.54) that error function is bounded as follows.

$$|E_n f(x)| \leq \frac{M_{n+1}}{(n+1)!}|(x-x_0)(x-x_0-h)\ldots(x-x_0-nh)|. \tag{2.55}$$

Clearly $\omega(x) = (x-x_0)(x-x_0-h)\ldots(x-x_0-nh)$ is independent of the function. Here we will try to find a bound for $\omega(x) = (x-x_0)(x-x_0-h)\ldots(x-x_0-nh)$ on the smallest interval containing all nodes, that is, $[x_0, x_0+nh]$.

We will find the bound for $\omega(x)$ in three different cases of linear ($n = 1$), quadratic ($n = 2$), and cubic ($n = 3$) interpolation.

In case of linear interpolation it is easy to observe that the maximum value of $|(x-x_0)(x-x_1)|$ is obtained at $x = (x_0 + x_1)/2$ and is given by $(x_1 - x_0)^2/4$. And thus the bound for error in interpolation by a linear polynomial (line) is given by $M_2(x_1 - x_0)^2/8$, that is,

$$|E_1 f(x)| \leq M_2(x_1 - x_0)^2/8. \tag{2.56}$$

In case of quadratic interpolation we need to find the maximum value of $|\omega(x)| = |(x-x_0)(x-x_1)(x-x_2)|$. This maximum value can be obtained at one of the roots of quadratic polynomial $|\omega'(x)|$, which certainly has two real roots. But in case of equidistant points we need to find the maximum value of $|\omega(x)| = |(x-x_0)(x-x_0-h)(x-x_0-2h)|$ in the interval $[x_0, x_0+2h]$. This extremum will be attained at some critical point given by $\omega'(x) = 0$ or $(x-x_0-h)(x-x_0-2h) + (x-x_0)(x-x_0-2h) + (x-x_0)(x-x_0-h) = 0$. For simplification we reparameterize the curve $\omega(x)$ with origin as the middle point of the interval $[x_0, x_0+2h]$ by assuming $x - x_0 = (t+1)h$, then $x \in [x_0, x_0+2h] \Leftrightarrow t \in [-1, 1]$ and we need to solve $t(t-1) + (t+1)(t-1) + (t+1)t = 0$, that is, $3t^2 - 1 = 0$. Thus $t = \pm 1/\sqrt{3}$ or $x = x_0 + h \pm h/\sqrt{3}$. But at both of these values of x, $|\omega(x)| = 2h^3/\sqrt{27}$. Thus from (2.55), It is clear that

$$|E_2 f(x)| \leq \frac{M_3}{3!} 2h^3/\sqrt{27}. \tag{2.57}$$

Now we can bound $E_2 f$ by any given positive small number by choosing step size h accordingly small.

Similarly in cubic case if we know M_4, then to bound $E_3 f$, we need to control $|\omega(x)| = |(x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)|$. Thus we need to find the solution for $\omega'(x) = 0$. For simplification we reparameterize the curve $\omega(x)$ with origin as the center of the interval $[x_0, x_0+3h]$. Let $x - x_0 = (t+3/2)h$, then $x \in [x_0, x_0+3h] \Leftrightarrow t \in [-3/2, 3/2]$ and we need to solve $d/dt[(t^2 - 9/4)(t^2 - 1/4)] = 0$, which gives $t = 0, \pm\sqrt{5}/2$. And the maximum value of $\omega(x)$ is 1. And hence

$$|E_3 f(x)| \leq \frac{M_4}{4!} h^4. \tag{2.58}$$

Problem 2.6. Determine the maximum step size that can be used in the tabulation of $f(x) = e^x$ in $[0, 1]$, so that the error for cubic interpolation will be less than 5×10^{-4}.

Solution. Since we want to approximate f by a three degree polynomial, we need to bound the error function $E_3 f$ using (2.58). For this we need to find M_4, that is, maximum of the fourth derivative of $f(x) = e^x$ on the interval $[0, 1]$. But $f^{4th}(x) = e^x$, so $M_4 = e$. Since we want or error to be bounded by 5×10^{-4}, we need the step size h such that

$$|E_3 f(x)| \leq \frac{M_4}{4!} h^4 = (e/24)h^4 \leq 5 \times 10^{-4}.$$

Or we want

$$h^4 \leq (5 \times 24 \times 10^{-4})/e = .004414553294,$$

or,

$$h \leq (.004414553294)^{.25} = .25776366.$$

Observation 2.3. Suppose if we approximate the function $f(x) = x^3 - x^2 + 2$ at the nodes $-1, 0, 1$, the two degree polynomial, which satisfies $f(-1) = 2, f(0) = 2, f(1) = 2$, is given by $P(x) = -x^2 + x + 2$. If we draw the graphs of these functions, we observe that the graphs intersect each other at $x = 1$, but goes away rapidly as soon as they leave $x = 1$. And this is because they intersect perpendicularly at $x = 1 (P'(1) = -1, f'(1) = 1)$. So it is desirable to approximate the function with a polynomial, which not only agrees with the function at nodes but derivative of the polynomial also agrees with the derivative of the function at node points.

2.8. Hermite Interpolation

Problem 2.7. Find a polynomial $P(x)$ of minimal degree which agrees with the function f along the data

$$P(x_0) = f(x_0), P'(x_0) = f'(x_0), P(x_1) = f(x_1), P'(x_1) = f'(x_1). \tag{2.59}$$

Solution. Since there are four conditions, it is reasonable to consider the three degree polynomial $P(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, in four unknown coefficients, (a_0, a_1, a_2, a_3), which are to be determined by the given data (2.59). Since $P(x)$ satisfies (2.59), we have

$$\begin{array}{llll} a_0(1)+ & a_1x_0+ & a_2x_0^2+ & a_3x_0^3 \quad = f(x_0) \\ a_0(0)+ & a_1(1)+ & a_2 2x_0+ & a_3 3x_0^2 \quad = f'(x_0) \\ a_0(1)+ & a_1x_1+ & a_2x_1^2+ & a_3x_1^3 \quad = f(x_1) \\ a_0(0)+ & a_1(1)+ & a_2 2x_1+ & a_3 3x_1^2 \quad = f'(x_1) \end{array}$$

Or in matrix form,

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 0 & 1 & 2x_0 & 3x_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 0 & 1 & 2x_1 & 3x_1^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f'(x_0) \\ f(x_1) \\ f'(x_1) \end{bmatrix} \tag{2.60}$$

One can check that the $\det \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 0 & 1 & 2x_0 & 3x_0^2 \\ 1 & x_1 & x_1^2 & x_1^3 \\ 0 & 1 & 2x_1 & 3x_1^2 \end{bmatrix} = (x_1 - x_0)^4 \neq 0$, so we will have a unique solution to the above system of four equations. But to find solution is a tedious task. (So we leave it to very enthusiastic reader.)

Question 2.7. Can we find above polynomial in terms of Lagrange's fundamental polynomials? Can we find the values of $a_i, b_i, c_i, d_i, i = 0, 1, 2, \ldots, n$, so that the polynomial

$$P(x) = \sum_{i=0}^{n} \left[(a_ix + b_i)L_i^2(x)f(x_i) + (c_ix + d_i)L_i^2(x)f'(x_i) \right], \tag{2.61}$$

satisfies $P(x_i) = f(x_i)$ and $P'(x_i) = f'(x_i)$ for all $i = 0, 1, 2, \ldots, n$.

Problem 2.8. Suppose we denote

$$H_i(x) = (a_ix + b_i)L_i^2(x), \tag{2.62}$$

and

$$K_i(x) = (c_ix + d_i)L_i^2(x), \tag{2.63}$$

then find these a_i, b_i, c_i, d_i such that

$$H_i(x_j) = \delta_{ij}, K_i(x_j) = 0, H_i'(x_j) = 0, K_i'(x_j) = \delta_{ij}, \tag{2.64}$$

so that the polynomial

$$P(x) = \sum_{i=0}^{n} \left[H_i(x)f(x_i) + K_i(x)f'(x_i) \right], \tag{2.65}$$

automatically satisfies the conditions satisfies $P(x_i) = f(x_i)$ and $P'(x_i) = f'(x_i)$ for all $i = 0, 1, 2, \ldots, n$.

Solution. We use the definition of H_i, K_i in (2.64) to obtain

$$\begin{aligned} (a_ix_i + b_i)L_i^2(x_i) &= 1 && (2.66) \\ (a_ix_j + b_i)L_i^2(x_j) &= 0 && (2.67) \\ a_iL_i^2(x_i) + (a_ix_i + b_i)2L_i(x_i)L_i'(x_i) &= 0 && (2.68) \\ a_iL_i^2(x_j) + (a_ix_j + b_i)2L_i(x_j)L_i'(x_j) &= 0 && (2.69) \\ (c_ix_i + d_i)L_i^2(x_i) &= 0 && (2.70) \\ (c_ix_j + d_i)L_i^2(x_j) &= 0 && (2.71) \\ c_iL_i^2(x_i) + (c_ix_i + d_i)2L_i(x_i)L_i'(x_i) &= 1 && (2.72) \\ c_iL_i^2(x_j) + (c_ix_j + d_i)2L_i(x_j)L_i'(x_j) &= 0 && (2.73) \end{aligned}$$

Clearly (2.67),(2.69),(2.71),(2.73) are obviously satisfied because $L_i(x_j) = 0$. And to satisfy other conditions, we need

$$\begin{aligned}(a_i x_i + b_i) &= 1 && (2.74)\\ a_i + (a_i x_i + b_i)2L_i'(x_i) &= 0 && (2.75)\\ (c_i x_i + d_i) &= 0 && (2.76)\\ c_i + (c_i x_i + d_i)2L_i'(x_i) &= 1 && (2.77)\end{aligned}$$

After solving them we get,

$$\begin{aligned}a_i &= -2L_i'(x_i) && (2.78)\\ b_i &= 1 + 2x_i L_i'(x_i) && (2.79)\\ c_i &= 1 && (2.80)\\ d_i &= -x_i && (2.81)\end{aligned}$$

Thus $P(x)$ is given by

$$P(x) = \sum_{i=0}^{n} \left[1 - 2(x - x_i)L_i'(x_i)\right] L_i^2(x) f(x_i) + \sum_{i=0}^{n} (x - x_i) L_i^2(x) f'(x_i). \tag{2.82}$$

This polynomial is called as Hermite polynomial. Here we can write $L_i'(x_i)$ in terms of $\omega(x)$ as follows

$$L_i'(x_i) = \frac{\omega''(x_i)}{2\omega'(x_i)}. \tag{2.83}$$

Remark 2.14. In Hermite interpolation there are $n + 1$ nodes, and interpolating polynomial should satisfy $2(n + 1)$ equations. This suggests that our $P(x)$ should be of degree at most $2n + 1$. And clearly from (2.82), $P(x)$ satisfies this minimal degree criteria.

Theorem 2.6. *(**Uniqueness of Hermite Interpolating Polynomial**) If we know the value of a real valued function f and its derivative f' at $n + 1$ distinct points $x_0 < x_1 < \ldots <, x_n$, then there exists exactly one polynomial of degree at most $2n+1$, which satisfies the data $P(x_i) = f(x_i)$ and $P'(x_i) = f'(x_i)$ for all $i = 0, 1, 2, \ldots, n$.*

Proof. Clearly, the existence of such a polynomial is given by (2.82). We only need to prove the uniqueness.

Suppose if there is some polynomial $Q(x) \neq P(x)$ of degree at most $2n + 1$ such that

$$P(x_i) = f(x_i) = Q(x_i), P'(x_i) = f'(x_i) = Q(x_i).$$

Then $\phi(x) = P(x) - Q(x) \neq 0$ is a non zero polynomial of degree at most $2n + 1$, such that $\phi(x_i) = 0$ and $\phi'(x_i) = 0$. But $\phi(x_i) = 0$ for all $i = 0, 1, 2, \ldots, n$ implies that $\phi'(\xi_i) = 0$, for some $\xi_i \in (x_{i-1}, x_i)$ for all $i = 1, 2, \ldots, n$. Thus we found n distinct zeros of ϕ' other than $x_i's$. This shows that ϕ' has $2n + 1$ distinct zeros, which is a contradiction to the fact that $\phi'(x)$ is nonzero polynomial of degree at most $2n$ (see Corollary 2.3). We note that $\phi'(x)$ can not be a zero polynomial because $\phi(x)$ can not be a non zero constant polynomial as $\phi(x_i) = 0$. □

We can also compute the error in Hermite interpolation.

Theorem 2.7. *Let f be $2n + 2$ times differentiable real valued function defined on interval $[a, b]$. Let P_{2n+1} be the unique Hermite interpolating polynomial of degree at most $2n + 1$, satisfying*

$$P_{2n+1}(x_i) = f(x_i), P_{2n+1}'(x_i) = f'(x_i), \quad i = 0, 1, 2, \ldots, n. \tag{2.84}$$

Then the error in interpolation at some point $x_i \neq x \in [a,b]$ is given by

$$E_{2n+1}f(x) = (x-x_0)^2(x-x_1)^2\dots(x-x_n)^2\frac{1}{(2n+2)!}f^{2n+2}(\xi) = \frac{\omega^2(x)}{(2n+2)!}f^{2n+2}(\xi). \quad (2.85)$$

Where ξ is some point in the smallest interval I containing all the node points and depends upon x. And error bound on I is given by

$$|E_{2n+1}f(x)| \leq \max_{x\in I}|\omega^2(x)|\frac{M_{2n+2}}{(2n+2)!}. \quad (2.86)$$

Proof. Let x be some point in the interval other than node points. Let $\phi(t)$ be a function defined on interval $[a,b]$ by

$$\phi(t) = f(t) - P_{2n+1}(t) - (t-x_0)^2(t-x_1)^2\dots(t-x_n)^2K, \quad (2.87)$$

where K is a constant determined by the equation $\phi(x) = 0$, that is,

$$f(x) - P_{2n+1}(x) - (x-x_0)^2(x-x_1)^2\dots(x-x_n)^2K = 0. \quad (2.88)$$

It is clear that $\phi(x_i) = 0$ for all $i = 0,1,2,\dots,n$. Thus function ϕ has $n+2$ zeros say $\{x, x_0, x_1, \dots, x_n\}$. Hence ϕ' must have at least $n+1$ zeros at some points other than $\{x, x_0, x_1, \dots, x_n\}$. Moreover $\phi'(x_i) = 0$ for all $i = 0,1,2,\dots,n$. Thus ϕ' has at least $2n+2$ distinct zeros. Hence $(2n+1)_{th}$ derivative of ϕ', that is, $\phi^{(2n+2)_{th}}$ must have at least one zero ξ in the interval I. Therefore,

$$0 = \phi^{2n+2}(\xi) = f^{2n+2}(\xi) - 0 - (2n+2)!K. \quad (2.89)$$

This implies that $K = \frac{1}{2n+2!}f^{2n+2}(\xi)$. And hence from (2.87) and (2.88),

$$E_{2n+1}f(x) = f(x) - P_{2n+1}(x) = (x-x_0)^2(x-x_1)^2\dots(x-x_n)^2\frac{1}{(2n+2)!}f^{2n+2}(\xi). \quad (2.90)$$

Clearly (2.86) follows if we consider $M_{2n+2} = \max_{x\in I}|f^{2n+2}(x)|$. □

Problem 2.9. Find Lagrange's fundamental polynomial to interpolate the following data.

$$f(0) = 2, f'(0) = 0, f(1) = 2, f'(1) = 1.$$

Solution. Let $x_0 = 0, x_1 = 1$. And $L_0(x) = \frac{(x-x_1)}{(x_0-x_1)} = \frac{(x-1)}{(0-1)}$. Similarly $L_1(x) = \frac{(x-x_0)}{(x_1-x_0)} = \frac{(x-0)}{(1-0)}$. Thus, $L_0'(x_0) = L_0'(0) = -1$ and $L_1'(x_1) = L_1'(1) = 1$. Using,

$$P(x) = \sum_{i=0}^{n}\left[1 - 2(x-x_i)L_i'(x_i)\right]L_i^2(x)f(x_i) + \sum_{i=0}^{n}(x-x_i)L_i^2(x)f'(x_i).$$

$$\begin{aligned}
P_3(x) &= \sum_{i=0}^{1}\left[1 - 2(x-x_i)L_i'(x_i)\right]L_i^2(x)f(x_i) + \sum_{i=0}^{1}(x-x_i)L_i^2(x)f'(x_i),\\
&= [1-2(x-0)L_0'(0)]L_0^2(x)f(0) + [1-2(x-1)L_1'(1)]L_1^2(x)f(1)\\
&\quad +(x-0)L_0^2(x)f'(0) + (x-1)L_1^2(x)f'(1)\\
&= [1-2x(-1)](1-x)^2 2 + [1-2(x-1)1]x^2 2\\
&\quad +x(1-x)^2(0) + (x-1)x^2(1)\\
&= 2(1+2x)(x-1)^2 + 2(3-2x)x^2 + (x-1)x^2\\
&= x^3 - x^2 + 2
\end{aligned}$$

2.9. Newton's method to find Hermite Interpolating Polynomial

We further consider Problem (2.59). And assume that the polynomial of degree three in the following form.

$$P(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + a_3(x - x_0)^2(x - x_1). \tag{2.91}$$

If this polynomial satisfy the data $P(x_0) = f(x_0), P'(x_0) = f'(x_0), P(x_1) = f(x_1), P,(x_1) = f,(x_1)$, then we must have

$$\begin{aligned} a_0 &= f(x_0) \\ a_1 &= f'(x_0) \\ a_0 + a_1(x_1 - x_0) + a_2(x_1 - x_0)^2 &= f(x_1) \\ a_1 + 2a_2(x_1 - x_0) + a_3(x_1 - x_0)^2 &= f'(x_1) \end{aligned}$$

If we assume that the function is smooth enough (as many times differentiable as we want), then $f'(x_0)$ can be written as $f[x_0, x_0]$, and $f'(x_1) = f[x_1, x_1]$. Hence, we have $f(x_0) + f[x_0, x_0](x_1 - x_0) + a_2(x_1 - x_0)^2 = f(x_1)$ or $f[x_0, x_0] + a_2(x_1 - x_0) = f[x_0, x_1]$, or $a_2 = f[x_0, x_0, x_1]$. Further from the last equation we have $a_1 + a_2(x_1 - x_0) + a_2(x_1 - x_0) + a_3(x_1 - x_0)^2 = f[x_1, x_1]$, or $f[x_0, x_1] + a_2(x_1 - x_0) + a_3(x_1 - x_0)^2 = f[x_1, x_1]$, or $f[x_0, x_0, x_1](x_1 - x_0) + a_3(x_1 - x_0)^2 = f[x_1, x_1] - f[x_0, x_1]$, or $f[x_0, x_0, x_1] + a_3(x_1 - x_0) = f[x_0, x_1, x_1]$, or $a_3 = f[x_0, x_0, x_1, x_1]$. Thus

$$P(x) = f(x_0) + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1). \tag{2.92}$$

Remark 2.15. In general if there are $n + 1$ nodes one can compute Newton's divided difference table with each node considered twice, that is, total $2n+2$ total number of nodes. And to compute the the first divided difference relative to repeated points x_i, x_i we directly use the data given to us because $f[x_i, x_i] = f'(x_i)$.

Example 2.9. We will use Newton's extended divided difference table to find a polynomial, which satisfies the following data in which at each node we need to match derivatives up to certain order.

$$P(0) = 1, P'(0) = 2, P''(0) = -2, P(1) = 3, P'(1) = 5, P''(1) = 18, P'''(1) = 60.$$

x	$f(x)$	$f[x,x]$	$f[x,x,x]$	$f[x,x,x,x]$
0	1	$f[0,0] = f'(0) = 2$	$f[0,0,0] = f''(0)/2 = -1$	$f[0,0,0,1] = 0 - (-1) = 1$
0	1	$f[0,0] = f'(0) = 2$	$f[0,0,1] = 0$	$f[0,0,1,1] = 3 - 0 = 3$
0	1	$f[0,1] = (3-1)/(1-0) = 2$	$f[0,1,1] = 3$	$f[0,1,1,1] = 9 - 3 = 6$
1	3	$f[1,1] = f'(1) = 5$	$f[1,1,1] = f''(1)/2 = 9$	$f[1,1,1,1] = f'''(1)/6 = 10$
1	3	$f[1,1] = f'(1) = 5$	$f[1,1,1] = f''(1)/2 = 9$	...
1	3	$f[1,1] = f'(1) = 5$	...	...
1	3	...	...	...

x	$f[x,x,x,x,x]$	$f[x,x,x,x,x,x]$	$f[x,x,x,x,x,x,x]$
0	$f[0,0,0,1,1] = 3 - 1 = 2$	$f[0,0,0,1,1,1] = 1$	$f[0,0,0,1,1,1,1] = 0$
0	$f[0,0,1,1,1] = 6 - 3 = 3$	$f[0,0,1,1,1,1] = 1$	...
0	$f[0,1,1,1,1] = 10 - 6 = 4$	...	...
1	...	...	...
1	...	...	...
1	...	...	...
1	...	...	...

$P(x) = f(0) + f[0,0]x + f[0,0,0]x^2 + f[0,0,0,1]x^3 + f[0,0,0,1,1]x^3(x-1) + f[0,0,0,1,1,1]x^3(x-1)^2 + f[0,0,0,1,1,1,1]x^3(x-1)^3 = x^5 - x^2 + 2x + 1.$

Remark 2.16. One can not use Newton's extended divided difference table if the value of certain order derivative is known at some node, but the value of lower order derivative is not known. We can not find a polynomial of degree three such that $P(0) = 1, P''(0) = 2, P''(1) = 0, P'''(1) = 1$.

2.10. Piecewise Linear Interpolation

In practical purpose it has been observed that the approximation of a given function by linear pieces is better than the one polynomial of higher degree. Let f be a nice function with $f(x_0), f(x_1), \ldots, f(x_n)$ known at $n+1$ distinct points $x_0, x_1, \ldots, x_n$. We aim to find n lines $s_i, i = 1, 2, ..., n$, such that each s_i interpolates the function at the end points of the interval $[x_{i-1}, x_i]$, that is,

$$s_i(x_{i-1}) = f(x_{i-1}), s_i(x_i) = f(x_i), \qquad i = 1, 2, ...n. \tag{2.93}$$

Using Lagrange's linear interpolating polynomial we get,

$$\begin{aligned} s_i(x) &= \frac{(x - x_i)}{(x_{i-1} - x_i)} f(x_{i-1}) + \frac{(x - x_{i-1})}{(x_i - x_{i-1})} f(x_i) \quad x \in [x_{i-1}, x_i], i = 1, 2, ...n. \\ &= 0 \qquad x \in [x_{i-1}, x_i]^c. \end{aligned} \tag{2.94}$$

Thus $P_1(x) = \sum_{i=1}^{n} s_i(x)$ is the desired piecewise linear interpolating polynomial satisfying the given data. If we write the shape function as

$$N_i(x) = \begin{cases} 0, & x \le x_{i-1} \\ \frac{(x - x_{i-1})}{(x_i - x_{i-1})}, & x_{i-1} \le x \le x_i \\ \frac{(x - x_{i+1})}{(x_i - x_{i+1})}, & x_i \le x \le x_{i+1} \\ 0, & x_{i+1} \le x \end{cases} . \tag{2.95}$$

Then $P_1(x) = \sum_{i=1}^{n} N_i(x) f(x_i)$. The error in piecewise linear interpolation is given by

$$E_1 f(x) = \frac{1}{2!}(x - x_{i-1})(x - x_i) f''(\xi), \xi \in [x_{i-1}, x_i].$$

Remark 2.17. Here it is important to note that on each subinterval the expression for error function is different. So to find the uniform error bound we need to take the maximum of error bounds on different subintervals. Thus

$$|Ef(x)| \le \frac{M_2}{2} \max_i \left\{ \frac{|x_i - x_{i-1}|^2}{2} \right\}. \tag{2.96}$$

2.11. Cubic Spline Interpolation

Here we aim to find n cubic polynomials s_i, $i = 1, 2, ..., n$, such that each s_i interpolates the function f at the node points x_{i-1}, x_i in the interval $[x_{i-1}, x_i]$ such that the spline s, which we get after adjoining each cubic curve at the node points, is not only continuous on the interval (x_0, x_n) but s' and s'' is also continuous on this interval. These conditions can also be stated mathematically as follows.

- $s_i(x_{i-1}) = f(x_{i-1})$ for all $i = 1, 2, ..., n$,
- $s_i(x_i) = f(x_i)$ for all $i = 1, 2, ..., n$,
- $s_i'(x_i) = s_{i+1}'(x_i)$ for all $i = 1, 2, ..., n-1$,
- $s_i''(x_i) = s_{i+1}''(x_i)$ for all $i = 1, 2, ..., n-1$.

These are total $4n-2$ equations, but to obtain polynomials uniquely we need two more conditions. These two conditions are known as boundary conditions. In practical purpose there are two types of boundary conditions. Firstly, free boundary conditions is given by

$$s_1''(x_0) = s_n''(x_n) = 0. \tag{2.97}$$

And clamped boundary conditions are given by

$$s_1'(x_0) = f'(x_0) \quad \text{and} \quad s_n'(x_n) = f'(x_n). \tag{2.98}$$

At node points we tie two distinct polynomial, so some times these node are also called as knots. Since $f(x_i)$ is given, in first two conditions both RHS and LHS is fixed. But in last two conditions RHS=LHS is to be determined so that our $s_i's$ should satisfy all the conditions simultaneously. Suppose for notational simplification we consider new variables

$$m_0 = s_1'(x_0), M_0 = s_1''(x_0) \tag{2.99}$$
$$m_i = s_i'(x_i) = s_{i+1}'(x_i), \quad i = 1, 2, ..., n-1, \tag{2.100}$$
$$M_i = s_i''(x_i) = s_{i+1}''(x_i), \quad i = 1, 2, ..., n-1. \tag{2.101}$$
$$m_n = s_n'(x_n), M_n = s_n''(x_n) \tag{2.102}$$

Thus all the four conditions will be automatically satisfied, if we demand our each cubic polynomial s_i to satisfy the following conditions.

$$\begin{aligned} s_i(x_{i-1}) &= f(x_{i-1}) \\ s_i(x_i) &= f(x_i) \\ s_i'(x_{i-1}) &= m_{i-1} \\ s_i'(x_i) &= m_i \\ s_i''(x_{i-1}) &= M_{i-1} \\ s_i''(x_i) &= M_i \end{aligned}$$

Since in the interval $[x_{i-1}, x_i]$ we want a cubic polynomial s_i, that is, we want s_i'' to be a linear polynomial to satisfy the last two conditions. Using Lagrange's formula we get,

$$s_i''(x) = \frac{(x-x_i)}{(x_{i-1}-x_i)} M_{i-1} + \frac{(x-x_{i-1})}{(x_i - x_{i-1}} M_i.$$

Now if we write $h_i = x_i - x_{i-1}$, then

$$s_i(x) = \frac{(x-x_i)^3}{-6h_i} M_{i-1} + \frac{(x-x_{i-1})^3}{6h_i} M_i + a_i x + b_i.$$

And since $s_i(x_{i-1}) = f(x_{i-1})$ and $s_i(x_i) = f(x_i)$, we must have

$$f(x_{i-1}) = h_i^2 M_{i-1}/6 + a_i x_{i-1} + b_i \quad \text{and} \quad f(x_i) = h_i^2 M_i/6 + a_i x_i + b_i.$$

Solving these equations we get

$$\begin{aligned} a_i &= \frac{f(x_i) - f(x_{i-1})}{h_i} - \frac{M_i - M_{i-1}}{6} h_i, \\ b_i &= \frac{x_i f(x_{i-1}) - x_{i-1} f(x_i)}{h_i} - \frac{x_i M_{i-1} - x_{i-1} M_i}{6} h_i. \end{aligned}$$

Substituting these values of a_i, b_i in the expression of s_i, we get

$$\begin{aligned} s_i(x) &= \frac{(x-x_i)^3}{-6h_i} M_{i-1} + \frac{(x-x_{i-1})^3}{6h_i} M_i + \frac{f(x_i) - f(x_{i-1})}{h_i} x - \frac{M_i - M_{i-1}}{6} h_i x \\ &\quad + \frac{x_i f(x_{i-1}) - x_{i-1} f(x_i)}{h_i} - \frac{x_i M_{i-1} - x_{i-1} M_i}{6} h_i. \end{aligned} \tag{2.103}$$

This gives that

$$s_i'(x) = \frac{(x-x_i)^2}{-2h_i}M_{i-1} + \frac{(x-x_{i-1})^2}{2h_i}M_i + \frac{f(x_i)-f(x_{i-1})}{h_i} - \frac{M_i - M_{i-1}}{6}h_i. \tag{2.104}$$

Further we use the condition $s_i'(x_i) = s_{i+1}'(x_i)$ to get

$$h_i M_{i-1} + 2(h_i + h_{i+1})M_i + h_{i+1}M_{i+1} = 6f[x_i, x_{i+1}] - 6f[x_{i-1} - x_i]. \tag{2.105}$$

By the assumption of clamped boundary conditions $f'(x_0) = s_1'(x_0)$ and $f'(x_n) = s_n'(x_n)$, we get

$$\begin{aligned} 2h_1M_0 + h_1M_1 &= 6f[x_0, x_1] - 6f[x_0, x_0] \\ h_nM_{n-1} + 2h_nM_n &= 6f[x_n, x_n] - 6f[x_{n-1}, x_n] \end{aligned}$$

Now using (2.105) and above two equations we write the matrix form of the system of linear equations.

$$\begin{bmatrix} 2h_1 & h_1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & h_2 & 2(h_2+h_3) & h_3 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & h_{i-1} & 2(h_{i-1}+h_i) & h_i & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & h_i & 2(h_i+h_{i+1}) & h_{i+1} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & h_{i+1} & 2(h_{i+1}+h_{i+2}) & h_{i+2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & h_n & 2h_n \end{bmatrix} \times$$
$$\begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = 6 \begin{bmatrix} f[x_0,x_1] - f[x_0,x_0] \\ f[x_1,x_2] - f[x_0,x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i, x_{i+1}] - f[x_{i-1} - x_i] \\ \dots \\ \dots \\ f[x_{n-1}, x_n] - f[x_{n-1}, x_{n-2}] \\ f[x_n, x_n] - f[x_{n-1}, x_n] \end{bmatrix} \tag{2.106}$$

Remark 2.18. This matrix equation (correspond to clamped boundary) takes simple form when step size is fixed, that is, $h_i = h$.

$$\begin{bmatrix} 2 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 4 & 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 4 & 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 2 \end{bmatrix} \times \begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = \frac{6}{h} \begin{bmatrix} f[x_0,x_1] - f[x_0,x_0] \\ f[x_1,x_2] - f[x_0,x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i, x_{i+1}] - f[x_{i-1} - x_i] \\ \dots \\ \dots \\ f[x_{n-1}, x_n] - f[x_{n-1}, x_{n-2}] \\ f[x_n, x_n] - f[x_{n-1}, x_n] \end{bmatrix} \tag{2.107}$$

Remark 2.19. Further for the free boundary condition $M_0 = 0, M_n = 0$ we replace the first and last row of the matrix with these two conditions and get

$$\begin{bmatrix} h_1 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & h_2 & 2(h_2+h_3) & h_3 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & h_{i-1} & 2(h_{i-1}+h_i) & h_i & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & h_i & 2(h_i+h_{i+1}) & h_{i+1} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & h_{i+1} & 2(h_{i+1}+h_{i+2}) & h_{i+2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & h_n \end{bmatrix} \times$$

$$\begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = 6 \begin{bmatrix} 0 \\ f[x_1,x_2]-f[x_0,x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i,x_{i+1}]-f[x_{i-1}-x_i] \\ \dots \\ \dots \\ f[x_{n-1},x_n]-f[x_{n-1},x_{n-2}] \\ 0 \end{bmatrix}. \tag{2.108}$$

This equation also takes a simple form when step size is fixed.

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 4 & 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 4 & 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 1 \end{bmatrix} \times \begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = \frac{6}{h} \begin{bmatrix} 0 \\ f[x_1,x_2]-f[x_0,x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i,x_{i+1}]-f[x_{i-1}-x_i] \\ \dots \\ \dots \\ f[x_{n-1},x_n]-f[x_{n-1},x_{n-2}] \\ 0 \end{bmatrix}. \tag{2.109}$$

Remark 2.20. In all above matrix equations first and last row of the matrix appear because of extra boundary condition at x_0 and x_n respectively, which might be either clamped boundary or the free boundary condition. But one can also think of having mixed boundary condition like free boundary at x_0, that is, $f''(x_0) = M_0 = 0$ and clamped boundary at x_n, that is $s'(x_n) = f'(x_n)$,

or equivalently, $h_nM_{n-1} + 2h_nM_n = 6f[x_n, x_n] - 6f[x_{n-1}, x_n]$ or viceversa.

$$\begin{bmatrix} h_1 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ h_1 & 2(h_1+h_2) & h_2 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & h_2 & 2(h_2+h_3) & h_3 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & h_{i-1} & 2(h_{i-1}+h_i) & h_i & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & h_i & 2(h_i+h_{i+1}) & h_{i+1} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & h_{i+1} & 2(h_{i+1}+h_{i+2}) & h_{i+2} & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & h_{n-1} & 2(h_{n-1}+h_n) & h_n \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & h_n & 2h_n \end{bmatrix} \times$$

$$\begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = 6 \begin{bmatrix} 0 \\ f[x_1, x_2] - f[x_0, x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i, x_{i+1}] - f[x_{i-1} - x_i] \\ \dots \\ \dots \\ f[x_{n-1}, x_n] - f[x_{n-1}, x_{n-2}] \\ f[x_n, x_n] - f[x_{n-1}, x_n \end{bmatrix}. \tag{2.110}$$

Further if the step size is fixed, we get

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 2 & 4 & 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 1 & 4 & 1 & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 4 & 1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 1 & 2 \end{bmatrix} \times \begin{bmatrix} M_0 \\ M_1 \\ \dots \\ \dots \\ M_{i-1} \\ M_i \\ M_{i+1} \\ \dots \\ \dots \\ M_n \end{bmatrix} = \frac{6}{h} \begin{bmatrix} 0 \\ f[x_1, x_2] - f[x_0, x_1] \\ \dots \\ \dots \\ \dots \\ f[x_i, x_{i+1}] - f[x_{i-1} - x_i] \\ \dots \\ \dots \\ f[x_{n-1}, x_n] - f[x_{n-1}, x_{n-2}] \\ f[x_n, x_n] - f[x_{n-1}, x_n] \end{bmatrix}. \tag{2.111}$$

2.12. Error in cubic spline interpolation

It can be shown that

$$|E(x)| = |f(x) - s(x)| \leq \frac{5M_4}{384}(\max_i\{|h_i|\})^4. \tag{2.112}$$

Further it can also be shown that

$$|f'(x) - s'(x)| \leq \frac{M_4}{24}(\max_i\{|h_i|\})^3. \tag{2.113}$$

This shows that we can also approximate f' using cubic spline just by knowing some functional values at nodes and M_4.

3 Numerical integration

Our aim in this chapter is to find the approximate value of some definite integral specially in the cases when the value of the integrand is known only at certain points, or the integral of the integrand is not known in terms of standard functions.

Question 3.1. Can we use interpolating polynomials of some function f to find the approximate integral of f on $[a,b]$, that is, $\int_a^b f(x)dx$? Is this approximation a good one?

If $P(x)$ is some interpolating polynomial to $f(x)$ such that the error is as small as some given positive number ϵ, that is, $|E(x)| = |f(x) - P(x)| \leq \epsilon$, then

$$\begin{aligned}\left|\int_a^b f(x)dx - \int_a^b P(x)dx\right| &= \left|\int_a^b (f(x) - P(x))dx,\right| \\ &\leq \int_a^b |f(x) - P(x)|dx, \\ &\leq \int_a^b \epsilon\, dx, \\ &= \epsilon\,(b-a). \end{aligned} \tag{3.1}$$

Thus we can approximate the integral of f by the integral of interpolating polynomial with the desirable accuracy.

3.1. Newton-Cotes Methods

In this method we use interpolating polynomials $P(x)$ to approximate $\int_a^b f(x)dx$ by $\int_a^b P(x)dx$ with error in integration as $\int_a^b [f(x) - P(x)]dx = \int_a^b E(x)dx$.

3.2. Trapezoidal Rule

Suppose f is known at the two end points of the interval $x_0 = a, x_1 = b$ and $P_1(x)$ is the linear approximation to the function f at nodes x_0, x_1, then by Lagrange's method we write

$$P_1(x) = \frac{(x-x_1)}{(x_0-x_1)}f(x_0) + \frac{(x-x_0)}{(x_1-x_0)}f(x_1). \tag{3.2}$$

Now, $\int_{x_0}^{x_1} P_1(x)dx = \int_{x_0}^{x_1} \left(\frac{(x-x_1)}{(x_0-x_1)}f(x_0)\right)dx + \int_{x_0}^{x_1} \left(\frac{(x-x_0)}{(x_1-x_0)}f(x_1)\right)dx = \frac{(x_1-x_0)}{2}f(x_0) + \frac{(x_1-x_0)}{2}f(x_1)$. And if $h_1 = x_1 - x_0$, then

$$\int_{x_0}^{x_1} P_1(x)dx = \frac{h_1}{2}[f(x_0) + f(x_1)]. \tag{3.3}$$

Thus by Trapezoidal rule the approximate value of $\int_a^b f(x)dx$ is given by (3.3). Further from (2.46)

$$f(x) - P_1(x) = (x-x_0)(x-x_1)f''(\xi_x)/2. \tag{3.4}$$

And hence,

$$\begin{aligned}\left|\int_{x_0}^{x_1} [f(x) - P_1(x)]\,dx\right| &\leq \int_{x_0}^{x_1} \left|\frac{1}{2}(x-x_0)(x-x_1)f''(x)\right| dx \\ &\leq \int_{x_0}^{x_1} \frac{1}{2}|(x-x_0)|\,|(x-x_1)| \max_{x\in[x_0,x_1]} |f''(x)|\,dx \\ &= \frac{1}{2}\max_{x\in[x_0,x_1]} |f''(\xi_x)| \int_{x_0}^{x_1} (x-x_0)\,(x_1-x)dx \\ &= \frac{1}{2}\max_{x\in[x_0,x_1]} |f''(x)|\,\frac{(x_1-x_0)^3}{6}. \end{aligned} \tag{3.5}$$

3.3. Composite Trapezoidal Rule

Now if we partition $[a,b]$ in n subintervals $[x_{i-1},x_i], i = 1,2,\ldots,n$. We want to apply trapezoidal rule in each subinterval to approximate $\int_a^b f(x)dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x)dx$ by $\sum_{i=1}^n \int_{x_{i-1}}^{x_i} P_{1,i}(x)dx$. Thus

$$\begin{aligned}\int_a^b f(x)dx = \sum_{i=1}^n \int_{x_{i-1}}^{x_i} f(x)dx &\approx \sum_{i=1}^n \int_{x_{i-1}}^{x_i} P_{1,i}(x)dx,\\ &\approx \sum_{i=1}^n \left[\frac{h_i}{2}[f(x_{i-1})+f(x_i)]\right],\\ &= \frac{h_1}{2}f(x_0) + \left[\sum_{i=1}^{n-1}\left(\frac{h_{i+1}+h_i}{2}\right)f(x_i)\right] + \frac{h_n}{2}f(x_n). \quad (3.6)\end{aligned}$$

Similar to (3.5), bound for the error of integration through trapezoidal rule in the interval $[x_{i-1},x_i]$ is given by

$$\left|\int_{x_{i-1}}^{x_{i+1}} [f(x)-P_{1,i}(x)]dx\right| \le \frac{h_i^3}{12}\max_{x\in[x_{i-1},x_i]}|f''(x)|. \quad (3.7)$$

And the total error in composite trapezoidal rule will be bounded by

$$\sum_{i=1}^n \left(\frac{h_i^3}{12}\max_{x\in[x_{i-1},x_i]}|f''(x)|\right) \le \max_{x\in[a,b]}|f''(x)|\sum_{i=1}^n\left(\frac{h_i^3}{12}\right) = \frac{M_2}{12}\sum_{i=1}^n h_i^3. \quad (3.8)$$

Further if our step size is fixed, that is, $x_i - x_{i-1} = h_i = h$, then

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{h}{2}f(x_0) + \left[\sum_{i=1}^{n-1} hf(x_i)\right] + \frac{h}{2}f(x_n)\\ &= \frac{h}{2}\left(f(x_0) + \left[\sum_{i=1}^{n-1}2f(x_i)\right] + f(x_n)\right)\\ &= \frac{h}{2}\left[f(x_0)+2f(x_1)+\ldots+2f(x_{n-1})+f(x_n)\right]. \quad (3.9)\end{aligned}$$

And by (3.8), total error in this case will be bounded by $\frac{n}{12}M_2h^3 = \frac{1}{12n^2}M_2(nh)^3 = \frac{(b-a)^3}{12n^2}M_2$.

3.4. Simpson's one-third rule

Suppose now we have three equidistant nodes $a = x_0, x_1 = x_0+h, x_2 = x_0+2h = b$ and want to approximate $\int_{x_0}^{x_2} f(x)dx$ by $\int_{x_0}^{x_2} P_2(x)dx$ with error as $\int_{x_0}^{x_2}[f(x)-P_2(x)]dx = \int_{x_0}^{x_2} E_2(x)dx$. Since there are three nodes we should approximate by a polynomial of degree at most two. According to Newton's method we can write the quadratic polynomial interpolating at given three node as follows

$$P_2(x) = f(x_0) + f[x_0,x_0+h](x-x_0) + f[x_0,x_0+h,x_0+2h](x-x_0)(x-x_0-h). \quad (3.10)$$

Now we want to compute the integral

$$\begin{aligned}\int_{x_0}^{x_2} P_2(x)dx &= f(x_0)2h + f[x_0,x_0+h]2h^2 + f[x_0,x_0+h,x_0+2h]\int_{x_0}^{x_0+2h}(x-x_0)(x-x_0-h)dx,\\ \int_{x_0}^{x_0+2h} P_2(x)dx &= 2hf(x_0) + 2h^2 f[x_0,x_0+h] + \frac{2}{3}h^3 f[x_0,x_0+h,x_0+2h],\\ &= 2hf(x_0) + 2h^2\left[\frac{f(x_0+h)-f(x_0)}{h}\right] + \frac{2}{3}h^3\left[\frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{2h^2}\right],\\ &= \frac{h}{3}\left[f(x_0+2h)+4f(x_0+h)+f(x_0)\right] = \frac{h}{3}\left[f(x_2)+4f(x_1)+f(x_0)\right]. \quad (3.11)\end{aligned}$$

Thus if the step size is fixed, then

$$\int_{x_0}^{x_2} f(x)dx \approx \frac{h}{3}\left[f(x_2) + 4f(x_1) + f(x_0)\right]. \tag{3.12}$$

Further since $E_2(x) = f(x) - P_2(x) = (x-x_0)(x-x_0-h)(x-x_0-2h)\frac{f'''(\xi)}{3!}$, error in integration should be

$$\int_{x_0}^{x_0+2h} E_2(x)dx = \int_{x_0}^{x_0+2h} (x-x_0)(x-x_0-h)(x-x_0-2h)\frac{f'''(\xi)}{6}dx. \tag{3.13}$$

Thus, if $M_3 = \max_{x\in[x_0,x_2]} |f'''(x)|$, then

$$\begin{aligned}
\left|\int_{x_0}^{x_0+2h} E_2(x)dx\right| &\leq \int_{x_0}^{x_0+2h} \left|(x-x_0)(x-x_0-h)(x-x_0-2h)\frac{f'''(\xi)}{6}\right| dx \\
&\leq \int_{x_0}^{x_0+2h} |(x-x_0)(x-x_0-h)(x-x_0-2h)|\,\frac{M_3}{6}dx \\
&= \frac{M_3}{6}\left[\int_{x_0}^{x_1}(x-x_0)(x_1-x)(x_2-x)dx + \int_{x_1}^{x_2}(x-x_0)(x-x_1)(x_2-x)dx\right] \\
&= \frac{M_3}{6}\left[\int_{x_0}^{x_0+h}(x-x_0)(x_0+h-x)(x_0+2h-x)dx + \right. \\
&\qquad \left. \int_{x_0+h}^{x_0+2h}(x-x_0)(x-x_0-h)(x_0+2h-x)dx\right]
\end{aligned}$$

By substitution $x - x_0 = hu$,

$$\begin{aligned}
\left|\int_{x_0}^{x_0+2h} E_2(x)dx\right| &\leq \frac{M_3h^4}{6}\left[\int_0^1 u(1-u)(2-u)du + \int_1^2 u(u-1)(2-u)du\right] \\
&= \frac{M_3h^4}{6}\left[(u^2 - u^3 - u^4/4)|_0^1 + (-u^2 + u^3 - u^4/4)|_0^1\right] \\
&= \frac{M_3h^4}{6}\left[(1-1+1/4) - 0 + (-4+8-4) - (-1/4+1-1)\right] \\
&\leq \frac{M_3h^4}{12}.
\end{aligned} \tag{3.14}$$

This shows that we can approximate $\int_{x_0}^{x_0+2h} f(x)dx$ by $\frac{h}{3}\left[f(x_0) + 4f(x_0+h) + f(x_0+2h)\right]$ with a desired accuracy if h is small enough.

Observation 3.1. Clearly because of three nodes, approximation to a two degree polynomial is exact. It can also be seen easily from (3.13) that if f is a three degree polynomial, that is, $f(x) = a_0 + a_1x + a_2x^2 + a_3x^3$, then

$$\begin{aligned}
\int_{x_0}^{x_0+2h} E_2(x)dx &= \int_{x_0}^{x_0+2h}(x-x_0)(x-x_0-h)(x-x_0-2h)\frac{f'''(\xi)}{6}dx \\
&= a_3\int_{x_0}^{x_0+2h}(x-x_0)(x-x_0-h)(x-x_0-2h)dx = 0.
\end{aligned} \tag{3.15}$$

Thus, because of odd number of equidistant nodes, the approximation to the integral of **three** degree polynomial is also exact.

3.5. Another form of truncation error of integration in Simpson's one-third rule

Since we are approximating the function at three node points x_0, x_0+h, x_0+2h, we have $f(x) = f(x_0) + f[x_0, x_0+h](x-x_0) + f[x_0, x_0+h, x_0+2h](x-x_0)(x-x_0-h) + f[x_0, x_0+h, x_0+2h, x](x-x_0)(x-x_0-h)(x-x_0-2h) = P_2(x) + \frac{f'''(\xi)}{6}(x-x_0)(x-x_0-h)(x-x_0-2h)$. And the approximation to the integral $\int_{x_0}^{x_0+2h} f(x)dx$ is $\int_{x_0}^{x_0+2h} P_2(x)dx = \frac{h}{3}\left[f(x_0) + 4f(x_0+h) + f(x_0+2h)\right]$ with truncation error of integration given by (3.13). But using (3.15), the crux of the observation 3.1, one can obtain the same approximation as (3.12) to the integral $\int_{x_0}^{x_0+2h} f(x)dx$ by approximating $f(x)$ with a three degree interpolating polynomial $P_3(x)$ satisfying the conditions $P_3(x_0) = f(x_0), P_3(x_0+h) = f(x_0+h), P_3(x_0+2h) = f(x_0+2h)$ and an extra condition $P_3'(x_0+h) = 0$. And this is because $P_3(x) = P_2(x) + f[x_0, x_0+h, x_0+2h, x_0+h](x-x_0)(x-x_0-h)(x-x_0-2h)$ and

$$\begin{aligned}
&\int_{x_0}^{x_0+2h} P_3(x)dx = \int_{x_0}^{x_0+2h} \big[f(x_0) + f[x_0, x_0+h](x-x_0) + f[x_0, x_0+h, x_0+2h](x-x_0) \\
&(x-x_0-h)\big]dx + f[x_0, x_0+h, x_0+2h, x_0+h]\int_{x_0}^{x_0+2h}(x-x_0)(x-x_0-h)(x-x_0-2h)dx, \\
&= \int_{x_0}^{x_0+2h} P_2(x)dx + f[x_0, x_0+h, x_0+2h, x_0+h](0) = \int_{x_0}^{x_0+2h} P_2(x)dx. \qquad (3.16)
\end{aligned}$$

Thus the error of interpolation in this case is given by $f(x) - P_3(x) = f[x_0, x_0+h, x_0+2h, x_0+h, x](x-x_0)(x-x_0-h)^2(x-x_0-2h) = \frac{f^{iv}(\xi)}{4!}(x-x_0)(x-x_0-h)^2(x-x_0-2h)$. Further the truncation error of integration will be

$$\int_{x_0}^{x_0+2h} [f(x) - P_3(x)]dx = \int_{x_0}^{x_0+2h} (x-x_0)(x-x_0-h)^2(x-x_0-2h)\frac{f^{iv}(\xi)}{4!}dx. \qquad (3.17)$$

And bound for error in integration can be obtained as

$$\begin{aligned}
\left|\int_{x_0}^{x_0+2h} [f(x) - P_3(x)]\,dx\right| &= \left|\int_{x_0}^{x_0+2h} (x-x_0)(x-x_0-h)^2(x-x_0-2h)\frac{f^{iv}(\xi)}{4!}dx\right|, \\
\left|\int_{x_0}^{x_0+2h} [f(x) - P_2(x)]\,dx\right| &\le \frac{M_4}{4!}\int_{x_0}^{x_0+2h} \left|(x-x_0)(x-x_0-h)^2(x-x_0-2h)\right| dx, \\
&= \frac{M_4h^5}{4!}\int_{-1}^{1} \left|(u+1)(u)^2(u-1)\right| du, \quad x-x_0-h = uh, \\
&= \frac{M_4h^5}{4!}\int_{-1}^{1} \left|(u^4-u^2)\right| du = \frac{2M_4h^5}{4!}\int_0^1 \left|(u^4-u^2)\right| du, \\
&= \frac{2M_4h^5}{4!}\int_0^1 (u^2-u^4)du = \frac{2M_4h^5}{4!}\left(\frac{1}{3}-\frac{1}{5}\right), \\
&= \frac{M_4h^5}{90}.
\end{aligned}$$

Here since $\int_{x_0}^{x_0+2h} [f(x) - P_3(x)]\,dx = \int_{x_0}^{x_0+2h} [f(x) - P_2(x)]\,dx$, the bound for the error of integration in Simpson one-third rule can also be obtained as

$$\left|\int_{x_0}^{x_0+2h} [f(x) - P_2(x)]\,dx\right| \le \frac{M_4h^5}{90}. \qquad (3.18)$$

Remark 3.1. Thus we see from (3.14) and (3.18) that there are two bounds of order h^4 and h^5 respectively for the error of integration in Simpson's one-third rule. But for small h we can infer the better accuracy of approximating integral from the later bound (3.18). So often we use $\frac{M_4 h^5}{90}$ as error bound of integration in Simpson's one-third rule.

3.6. Composite Simpson's one-third rule

Now if we want to apply Simpson's one-third rule on a given interval $[a, b]$, we need to partition $[a, b]$ in to $2n$ number of equal length subintervals with nodes $a = x_0, x_1, x_2, \ldots, x_{2n-2}, x_{2n-1}, x_{2n} = b$, where $x_i = x_0 + ih$. In each two adjacent subintervals of the form $[x_{2i-2}, x_{2i-1}], [x_{2i-1}, x_{2i}]$ we approximate the integral $\int_{x_{2i-2}}^{x_{2i}} f(x)dx$ using Simpson's one-third rule. So that

$$\begin{aligned}\int_a^b f(x)dx &= \int_{x_0}^{x_{2n}} f(x)dx = \sum_{i=1}^{n} \int_{x_{2i-2}}^{x_{2i}} f(x)dx \\ &\approx \sum_{i=1}^{n} \int_{x_{2i-2}}^{x_{2i}} P_{2,i}(x)dx\end{aligned}$$

Now using (3.12), we get

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_{i=1}^{n} \frac{h}{3}\left[f(x_{2i-2}) + 4f(x_{2i-1}) + f(x_{2i})\right] \\ &= \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + 2f(x_4) + \ldots + 2f(x_{2i-2})\right. \\ &\quad \left. + 4f(x_{2i-1}) + 2f(x_{2i}) + \ldots + 2f(x_{2n-2}) + 4f(x_{2n-1}) + f(x_{2n})\right] \qquad (3.19)\end{aligned}$$

Further using (3.14) error will bounded as follows

$$\begin{aligned}\left|\int_a^b E(x)dx\right| &= \left|\sum_{i=1}^{n} \int_{x_{2i-2}}^{x_{2i}} [f(x) - P_{2,i}(x)]dx\right|, \\ &\leq \sum_{i=1}^{n} \left|\int_{x_{2i-2}}^{x_{2i}} [f(x) - P_{2,i}(x)]dx\right|, \\ &\leq \sum_{i=1}^{n} \frac{M_3 h^4}{12} = \frac{nM_3h^4}{12}, \qquad 2nh = b - a, \\ &= \frac{M_3(b-a)^4}{192n^3}. \qquad (3.20)\end{aligned}$$

Further using (3.18) one can also obtain the error bound as

$$\begin{aligned}\left|\int_a^b E(x)dx\right| &= \left|\sum_{i=1}^{n} \int_{x_{2i-2}}^{x_{2i}} [f(x) - P_{3,i}(x)]dx\right|, \\ &\leq \sum_{i=1}^{n} \left|\int_{x_{2i-2}}^{x_{2i}} [f(x) - P_{3,i}(x)]dx\right|, \\ &\leq \sum_{i=1}^{n} \frac{M_4 h^5}{90} = \frac{nM_4h^5}{90}, \qquad 2nh = b - a,, \\ &= \frac{M_4(b-a)^5}{2880n^4}. \qquad (3.21)\end{aligned}$$

Remark 3.2. As we discussed in Remark 3.1, in composite Simpson's one-third rule we also get two bounds for error of integration as given above. But we always try to subdivided interval $[a, b]$ in to least number of subintervals to obtain certain accuracy. Thus to obtain the error bound as ϵ, we need either $\frac{M_3(b-a)^4}{192n^3} < \epsilon$ or $\frac{M_4(b-a)^5}{2880n^4} < \epsilon$. And hence the minimum n, required for obtaining the desired error bound of integration, should satisfy

$$\min\left\{\left(\frac{M_3(b-a)^4}{192\epsilon}\right)^{1/3}, \left(\frac{M_4(b-a)^5}{2880\epsilon}\right)^{1/4}\right\} \leq n. \tag{3.22}$$

3.7. Simpson's 3/8 rule

Now our aim is to approximate $\int_a^b f(x)dx$ by approximating $f(x)$ with an interpolating polynomial $P_3(x)$ of degree at most three at four equidistant nodes $a = x_0, x_0 + h, x_0 + 2h, x_0 + 3h$. Newton's form of the polynomial $P_3(x)$ is as follows

$$\begin{aligned} P_3(x) &= f(x_0) + f[x_0, x_0+h](x-x_0) + f[x_0, x_0+h, x_0+2h](x-x_0)(x-x_0-h) \\ &\quad + f[x_0, x_0+h, x_0+2h, x_0+3h](x-x_0)(x-x_0-h)(x-x_0-2h). \end{aligned} \tag{3.23}$$

So that,

$$\begin{aligned} \int_{x_0}^{x_0+3h} P_3(x)dx &= f(x_0)\int_{x_0}^{x_0+3h} dx + f[x_0, x_0+h]\int_{x_0}^{x_0+3h}(x-x_0)dx \\ &\quad + f[x_0, x_0+h, x_0+2h]\int_{x_0}^{x_0+3h}(x-x_0)(x-x_0-h)dx \\ &\quad + f[x_0, x_0+h, x_0+2h, x_0+3h]\int_{x_0}^{x_0+3h}(x-x_0)(x-x_0-h)(x-x_0-2h)dx, \\ &= 3h\ f(x_0) + \left[\frac{f(x_0+h)-f(x_0)}{h}\right]\frac{9h^2}{2} \\ &\quad + \left[\frac{f(x_0+2h)-2f(x_0+h)+f(x_0)}{2h^2}\right]\frac{9h^3}{2} \\ &\quad + \left[\frac{f(x_0+3h)-3f(x_0+2h)+3f(x_0+h)-f(x_0)}{6h^3}\right]\frac{9h^4}{4}, \\ &= \frac{3h}{8}\left[f(x_0)+3f(x_0+h)+3f(x_0+2h)+f(x_0+3h)\right]. \end{aligned} \tag{3.24}$$

The above expression is the approximation to the $\int_{x_0}^{x_0+3h} f(x)\ dx$ in Simpson's three-eighth rule.

$$\int_{x_0}^{x_0+3h} f(x)\ dx \approx \frac{3h}{8}\left[f(x_0)+3f(x_0+h)+3f(x_0+2h)+f(x_0+3h)\right]. \tag{3.25}$$

Using (2.46) the error in interpolation is given by

$$E_3 f(x) = (x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)\frac{f^{\text{iv}}(\xi)}{4!}. \tag{3.26}$$

And the error in integration will be bounded as follows

$$\begin{aligned} \left|\int_{x_0}^{x_0+3h} E_3(x)dx\right| &= \left|\int_{x_0}^{x_0+3h}(x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)\frac{f^{\text{iv}}(\xi)}{4!}dx\right|, \\ &\leq \int_{x_0}^{x_0+3h}\left|(x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)\frac{f^{\text{iv}}(\xi)}{4!}\right|dx, \\ &\leq \frac{M_4}{24}\int_{x_0}^{x_0+3h}|(x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)|\,dx. \end{aligned}$$

Here M_4 is the maximum of $|f^{\text{iv}}|$ on the interval $[a, b]$. Now

$$\begin{aligned}
&\int_{x_0}^{x_0+3h} |(x-x_0)(x-x_0-h)(x-x_0-2h)(x-x_0-3h)|\,dx \\
&= \int_{-\frac{3}{2}}^{\frac{3}{2}} \left|\left(u+\frac{3}{2}\right)\left(u+\frac{1}{2}\right)\left(u-\frac{1}{2}\right)\left(u-\frac{3}{2}\right)\right| du \\
&= \int_{-\frac{3}{2}}^{\frac{3}{2}} \left|\left(u^2-\frac{9}{4}\right)\left(u^2-\frac{1}{4}\right)\right| du = 2\int_0^{\frac{3}{2}} \left|\left(u^2-\frac{9}{4}\right)\left(u^2-\frac{1}{4}\right)\right| du \\
&= 2\int_0^{\frac{1}{2}} \left(\frac{9}{4}-u^2\right)\left(\frac{1}{4}-u^2\right) du + 2\int_{\frac{1}{2}}^{\frac{3}{2}} \left(\frac{9}{4}-u^2\right)\left(u^2-\frac{1}{4}\right) du \\
&= 2\int_0^{\frac{1}{2}} \left(\frac{9}{16}-\frac{10}{4}u^2+u^4\right) du + 2\int_{\frac{1}{2}}^{\frac{3}{2}} \left(-\frac{9}{16}+\frac{10}{4}u^2-u^4\right) du, \\
&= 2\left(\frac{9}{16}u-\frac{10}{12}u^3+\frac{u^5}{5}\right)_0^{\frac{1}{2}} + 2\left(-\frac{9}{16}u+\frac{10}{12}u^3-\frac{u^5}{5}\right)_{\frac{1}{2}}^{\frac{3}{2}} = \frac{49}{30}
\end{aligned}$$

Thus the error bound for Simpson's 3/8 rule can be given as

$$\left|\int_{x_0}^{x_0+3h} f(x) - P_3(x)]dx\right| \leq \frac{M_4 h^5}{24}\frac{49}{30} = \frac{49M_4h^5}{720}. \tag{3.27}$$

3.8. Composite Simpson's 3/8 rule

Here we first subdivide the interval $[a, b]$ in to n number of equal length subintervals and aim to apply Simpson's 3/8 rule in each of these n subintervals. Thus we further need to subdivided these intervals into three equal length subintervals of width $h = (b-a)/3$. Now we have total $3n$ subintervals with $3n+1$ nodes as

$$x_0, x_1, x_2, x_3, \ldots, x_{3i}, x_{3i+1}, x_{3i+2}, x_{3(i+1)}, \ldots x_{3(n-1)}, x_{3(n-1)+1}, x_{3(n-1)+2}, x_{3n}$$

and apply Simpson's 3/8 rule in the interval of the form $[x_{3i}, x_{3(i+1)}]$ to obtain

$$\int_{x_{3i}}^{x_{3(i+1)}} f(x)\,dx \approx \frac{3h}{8}\left[f(x_{3i}) + 3f(x_{3i+1}) + 3f(x_{3i+2}) + f(x_{3(i+1)})\right]. \tag{3.28}$$

And the approximation for the integral on whole interval $[a, b]$ can be obtained as follows.

$$\begin{aligned}
&\int_a^b f(x)\,dx = \sum_{i=0}^{n-1}\int_{x_{3i}}^{x_{3(i+1)}} f(x)\,dx \approx \sum_{i=0}^{n-1}\frac{3h}{8}\left[f(x_{3i}) + 3f(x_{3i+1}) + 3f(x_{3i+2}) + f(x_{3(i+1)})\right] \\
&\approx \frac{3h}{8}\left[f(x_0) + 3\sum_{i=0}^{n-1} f(x_{3i+1}) + 3\sum_{i=0}^{n-1} f(x_{3i+2}) + 2\sum_{i=1}^{n-1} f(x_{3i}) + f(x_{3n})\right].
\end{aligned} \tag{3.29}$$

The total error in composite Simpson's 3/8 rule will be bounded as follows

$$\begin{aligned}
&\left|\int_a^b [f(x) - P(x)]\,dx\right| = \left|\sum_{i=0}^{n-1}\int_{x_{3i}}^{x_{3(i+1)}} [f(x) - P(x)]\,dx\right| \leq \sum_{i=0}^{n-1}\left|\int_{x_{3i}}^{x_{3(i+1)}} [f(x) - P(x)]\,dx\right| \\
&\leq \sum_{i=0}^{n-1}\frac{49M_4h^5}{7200} = n\frac{49M_4h^5}{720} = \frac{49M_4}{720\times 3^5}n^4(b-a)^5.
\end{aligned} \tag{3.30}$$

4 Solution of a system of linear equations

Here we aim to solve the following system of linear equations.

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & a_{12}x_2 & \dots & \dots & \dots & a_{1i}x_i & \dots & \dots & \dots & a_{11}x_n & = & b_1 \\
a_{21}x_1 & a_{22}x_2 & \dots & \dots & \dots & a_{2i}x_i & \dots & \dots & \dots & a_{21}x_n & = & b_2 \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a_{i1}x_1 & a_{i2}x_2 & \dots & \dots & \dots & a_{ii}x_i & \dots & \dots & \dots & a_{in}x_n & = & b_i \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
\dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\
a_{n1}x_1 & a_{n2}x_2 & \dots & \dots & \dots & a_{ni}x_i & \dots & \dots & \dots & a_{nn}x_n & = & b_n
\end{array} \tag{4.1}$$

In this system of linear equations $a'_{ij}s, i, j = 1, 2, \dots, n$ are known coefficients, $b'_i s, i = 1, 2, \dots, n$ are known values and $x'_i s, i = 1, 2, \dots, n$ are unknowns to be determined. This system can also be written in matrix form as

$$Ax = b, \tag{4.2}$$

where A is $n \times n$ matrix and x and b are column vectors of order $n \times 1$ of unknowns and known values respectively. If $b = 0$, then (4.2) is called homogeneous system of equations. For the solution of the system of linear equations (4.2) we recall an important theorem from linear algebra.

Theorem 4.1. *If A is real matrix of order $n \times n$, then the following statements are equivalent.*

- *$Ax = 0$ has only trivial solution.*
- *For each b, $Ax = b$ has a solution.*
- *A is invertible.*
- $\det(A) \neq 0$.

Remark 4.1. If any of the four equivalent conditions are satisfied in the above theorem, then one can find the solution of the system of linear equations by multiplying the inverse of the matrix A on left of both side of the equation $Ax = b$ to get

$$x = A^{-1}b. \tag{4.3}$$

And to find A^{-1} we know the standard method (learned in $10 + 2$ standard) of finding the adjoint of the matrix A and $A^{-1} = \text{Adj}(A)/\det(A)$. One can also recall the Cramer's method of finding the solution.

Observation 4.1. Both the above method of finding the solution involve an intermediate step of finding the determinant of matrix A. But finding the determinant of a large matrix is not an easy task.

Question 4.1. Can we find the solution of the system of linear equations without finding the inverse of the coefficient matrix?

4.1. Direct method for some special form of the coefficient matrix A

Here we will try to answer the question (4.1) positively if the coefficient matrix is of some easy form such that the solution can be obtained by direct computations.

1. **Diagonal case:** $A = D$
Since matrix A is assumed to be nonsingular, $\det(A) = \prod_{i=1}^{n} a_{ii} \neq 0$. And we have

$$\begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 & \dots & \dots & 0 & 0 \\ 0 & a_{22} & \dots & \dots & 0 & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & a_{ii} & \dots & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \dots & 0 & \dots & \dots & a_{n-1,n-1} & 0 \\ 0 & 0 & \dots & \dots & 0 & \dots & \dots & 0 & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_i \\ \dots \\ \dots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ \dots \\ b_i \\ \dots \\ \dots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

The solution is obvious in this case and can be written as

$$x_{ii} = \frac{b_i}{a_{ii}}, \qquad i = 1, 2 \dots, n. \tag{4.4}$$

Note that the only n divisions are required as computer operations.

2. **Lower triangular case:** $A = LT$
We have the following matrix equation.

$$\begin{bmatrix} a_{11} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & 0 & 0 & 0 & 0 & 0 \\ a_{i1} & a_{i2} & \dots & \dots & a_{ii} & 0 & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 & 0 \\ a_{n-1,1} & a_{n-1,2} & \dots & \dots & a_{n-1,i} & \dots & \dots & a_{n-1,n-1} & 0 \\ a_{n1} & a_{n2} & \dots & \dots & a_{ni} & \dots & \dots & a_{n,n-1} & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_i \\ \dots \\ \dots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ \dots \\ b_i \\ \dots \\ \dots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

In this case also we assume the solution to exist and hence, $\det(A) = \prod_{i=1}^{n} a_{ii} \neq 0$. Here it is easy to compute x_1 but to compute x_2, we need to substitute the value of x_1 in the second equation. Similarly to compute x_k we need to substitute the values of $x_1, x_2, \dots, x_{k-1}$ (which are already obtained) in the following equation

$$x_k = \frac{(b_k - \sum_{j=1}^{k-1} a_{kj} x_j)}{a_{kk}}, \qquad k = 2, \dots, n. \tag{4.5}$$

Thus we need forward substitution of the entries so we call this method as forward substitution. For computation of x_k in the above equation we require $(k-1)$ multiplications, $(k-1)$ additions and one division. Thus the total number of computer operations are $\sum_{k=1}^{n}(k-1) + (k-1) + 1 = n^2 + 2n$.

3. **Upper triangular case:** $A = UT$
In this case we also have $a_{ii} \neq 0$. And we have

$$\begin{bmatrix} a_{11} & a_{12} & \dots & \dots & a_{1i} & \dots & \dots & a_{1,n-1} & a_{1n} \\ 0 & a_{22} & \dots & \dots & a_{2i} & \dots & \dots & a_{2,n-1} & a_{2n} \\ 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & a_{ii} & \dots & \dots & a_{i,n-1} & a_{in} \\ 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{nn} \end{bmatrix} \times \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ \dots \\ x_i \\ \dots \\ \dots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \dots \\ \dots \\ b_i \\ \dots \\ \dots \\ b_{n-1} \\ b_n \end{bmatrix}.$$

But here the computation of x_n is easy compared to other unknowns. So we first compute x_n from the last equation and substitute its value in the second last equation to compute x_{n-1}. And to compute x_k we substitute the values of unknowns $x_n, \ldots, x_{k+1}$ obtained from the last $n-k$ equations in the k_{th} equation as follows

$$x_k = \frac{(b_k - \sum_{j=k+1}^{n} a_{kj}x_j)}{a_{kk}}, \qquad k = 2, \ldots, n-1. \tag{4.6}$$

Because of substitution in the earlier equations this method is known as back substitution. And similar to forward substitution we need total $n^2 + 2n$ computer operations for the complete solution.

Remark 4.2. Above methods are applicable only to diagonal or triangular matrices and do not give the general answer to the question (4.1).

4.2. Certain Decomposition Methods for solving the system of linear equations

Here our aim is to write the coefficient matrix A as the product of two matrices B and C of some simpler form. And then go for solving $Ax = b$ or $BCx = b$, by solving two different systems. First $Bz = b$ for z and then $Cx = z$ for x.

4.3. Doolittle's Method

In this method A is decomposed as $A = LU$, where

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}.$$

So that

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = LU = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{bmatrix} \tag{4.7}$$

Now one has to solve nine equations to find the all the total nine unknown coefficients of lower triangular matrix L and upper triangular matrix U. But these are easy to solve more or less only substitutions are needed. The solution is obtained by first solving $Lz = b$ for z by direct methods and then solving $Ux = z$ for x again by direct methods.

4.4. Crout's Method

Here one decomposes $A = LU$, where

$$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}, \qquad U = \begin{bmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{bmatrix}.$$

Thus one has

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = LU = \begin{bmatrix} l_{11} & l_{11}u_{12} & l_{11}u_{13} \\ l_{21} & l_{21}u_{12} + l_{22} & l_{21}u_{13} + l_{22}u_{23} \\ l_{31} & l_{31}u_{12} + l_{32} & l_{31}u_{13} + l_{32}u_{23} + l_{33} \end{bmatrix} \tag{4.8}$$

Here we again need to solve nine equations to determine all the nine unknown coefficients. And similar to previous method we first solve $Lz = b$ for z and then $Ux = z$ for x.

4.5. Positive definite matrix

A real square matrix A is said to be positive definite if $\det A > 0$ and all leading principal minors are positive.

4.6. A matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

is positive definite if

- $a_{11} > 0$,
- $\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} > 0$.
- $\det A > 0$.

Example 4.1. The matrix $\begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}$ is positive definite, while the matrix $\begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 2 \\ 3 & 2 & 5 \end{bmatrix}$ is not positive definite because second leading principal minor is not positive.

Cholesky's Method

Cholesky's method is applicable for symmetric and positive definite matrix A. In this case the decomposition of A is $A = LL^T$, where

$$L = \begin{bmatrix} d_1 & 0 & 0 \\ l_{21} & d_2 & 0 \\ l_{31} & l_{32} & d_3 \end{bmatrix}.$$

So that

$$A = \begin{bmatrix} a_{11} & a_{21} & a_{31} \\ a_{21} & a_{22} & a_{32} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = LU = \begin{bmatrix} d_1^2 & d_1 l_{12} & d_1 l_{31} \\ d_1 l_{21} & l_{21}^2 + d_2^2 & l_{21} l_{31} + d_2 l_{32} \\ d_1 l_{31} & l_{31} l_{21} + l_{32} d_2 & l_{31}^2 + l_{32}^2 + d_3^2 \end{bmatrix} \tag{4.9}$$

Here we only need to solve six equations in six unknowns. To solve $Ax = b$ we first solve $Lz = b$, for z and then $L^T x = z$ for x.

Question 4.2. Can we think of some generalization of the elimination method, which we learned in 10^{th} standard?

4.7. Gauss elimination method

Here our aim is to convert the given system of linear equations $Ax = b$ in another equivalent (having the same solution) system of linear equations in which the coefficient matrix is of triangular or diagonal form. So that we can use the direct methods to solve it. Gauss elimination method involves three types of elementary row transformations, which does not change the solution of the system. These transformations can be described as follows.

1. **Multiplication of a row by a non zero constant.**
   Suppose i_{th} row is multiplied by a non-zero constant c. We denote this transformation by $R_i(c)$. If we multiply the i_{th} row of the identity matrix by the same constant c and name this new matrix by $E_{i(c)}$, then it is easy to observe that the inverse of this matrix is $E_{i(c^{-1})}$. Now it is important to observe that this row transformation can also be operated to the system of linear equations by multiplying the system $Ax = b$ by elementary matrix $E_{i(c)}$ on the left to obtain $E_{i(c)}Ax = E_{i(c)}b$. And clearly this new system also have the solution as $x = A^{-1}(E_{i(c)})^{-1}E_{i(c)}b = A^{-1}b$.

2. **Interchanging two rows.**
   The row transformation of interchanging the i_{th} row with the j_{th} row is denoted by $R_i \leftrightarrow R_j$. If we interchange the i_{th} row with the j_{th} row of the identity matrix and obtain a new matrix denoted by $E_{i\leftrightarrow j}$. This is a self inverse matrix. And the same row transformation can also be obtained by left multiplication of the elementary matrix $E_{i\leftrightarrow j}$ on both sides of the system $Ax = b$.

3. **Addition of non-zero multiple of one row to other.**
   Suppose we multiply a constant c to i_{th} row and add it to j_{th} row. This row transformation is denoted by $R_{ji(c)}$. The corresponding elementary matrix obtained from identity by the same row transformation is denoted by $E_{ji(c)}$. The inverse of this elementary matrix is $E_{ji(-c)}$.

Thus we note that at each step of elementary row transformation to a system of linear equations we get an equivalent system of linear equations having the same solution.

Steps of Gauss elimination method to convert the coefficient matrix in to upper triangular (or identity) matrix can be described as follows.

- If $a_{11} = 0$, interchange the first row with j_{th} row for which $a_{j1} \neq 0$.
- Use transformations $R_{i1(-a_{i1}/a_{11})}$ to eliminate all $a_{i1}, i > 1$ with the pivot element a_{11}.
- If $a_{22} = 0$, interchange the second row with $j_{th}, j > 2$ row for which $a_{j2} \neq 0$
- Use transformations $R_{i2(-a_{i2}/a_{22})}$ to eliminate all $a_{i2}, i > 2$ with the pivot element a_{22}.
- Use similar transformations to convert the coefficient matrix in to upper triangular matrix.

If we know already that the matrix A is invertible we can further use elementary row transformations to convert this upper triangular matrix to identity matrix in following steps. Also note that in this upper triangular matrix each $a_{ii} \neq 0$.

- Now by first using $R_{n(1/a_{nn})}$, we get $a_{nn} = 1$.
- Use transformations $R_{in(-a_{in})}, i = n-1, \ldots, 1$ to eliminate $a_{in}, i = n-1, \ldots, 1$ with pivot as $a_{nn} = 1$.
- We further do similar transformations in $(n-1)_{th}$ column after that in $(n-2)_{th}$ column and so on to second and then in first column. In general in k_{th} column we first apply $R_{k(1/a_{kk})}$ to get $a_{kk} = 1$ and then $R_{ik(-a_{ik})}, i = k-1, \ldots, 1$ to eliminate $a_{ik}, i = k-1, \ldots, 1$ with pivot as $a_{kk} = 1$.

To implement Gauss elimination method in general we operate a number of elementary row transformations to a system of linear equations, required to convert A in to triangular form (or identity form), or equivalently the same elementary row transformations on the augmented matrix $[A|b]$ to convert it into $[T|b_T]$ (or $[I|b_I]$). Theoretically it can be understood by successive left multiplication of corresponding elementary metrics on both sides of matrix form of the system $Ax = b$, that is, $E_l E_{l-1} \ldots E_2 E_1 Ax = E_l E_{l-1} \ldots E_2 E_1 b$ to obtain the new equivalent matrix form of the system as $Tx = b_T$ (or $Ix = b_I$). Since the product of invertible metrics is an invertible matrix, the matrix $E = E_l E_{l-1} \ldots E_2 E_1 Ax = E_l E_{l-1} \ldots E_2 E_1$ is invertible and hence the solution x for $EAx = Tx = b_T = Eb$ (or $EAx = Ix = b_I = Eb$) is same as the solution for $E^{-1}EAx = Ax = b = E^{-1}Eb$.

Observation 4.2. It might happen in general that the coefficient matrix is not invertible or it is not a square matrix, that is, the number of equations is different from number of unknowns. In these situations we can not use Theorem 4.1 and the steps of Gauss elimination method might not work fine.

Question 4.3. Can we convert $Ax = b$ in to some form, which can be solved directly when A is rectangular or singular matrix.

4.8. Row Echelon Form

Yes! One can still use elementary row transformations to convert A in to Echelon form, which can be described as follows.

A matrix A is said to be of Echelon form if it satisfies the following properties.

- Each of the zero rows (a row in which each entry is zero), if it occurs in the matrix A, must occur below every non-zero row (a row in which at least one entry is non-zero).
- If the leading non-zero entry (first non-zero entry of the row) in i_{th} row occurs in k_{ith} column, that is, $a_{ij} = 0, j = 1, 2, \ldots, k_i - 1$ and $a_{ik_i} \neq 0$, and there are only r many non-zero rows, then one must have $k_1 < k_2 < \cdots < k_r$.

To obtain the row Echelon form of a matrix A of order $m \times n$, one can use the following steps.

- If first non-zero column is l_1, by a row transformation (row interchange) bring one non-zero entry of l_{1th} column in the first row. And use this entry to eliminate all other entries of the l_{1th} column by the row transformations $R_{il_1(-a_{il_1}/a_{1l_1})}$. This l_1 is k_1 of the definition. In further row transformations we will not use the first row at all.
- Now consider the sub matrix of order $m - 1 \times n$ after ignoring the first row. Search for the first non-zero column l_2 of this sub-matrix. If first entry of this l_{2th} column of order $m-1 \times 1$ is zero, bring one non-zero entry of the column to the first row and use this entry to eliminate all other non-zero entries of the column. This l_2 is k_2 of the definition. In further row transformations we will not use the first two rows of the parent matrix.
- Now we further consider the sub matrix of order $m-2 \times n$ of the parent matrix. And search for first non-zero column l_3 of this sub-matrix. Use similar row transformations to eliminate all the entries of l_{3th} below the first row using the non-zero entry of the first row as pivot.
- Use similar transformations successively to find a column l_r of the parent matrix satisfying
  i) l_{rth} column is the first non-zero column of the sub-matrix (obtained by ignoring first $r-1$ rows) of order $m - r + 1 \times n$,
  ii) only non-zero entry of this column of sub-matrix is in the first row.
- This process will stop if i) either $r = m$, in this case $m \leq n$ (and $m = n$ only when Echelon form is upper triangular and invertible),
  ii) or last $r - m$ rows of the parent matrix are zero rows.

4.9. Solution of system through Echelon form

For solving the system Ax=b, where A is coefficient matrix of order $m \times n$, x is column of unknowns of order $n \times 1$ and b is column of order $m \times 1$, we apply same sequence of elementary row transformations to augmented matrix $[A|b]$ to convert it in to $[A_E|b_E]$, where A_E is the Echelon form of matrix A. Because of elementary row transformations the solution x for both the systems $Ax = b$ and $A_Ex = b_E$ are same.

We can also conclude about the nature of the solution by some observations of augmented matrix $[A_E|b_E]$as follows.

- If $r = m$ and $n = m$, then unique solution by direct method.
- If $r = m$ and $n > m$, then infinite solutions with $n - m$ order of independency, that is, out of n unknowns the system can be solved for any chosen m unknowns in terms of remaining $n - m$ unknowns, which can be given any values.

- If $r < m$ and there exists some non-zero row in last $m - r$ rows of augmented matrix $[A_E|b_E]$. This happens because of only non-zero entry of the row as the entry corresponding to column b_E. If this entry is $b_{Er+p} \neq 0$, we end up with solving an equation of the form $0 \times x_{r+p} = b_{Er+p}$, which shows the inconsistency of the system and leads to **NO Solution**.
- Further if $r < m$ and last $m - r$ rows of augmented matrix $[A_E|b_E]$ are zero rows, then the system has unique solution if $n = r$, and infinite solutions with $n - r$ degree of freedom if $n > r$. Note that the case $n < r$ does not occur.

4.10. Partial Pivoting

Look at the class example for the need of partial pivoting.

One can ensure partial pivoting in Gauss Elimination or in Echelon form just by making sure that the pivot entry in the column is largest in magnitude compared to the entries, which are to be eliminated.

4.11. Norm

Norm is the notion of generalization of the modulus. Modulus of a number measures the distance of the position of that number from the origin. Now to measure the distances in a plane or in $\mathbb{R}^n$, one considers a function

$$\|\cdot\| : \mathbb{R}^n \to \mathbb{R}^+ \cup \{0\}, \tag{4.10}$$

satisfying,

- $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$, for all $x \in \mathbb{R}^n$.
- $\|\alpha x\| = |\alpha|\|x\|$, for any scalar $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$. This also implies $\|x - y\| = \|y - x\|$.
- **Triangle Inequality:** For any $x, y, z, u, v \in \mathbb{R}^n$,

  $$\|u + v\| \leq \|u\| + \|v\|. \tag{4.11}$$

  This is also equivalent to

  $$\|x - z\| \leq \|x - y\| + \|y - z\|. \tag{4.12}$$

  Equation (4.12) shows that distance of any two points is less or equal to sum of the distances of these two points from any third point.

Example 4.2. Consider the following function on $\mathbb{R}^n$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}. \tag{4.13}$$

This defines a norm on $\mathbb{R}^n$, known as $p-$norm of the vector $x \in \mathbb{R}^n$. Note that in daily life we use $2-$norm for measuring the distance in plane and space.

Example 4.3. Consider the function on $\mathbb{R}^n$

$$\|x\|_\infty = \max_{1 \leq i \leq \infty} \{|x_i|\}. \tag{4.14}$$

This also defines a norm on $\mathbb{R}^n$.

Exercise 4.1. Draw the locus of following in plane

- $\{x \in \mathbb{R}^2 : \|x\|_1 = 1\}$,

- $\{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$,
- $\{x \in \mathbb{R}^2 : \|x\|_\infty = 1\}$.

Exercise 4.2. Show that norm is a continuous function from $\mathbb{R}^n \to [0, \infty]$.

4.12. Matrix Norm

We know that the set of all real valued matrices of order $m \times n$, denoted by M_{mn}, forms a vector space over real numbers. One can define a norm on M_{mn}, which is also compatible with matrix multiplication as

$$\|\cdot\| : M_{mn} \to \mathbb{R}^+ \cup \{0\}, \tag{4.15}$$

satisfying,

- $\|A\| \geq 0$ and $\|A\| = 0$ if and only if A is null matrix, for all $A \in M_{mn}$.
- $\|\alpha A\| = |\alpha|\|A\|$, for any scalar $\alpha \in \mathbb{R}$ and $A \in M_{mn}$.
- **Triangle Inequality:** For any $A, B \in M_{mn}$,

$$\|A + B\| \leq \|A\| + \|B\|. \tag{4.16}$$

- This norm is also compatible with multiplication by a column vector of order $n \times 1$

$$\|Ax\| \leq \|A\|\|x\|, \tag{4.17}$$

Example 4.4. Consider the following function on M_{mn}

$$\|A\|_1 = \max_{1 \leq j \leq m} \left\{ \left(\sum_{i=1}^{n} |a_{ij}| \right) : j = 1, \ldots, m \right\}. \tag{4.18}$$

This defines a norm on M_{mn}, which is compatible with multiplication by vectors $x \in \mathbb{R}^n$ as

$$\|Ax\|_1 \leq \|A\|_1 \|x\|_1, \tag{4.19}$$

where $\|Ax\|_1$ and $\|x\|_1$ are defined in (4.13) for $p = 1$.

Example 4.5. Consider the following function on M_{mn}

$$\|A\|_2 = \left(\sum_{i,j=1}^{m,n} |a_{ij}|^2 \right)^{1/2}. \tag{4.20}$$

This also defines a norm on M_{mn}, which is compatible with multiplication by vectors $x \in \mathbb{R}^n$ as

$$\|Ax\|_2 \leq \|A\|_2 \|x\|_2, \tag{4.21}$$

where $\|Ax\|_2$ and $\|x\|_2$ are defined in (4.13) for $p = 2$.

Example 4.6. Consider the following function on M_{mn}

$$\|A\|_\infty = \max_{1 \leq i \leq n} \left\{ \left(\sum_{j=1}^{m} |a_{ij}| \right) : i = 1, \ldots, n \right\}. \tag{4.22}$$

This defines a norm on M_{mn}, which is compatible with multiplication by vectors $x \in \mathbb{R}^n$ as

$$\|Ax\|_\infty \leq \|A\|_\infty \|x\|_\infty, \tag{4.23}$$

where $\|Ax\|_\infty$ and $\|x\|_\infty$ are defined in (4.14).

4.13. Convergence of vectors

Similar to convergence of sequence of numbers to a point, one can define convergence of sequence of vectors to another vector. A sequence of vector (or matrices) converges to some vector (or matrices) if and only if sequence of components converges to the corresponding component of limit vector (or matrix) for each component.

In $\mathbb{R}^m$ a sequence of vectors $y^{(n)}$ converges to a vector y **if and only if** $y_i^{(n)} \to y_i$, for all $i = 1, \ldots, m$.

Example 4.7. The vector $[1/n, 2n/n-1, n+1/n-1]^t \to [0, 2, 1]^t$, while $[1/n, 2n/n-1, n]^t$ does not converse because the sequence corresponding to third component is not convergent.

4.14. Gauss-Jacobi iterative method

Now our aim is to solve (4.1) using some iteration method. For this we first assume $a_{ii} \neq 0$ and rewrite the system of linear equations as follows

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & = & 0x_1 & -a_{12}x_2 & \ldots & \ldots & \ldots & -a_{1i}x_i & \ldots & \ldots & \ldots & -a_{1n}x_n & +b_1 \\
a_{22}x_2 & = & -a_{21}x_1 & 0x_2 & \ldots & \ldots & \ldots & -a_{2i}x_i & \ldots & \ldots & \ldots & -a_{2n}x_n & +b_2 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
a_{ii}x_i & = & -a_{i1}x_1 & -a_{i2}x_2 & \ldots & \ldots & \ldots & 0x_i & \ldots & \ldots & \ldots & -a_{in}x_n & +b_i \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
a_{nn}x_n & = & -a_{n1}x_1 & -a_{n2}x_2 & \ldots & \ldots & \ldots & -a_{ni}x_i & \ldots & \ldots & \ldots & 0x_n & +b_n
\end{array}$$

Note that if we decompose $A = L + D + U$, where L, D, U are lower triangular, diagonal and upper triangular matrices respectively, then above system of equations in matrix form can be written as $Dx = -(L+U)x + b$, where x is the column vector of unknowns. Now since D is invertible by our assumption we have

$$x = -D^{-1}(L+U)x + D^{-1}b. \tag{4.24}$$

And this can also be written component wise as

$$\begin{array}{ccccccccccccc}
x_1 & = & 0x_1 & -\frac{a_{12}}{a_{11}}x_2 & \ldots & \ldots & \ldots & -\frac{a_{1i}}{a_{11}}x_i & \ldots & \ldots & \ldots & -\frac{a_{1n}}{a_{11}}x_n & +\frac{b_1}{a_{11}} \\
x_2 & = & -\frac{a_{21}}{a_{22}}x_1 & 0x_2 & \ldots & \ldots & \ldots & -\frac{a_{2i}}{a_{22}}x_i & \ldots & \ldots & \ldots & -\frac{a_{2n}}{a_{22}}x_n & +\frac{b_1}{a_{11}} \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
x_i & = & -\frac{a_{i1}}{a_{ii}}x_1 & -\frac{a_{i2}}{a_{ii}}x_2 & \ldots & \ldots & \ldots & 0x_i & \ldots & \ldots & \ldots & -\frac{a_{in}}{a_{ii}}x_n & +\frac{b_i}{a_{ii}} \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & = & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
x_n & = & -\frac{a_{n1}}{a_{nn}}x_1 & -\frac{a_{n2}}{a_{nn}}x_2 & \ldots & \ldots & \ldots & -\frac{a_{ni}}{a_{nn}}x_i & \ldots & \ldots & \ldots & 0x_n & +\frac{b_n}{a_{nn}}
\end{array} \tag{4.25}$$

Denoting $-D^{-1}(L+U) = B$ and $D^{-1}b = C$, we have $x = Bx + C$. Now if we assume the initial approximation to the solution as

$$x^{(0)} = \begin{bmatrix} x_1^{(0)} & x_2^{(0)} & \ldots & \ldots & \ldots & x_i^{(0)} & \ldots & \ldots & \ldots & x_n^{(0)} \end{bmatrix}^t, \tag{4.26}$$

then the first approximation can be obtained by the equation $x^{(1)} = Bx^{(0)} + C$, and in general $(k+1)_{th}$ approximation is obtained by

$$x^{(k+1)} = Bx^{(k)} + C, \tag{4.27}$$

where matrix B and C are as follows

$$B = \begin{bmatrix} 0 & -\frac{a_{12}}{a_{11}} & \dots & \dots & \dots & -\frac{a_{1i}}{a_{11}} & \dots & \dots & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & \dots & \dots & \dots & -\frac{a_{2i}}{a_{22}} & \dots & \dots & \dots & -\frac{a_{2n}}{a_{22}} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\frac{a_{i1}}{a_{ii}} & -\frac{a_{i2}}{a_{ii}} & \dots & \dots & \dots & 0 & \dots & \dots & \dots & -\frac{a_{in}}{a_{ii}} \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \end{bmatrix}, \qquad C = \begin{bmatrix} -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{2n}}{a_{22}} \\ \dots \\ \dots \\ \dots \\ -\frac{a_{in}}{a_{ii}} \\ \dots \\ \dots \\ \dots \\ 0 \end{bmatrix}. \tag{4.28}$$

Using (4.27) we can write i_{th} component of the $(k+1)_{th}$ approximation of the solution vector as

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} + \sum_{j=1, j\neq i}^{n} \frac{-a_{ij}}{a_{ii}} x_j^{(k)}. \tag{4.29}$$

4.15. Error analysis of Gauss-Jacobi Iteration Method

Suppose x is the exact solution to the system 4.27. Then the k_{th} error vector can be defined as

$$e^{(k)} = x - x^{(k)} = \begin{bmatrix} x_1 - x_1^{(k)} & x_2 - x_2^{(2)} & \dots & \dots & x_i - x_i^{(k)} \dots & \dots & x_n - x_n^{(k)} \end{bmatrix}^t, \tag{4.30}$$

where $[a_{ij}]^t$ represents the transpose of the corresponding matrix. Using (4.30) and (4.29), one has

$$e^{(k+1)} = x - x^{(k+1)} = (Bx + C) - (Bx^{(k)} + C) = B(x - k^k) = Be^{(k)}, \tag{4.31}$$

and component-wise

$$e_i^{(k+1)} = \sum_{j=1, j\neq i}^{n} \frac{-a_{ij}}{a_{ii}} e_j^{(k)}$$

or,

$$\begin{aligned} |e_i^{(k+1)}| &\leq \sum_{j=1, j\neq i}^{n} \left| \frac{-a_{ij}}{a_{ii}} \right| |e_j^{(k)}| \\ &\leq \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} \max_{1\leq j\leq n} \{|e_j^{(k)}| : j = 1, \dots, n\} = \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} \|e^{(k)}\|_\infty \end{aligned} \tag{4.32}$$

If we define

$$\alpha_i = \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|}, \qquad \beta_i = \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|}, \tag{4.33}$$

and

$$\mu = \max_{1\leq i\leq n} \{(\alpha_i + \beta_i) : i = 1, \dots, n\}. \tag{4.34}$$

Using (4.32) and (4.34),

$$\begin{aligned} |e_i^{(k+1)}| &\leq \|e^{(k)}\|_\infty \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|} \\ &\leq \|e^{(k)}\|_\infty \ (\alpha_i + \beta_i) \\ &\leq \|e^{(k)}\|_\infty \ \mu. \end{aligned}$$

Above inequality is true for all $i = 1, \ldots, n$, hence

$$\begin{aligned} \max_{1\leq i\leq n} \{|e_i^{(k+1)}| : i = 1, \ldots, n\} &\leq \|e^{(k)}\|_\infty \ \mu, \\ \|e^{(k+1)}\|_\infty &\leq \mu \ \|e^{(k)}\|_\infty, \end{aligned} \tag{4.35}$$

Above inequality (4.35) is true for all $k \in \mathbb{N}$. And hence by repeated application of (4.35), we have

$$\|e^{(k+1)}\|_\infty \leq \mu \|e^{(k)}\|_\infty \leq \mu\mu \|e^{(k-1)}\|_\infty \leq (\mu)^{k+1} \|e^{(0)}\|_\infty, \tag{4.36}$$

where $e^{(0)} = x - x^{(0)}$ is the zeroth error vector.

If we assume $\mu < 1$, then $\|e^{(k+1)}\|_\infty \to 0$, that is, $|x_i - x_i^{(k+1)}| = |e_i^{(k+1)}| \to 0$ for all $i = 1, \ldots, n$, which implies that the sequence of approximating vectors $x^{(k)}$ converges to the exact solution x. This μ defined by (4.34) is known as convergence factor of Gauss-Jacobi iterative method.

Remark 4.3. Now we collect all the assumptions, which are assumed for the convergence. These assumptions are

- $a_{ii} \neq 0$, for all $i = 1, \ldots, n$,
- and

$$\begin{aligned} \mu &= \max_{1\leq i\leq n} \{(\alpha_i + \beta_i) : i = 1, \ldots, n\} \\ &= \max_{1\leq i\leq n} \left\{ \left(\sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) : i = 1, \ldots, n \right\} < 1. \end{aligned} \tag{4.37}$$

Let us first consider the second condition (4.37), which will be valid if and only if

$$\begin{aligned} & (\alpha_i + \beta_i) < 1 \quad \text{for all} \quad i = 1, \ldots, n \\ \text{or} \quad & \left(\sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) < 1 \quad \text{for all} \quad i = 1, \ldots, n \\ \text{or} \quad & \left(\sum_{j=1, j\neq i}^{n} |a_{ij}| \right) < |a_{ii}| \quad \text{for all} \quad i = 1, \ldots, n \end{aligned} \tag{4.38}$$

The condition (4.38) is known as strict row diagonally dominant and also implies the first condition. Thus for the convergence of Gauss-Jacobi iteration method we only need the coefficient matrix to be strict row diagonally dominant.

Remark 4.4. Note that $\|B\|_\infty = \mu$. Then from (4.31)

$$\|e^{(k+1)}\| = \|Be^{(k)}\| \leq \|B\| \times \|e^{(k)}\| \leq \|B\| \times \|B\| \times \|e^{(k-1)}\| \leq \ldots \leq \|B\|^{k+1} \|e^{(0)}\| = \mu^{k+1} \|e^{(0)}\|.$$

Now the convergence follows if $\mu < 1$. Moreover, we have $x^{(k+1)} - x^{(k)} = Bx^{(k)} + C - Bx^{(k-1)} - C$ and hence

$$\|x^{(k+1)} - x^{(k)}\| = \|B(x^{(k)} - x^{(k-1)})\| \leq \|B\| \times \|x^{(k)} - x^{(k-1)}\| = \mu \|x^{(k)} - x^{(k-1)}\|, \qquad \text{for all } k \in \mathbb{N}.$$

Thus $\|x^{(k+1)} - x^{(k)}\| \leq \mu\|x^{(k)} - x^{(k-1)}\| \leq \mu \times \mu\|x^{(k-1)} - x^{(k-2)}\| \leq \ldots \leq \mu^k\|x^{(1)} - x^{(0)}\|$. Now for any $m > k > 1$ we have

$$\begin{aligned}
\|x^{(m)} - x^{(k)}\| &= \|x^{(m)} - x^{(m-1)} + x^{(m-1)} - x^{(m-2)} + x^{(m-2)} \ldots - x^{(k+1)} + x^{(k+1)} - x^{(k)}\| \\
&\leq \|x^{(m)} - x^{(m-1)}\| + \|x^{(m-1)} - x^{(m-2)}\| + \ldots + \|x^{(k+1)} - x^{(k)}\| \\
&\leq (\mu^{m-1} + \mu^{m-2} + \ldots + \mu^k)\|x^{(1)} - x^{(0)}\| \\
&< \mu^k(\sum_{i=0}^{\infty}\mu^i)\|x^{(1)} - x^{(0)}\| = \frac{\mu^k}{1-\mu}\|x^{(1)} - x^{(0)}\| \qquad \text{since } \mu < 1.
\end{aligned}$$

Further since $\|x^{(m)} - x\| \to 0$, for any given ϵ, there exists an integer $m_0 > k$ such that $\|x^{(m_0)} - x\| < \epsilon$. And hence

$$\|x - x^{(k)}\| \leq \|x - x^{(m_0)}\| + \|x^{(m_0)} - x^{(k)}\| < \epsilon + \frac{\mu^k}{1-\mu}\|x^{(1)} - x^{(0)}\|.$$

Since above inequality is true for any arbitrary $\epsilon > 0$, we can assume $\epsilon \to 0$ and hence

$$\|e^{(k)}\| = \|x - x^{(k)}\| \leq \frac{\mu^k}{1-\mu}\|x^{(1)} - x^{(0)}\|. \tag{4.39}$$

4.16. Gauss-Seidel iterative method

We rewrite the system of equations as follows

$$\begin{array}{ccccccccccccc}
a_{11}x_1 & = & -a_{12}x_2 & -a_{13}x_3 & \ldots & \ldots & -a_{1i}x_i & \ldots & \ldots & \ldots & \ldots & -a_{1n}x_n & +b_1 \\
a_{21}x_1 & +a_{22}x_2 & = & -a_{23}x_3 & \ldots & \ldots & -a_{2i}x_i & \ldots & \ldots & \ldots & \ldots & -a_{2n}x_n & +b_2 \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
a_{i1}x_1 & +a_{i2}x_2 & \ldots & \ldots & \ldots & +a_{ii}x_i & = & -a_{i,i+1}x_{i+1} & \ldots & \ldots & \ldots & -a_{in}x_n & +b_i \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\
a_{n1}x_1 & +a_{n2}x_2 & \ldots & \ldots & \ldots & +a_{ni}x_i & \ldots & \ldots & \ldots & \ldots & +a_{nn}x_n & = & +b_n
\end{array}$$

As in case of Gauss-Jacobi method we assume $a_{ii} \neq 0$ and decompose $A = L + D + U$, then above system of equations in matrix form can be written as $(L + D)x = -Ux + b$ and the matrix $L + D$ turns out to be invertible. Further if $x^{(0)}$ is the initial approximation, we define the sequence of iterates by

$$(L + D)x^{(k+1)} = -Ux^{(k)} + b, \qquad \text{or} \qquad x^{(k+1)} = -(L + D)^{-1}Ux^{(k)} + (L + D)^{-1}b. \tag{4.40}$$

Component-wise it can be written as follows

$$\sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} + a_{ii}x_i^{(k+1)} = -\sum_{j=i+1}^{n} a_{ij}x_j^{(k)} + b_i, \qquad \text{or equivalently}$$

$$x_i^{(k+1)} = \frac{b_i}{a_{ii}} + \sum_{j=1}^{i-1}\frac{-a_{ij}}{a_{ii}}x_j^{(k+1)} + \sum_{j=i+1}^{n}\frac{-a_{ij}}{a_{ii}}x_j^{(k)}. \tag{4.41}$$

4.17. Error analysis of Gauss-Seidel Iteration Method

Suppose x is the exact solution to the system 4.25. Then the k_{th} error vector can be similar to (4.30) defined as

$$e^{(k)} = x - x^{(k)} = \left[x_1 - x_1^{(k)} \quad x_2 - x_2^{(2)} \quad \dots \quad \dots \quad x_i - x_i^{(k)} \dots \quad \dots \quad x_n - x_n^{(k)}\right]^t,$$

and using (4.41) component-wise we can write

$$\begin{aligned} e_i^{(k+1)} &= x_i - x_i^{(k+1)} = \left(\frac{b_i}{a_{ii}} + \sum_{j=1}^{i-1} \frac{-a_{ij}}{a_{ii}} x_j \sum_{j=i+1}^{n} \frac{-a_{ij}}{a_{ii}} x_j\right) - \left(\frac{b_i}{a_{ii}} + \sum_{j=1}^{i-1} \frac{-a_{ij}}{a_{ii}} x_j^{(k+1)} + \sum_{j=i+1}^{n} \frac{-a_{ij}}{a_{ii}} x_j^{(k)}\right), \\ &= \sum_{j=1}^{i-1} \frac{-a_{ij}}{a_{ii}} (x_j - x_j^{(k+1)}) + \sum_{j=i+1}^{n} \frac{-a_{ij}}{a_{ii}} (x_j - x_j^{(k)}) \\ &= \sum_{j=1}^{i-1} \frac{-a_{ij}}{a_{ii}} (e_j^{(k+1)}) + \sum_{j=i+1}^{n} \frac{-a_{ij}}{a_{ii}} (e_j^{(k)}), \end{aligned} \tag{4.42}$$

Above equation implies

$$\begin{aligned} |e_i^{(k+1)}| &= \left|\sum_{j=1}^{i-1} \frac{-a_{ij}}{a_{ii}} (e_j^{(k+1)}) + \sum_{j=i+1}^{n} \frac{-a_{ij}}{a_{ii}} (e_j^{(k)})\right| \\ &\leq \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} |e_j^{(k+1)}| + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} |e_j^{(k)}| \\ &\leq \sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} \|e^{(k+1)}\|_\infty + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \|e^{(k)}\|_\infty \\ &= \alpha_i \|e^{(k+1)}\|_\infty + \beta_i \|e^{(k)}\|_\infty. \end{aligned}$$

Or,

$$\begin{aligned} |e_i^{(k+1)}| - \alpha_i \|e^{(k+1)}\|_\infty &\leq \beta_i \|e^{(k)}\|_\infty \\ |e_i^{(k+1)}| - \|e^{(k+1)}\|_\infty + (1-\alpha_i)\|e^{(k+1)}\|_\infty &\leq \beta_i \|e^{(k)}\|_\infty \end{aligned} \tag{4.43}$$

If we assume $(1-\alpha_i) > 0$ for all $i = 1, \dots, n$ and define

$$\eta = \max_{1 \leq i \leq n} \left\{ \frac{\beta_i}{(1-\alpha_i)} : i = 1, \dots, n \right\}. \tag{4.44}$$

Then from (4.43),

$$\frac{|e_i^{(k+1)}| - \|e^{(k+1)}\|_\infty}{(1-\alpha_i)} + \|e^{(k+1)}\|_\infty \leq \frac{\beta_i}{(1-\alpha_i)} \|e^{(k)}\|_\infty \leq \eta \|e^{(k)}\|_\infty \tag{4.45}$$

Let $\|e^{(k+1)}\|_\infty = |e_{i_0}^{(k+1)}|$. Since $(1-\alpha_i) > 0$, the expression $\frac{|e_i^{(k+1)}| - \|e^{(k+1)}\|_\infty}{(1-\alpha_i)}$ is less or equal to zero for all $i = 1, \dots, n$ and equal to zero for at least one $i = i_0$. Further since the above inequality (4.45) is true for all $i = 1, \dots, n$ and so it is true for $i = i_0$. Thus from (4.45) for $i = i_0$

$$\|e^{(k+1)}\|_\infty \leq \eta \|e^{(k)}\|_\infty. \tag{4.46}$$

Above inequality is true for all $k \in \mathbb{N}$. And hence we have from (4.46)

$$\|e^{(k+1)}\|_\infty \leq \eta \|e^{(k)}\|_\infty \leq \eta\eta \|e^{(k-1)}\|_\infty \leq (\eta)^{k+1} \|e^{(0)}\|_\infty. \tag{4.47}$$

Further if we assume $\eta < 1$, we conclude from (4.47) that $\|e^{(k)}\|_\infty \to 0$ or equivalently $x^{(k)}$ converges to exact solution x. This η defined by (4.44) is known as convergence factor of Gauss-Seidel iterative method.

Remark 4.5. Now we collect all the assumptions, which are assumed for the convergence of Gauss-Seidel method. These assumptions are

- $a_{ii} \neq 0$, for all $i = 1, \ldots, n$,
- $(1 - \alpha_i) > 0$ for all $i = 1, \ldots, n$,
- and

$$\eta = \max_{1 \leq i \leq n} \left\{ \frac{\beta_i}{(1 - \alpha_i)} : i = 1, \ldots, n \right\} < 1. \tag{4.48}$$

Let us first consider the third condition (4.48), which will be valid if and only if

$$\begin{aligned} & \left(\frac{\beta_i}{(1 - \alpha_i)} \right) < 1 \quad \text{for all} \quad i = 1, \ldots, n \\ \text{or} \quad & \beta_i < (1 - \alpha_i) \quad \text{for all} \quad i = 1, \ldots, n \\ \text{or} \quad & \left(\sum_{j=1}^{i-1} \frac{|a_{ij}|}{|a_{ii}|} + \sum_{j=i+1}^{n} \frac{|a_{ij}|}{|a_{ii}|} \right) < 1 \quad \text{for all} \quad i = 1, \ldots, n \\ \text{or} \quad & \left(\sum_{j=1, j \neq i}^{n} |a_{ij}| \right) < |a_{ii}| \quad \text{for all} \quad i = 1, \ldots, n \end{aligned} \tag{4.49}$$

Thus if we assume the coefficient matrix to be strict row diagonally dominant, the third assumption will be satisfied. Further in this case $1 - \alpha_i > \beta_i \geq 0$ implies the second condition and first is obviously true.

Exercise 4.3. If $\mu < 1$, prove that $\eta \leq \mu$.

Remark 4.6. Above exercise shows that if the coefficient matrix is strict row diagonally dominant, then the convergence factor of Gauss-Jacobi method is less or equal to convergence factor of Gauss-Seidel method. And in case if $\eta < \mu$, Gauss-Seidel method should converges faster than Gauss-Jacobi method.

Remark 4.7. To apply any of the two methods, we need the matrix of coefficients as the strict row diagonally dominant. Then for any initial approximation sequence of iterates converges to some exact solution. Now the question arise that, whether the sequences corresponding to two different initial approximations can converge to two different exact solutions. Note that it can happen only when the coefficient matrix, which is strictly row diagonally dominant, is singular. In the following discussion we will see that any strictly row diagonally dominant matrix is non singular and hence we conclude that any sequence of iterates with a given initial approximation converges to the exact solution.

Theorem 4.2. *Suppose A and B are two square matrices of order $n \times n$. If A is invertible and*

$$\|A - B\| < \frac{1}{\|A^{-1}\|},$$

then B is also invertible.

Proof. Suppose if B is not invertible, then the rank of B is less than n and hence by Rank-Nullity theorem the null space of B is not equal to $\{0\}$. Thus if $0 \neq x$ is in the Null space of (B), then $Bx = 0$ and

$$\frac{\|x\|}{\|A^{-1}\|} = \frac{\|A^{-1}Ax\|}{\|A^{-1}\|} \leq \|Ax\| = \|Ax - Bx\| \leq \|A - B\| \times \|x\|.$$

But $\|x\| \neq 0$, we can conclude the contradiction $\frac{1}{\|A^{-1}\|} \leq \|A - B\|$ (to the assumption). □

Problem 4.1. Show that a strictly row diagonally dominant matrix is invertible.

Solution. Let A be a strictly row diagonally dominant matrix. If $A = L+D+U$ is the decomposition of A, then the diagonal matrix D consists of non zero entries in the diagonal and hence invertible. Since $A = D \times D^{-1}A$, it is sufficient to prove that $D^{-1}A$ is invertible. Note that I is invertible matrix with $\|I^{-1}\|_\infty = 1$. Thus if we can show that $\|I - D^{-1}A\|_\infty < 1$, then by the application of the previous theorem it follows that $D^{-1}A$ is invertible. But it is easy to see that $\|I - D^{-1}A\|_\infty = \mu < 1$.

4.18. Ill-conditioned matrix

We first solve the following system of two linear equations in two unknowns.

$$\begin{aligned} x_1 + 3x_2 &= 19 \\ 2.5x_1 + 7.857x_2 &= 47.499 \end{aligned}$$

The exact solution to this system is $x_1 = -3$ and $x_2 = 7$. But if we round off 47.499 by 47.500, then the solution changes drastically to $x_1 = 19$ and $x_2 = 0$. Or if we round off 7.857 by 7.86, the solution changes to $x_1 = 98.17$ and $x_2 = -26.39$.

We observe that a small change in the coefficient matrix A or the constant vector b leads to a large change in the solution vector. Such system is called ill-conditioned, otherwise the system is called well-conditioned.

4.19. Small change in b vector

We want to solve

$$Ax = b \tag{4.50}$$

Suppose the change δb in b leads to change δx in solution vector so that $A(x + \delta x) = Ax + A\delta x = b + \delta b$. This implies $A\delta x = \delta b$, or $\delta x = A^{-1}\delta b$. Thus

$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\| \tag{4.51}$$

Above equation shows that the δx is controlled if $\|A^{-1}\|$ is controlled. Further since

$$\|b\| = \|Ax\| \leq \|A\|\|x\|, \tag{4.52}$$

we can control the relative error in solution vector $\|\delta x\|/\|x\|$. Using (4.51),

$$\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\| \times \|\delta b\| \frac{1}{\|x\|} \leq \|A^{-1}\| \times \|\delta b\| \frac{\|A\|}{\|b\|} = (\|A^{-1}\| \times \|A\|) \frac{\|\delta b\|}{\|b\|}. \tag{4.53}$$

Thus we see that the relative error in x is controlled by relative error in b if one has control over the quantity $(\|A^{-1}\|\|A\|)$, which is known as condition number of matrix A.

Exercise 4.4. Show that for any invertible matrix A the condition number is always greater or equal to 1, when the considered norm is infinity norm.

5 The Eigen-Value Problem

Let B be a square matrix. A number λ (real or complex) is said to be an eigenvalue of the matrix B if there exists a nonzero vector y such that

$$By = \lambda y.$$

This is equivalent to say that $(B - \lambda I)y = 0$ for some non-zero vector y,
$\Leftrightarrow$ the null space of $(B - \lambda I)$ is not equal to $\{0\}$,
$\Leftrightarrow$ the dimension of the null space of $(B - \lambda I)$ is greater than or equal to 1,
$\Leftrightarrow$ $(B - \lambda I)$ is singular,
$\Leftrightarrow$ determinant of $(B - \lambda I)$ is zero.

Clearly if y is an eigenvector for the matrix B corresponding to eigenvalue λ, then so is αy for any nonzero α

Note that any matrix B of order $m \times n$ represents a linear transformation from an n-dimensional vector space to an m-dimensional vector space with respect to some fixed ordered basis of these vector spaces. it can be proved that the linear map defined by the matrix B as discussed is a continuous map from n-dimensional vector space to an m-dimensional vector space.

Moreover, any square matrix of order $n \times n$ can be viewed as a linear transformation from $\mathbb{R}^n$ to $\mathbb{R}^n$ with respect to the standard basis.

The importance of eigenvector is that the image of an eigenvector of an square matrix is again a vector in the same direction but scaled by a factor of its eigenvalue.

Moreover, if there is a basis of eigenvectors say $\{v_1, v_2, \ldots, v_n\}$, with corresponding eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, then image of any vector is completely known. Suppose if $z \in \mathbb{R}^n$, then there are scalars $\alpha_1, \alpha_2, \ldots, \alpha_n$ such that

$$z = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_n v_n, \tag{5.1}$$

so that

$$Bz = \alpha_1 Bv_1 + \alpha_2 Bv_2 + \ldots + \alpha_n Bv_n = \alpha_1 \lambda_1 v_1 + \alpha_2 \lambda_2 v_2 + \ldots + \alpha_n \lambda_n v_n.$$

In this case the matrix of linear transformation with respect to the basis $\{v_1, v_2, \ldots, v_n\}$ turns out to be diagonal with diagonal entries as $\lambda_1, \lambda_2, \ldots, \lambda_n$ and we say that the matrix B is diagonalizable.

Exercise 5.1. Show that if a matrix B is diagonalizable, with eigenvalues $|\lambda_1| \geq |\lambda_2| \geq \ldots \geq |\lambda_n|$, then the image of the unit ball $\{x : \|x\| \leq 1\}$ is contained in a ball of radius $|\lambda_1|$ with center at origin.

This exercise shows the importance of eigenvalue of largest magnitude.

5.1. The Power Method

This method is useful to find the dominant eigenvalue among a collection of eigenvalues of a matrix and an eigenvector corresponding to the dominant eigenvalue. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be a set of eigenvalues of an square matrix of order $n \times n$, with corresponding eigenvectors $v_1, v_2, \ldots, v_m$, $m \leq n$ such that $z = \alpha_1 v_1 + \alpha_2 v_2 + \ldots + \alpha_m v_m$, with $\alpha_1 \neq 0$ and $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_m|$.

Thus,

$$B^k z = \alpha_1 \lambda_1^k v_1 + \alpha_2 \lambda_2^k v_2 + \ldots + \alpha_m \lambda_m^k v_m. \tag{5.2}$$

Note that if u is a vector such that $\langle v_1, u\rangle \neq 0$, then $\langle z, u\rangle \neq 0$ and

$$\frac{\langle B^{k+1} z, u\rangle}{\langle B^k z, u\rangle} = \lambda_1 \frac{\alpha_1 \langle v_1, u\rangle + \alpha_2 (\lambda_2/\lambda_1)^{k+1} \langle v_2, u\rangle + \ldots + \alpha_m (\lambda_m/\lambda_1)^{k+1} \langle v_m, u\rangle}{\alpha_1 \langle v_1, u\rangle + \alpha_2 (\lambda_2/\lambda_1)^{k} \langle v_2, u\rangle + \ldots + \alpha_m (\lambda_m/\lambda_1)^{k} \langle v_m, u\rangle}. \tag{5.3}$$

So that in the limiting case

$$\lim_{k \to \infty} \frac{\langle B^{k+1} z, u\rangle}{\langle B^k z, u\rangle} = \lambda_1. \tag{5.4}$$

Moreover, from (5.2) $\lambda_1^{-k}B^kz = \alpha_1v_1 + \alpha_2(\lambda_2/\lambda_1)^kv_2 + \ldots + \alpha_m(\lambda_m/\lambda_1)^kv_m$. Thus

$$\lim_{k\to\infty} \lambda_1^{-k}B^kz = \alpha_1v_1. \tag{5.5}$$

Thus by (5.4), we can first find the largest eigenvalue and then by (5.2) the eigenvector, corresponding to the largest eigenvalue, involved in the representation of z. Note that the eigenvector λ_1v_1 is not necessarily of unit length.

Problem 5.1. Use power method to find the largest eigenvalue and the corresponding eigenvector of the matrix $\begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}$ with the initial vector as $\begin{pmatrix} 3 \\ 1 \end{pmatrix}$.

Solution. Clearly, since $z = \begin{pmatrix} 3 \\ 1 \end{pmatrix}$ is not the eigenvalue of the matrix $B = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}$. So we can infer that if there is a basis of eigenvectors, z is a linear combination of both of them. Further since $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is not an eigenvector, its inner product with both the eigenvectors has to be non-zero and hence we can use this vector as u vector. Now

$$\begin{aligned}
Bz = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 8 \\ 6 \end{pmatrix} &\Rightarrow \langle B^1z, u\rangle = 8,\\
B^2z = BBz = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 8 \\ 6 \end{pmatrix} = \begin{pmatrix} 18 \\ 16 \end{pmatrix} &\Rightarrow \langle B^2z, u\rangle = 18,\\
B^3z = BB^2z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 18 \\ 16 \end{pmatrix} = \begin{pmatrix} 38 \\ 36 \end{pmatrix} &\Rightarrow \langle B^3z, u\rangle = 38,\\
B^4z = BB^3z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 38 \\ 36 \end{pmatrix} = \begin{pmatrix} 78 \\ 76 \end{pmatrix} &\Rightarrow \langle B^4z, u\rangle = 78,\\
B^5z = BB^4z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 78 \\ 76 \end{pmatrix} = \begin{pmatrix} 158 \\ 156 \end{pmatrix} &\Rightarrow \langle B^5z, u\rangle = 158,\\
B^6z = BB^5z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 158 \\ 156 \end{pmatrix} = \begin{pmatrix} 318 \\ 316 \end{pmatrix} &\Rightarrow \langle B^6z, u\rangle = 318,\\
B^7z = BB^6z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 318 \\ 316 \end{pmatrix} = \begin{pmatrix} 638 \\ 636 \end{pmatrix} &\Rightarrow \langle B^7z, u\rangle = 638,\\
B^8z = BB^7z = \begin{pmatrix} 3 & -1 \\ 2 & 0 \end{pmatrix}\begin{pmatrix} 638 \\ 636 \end{pmatrix} = \begin{pmatrix} 1278 \\ 1276 \end{pmatrix} &\Rightarrow \langle B^8z, u\rangle = 1278.
\end{aligned}$$

Thus first initial terms of the sequence $\left\{\frac{\langle B^{k+1}z,u\rangle}{\langle B^kz,u\rangle}\right\}$ are $\frac{18}{8} = 2.25$, $\frac{38}{18} = 2.1111$, $\frac{78}{38} = 2.0526$, $\frac{158}{78} = 2.0256$, $\frac{318}{158} = 2.0126$, $\frac{638}{318} = 2.0062$, $\frac{1278}{638} = 2.0031$. Thus up to to two decimal places largest eigenvalue is 2. Moreover first few terms of the sequence $\{\lambda_1^{-k}B^kz\}$ are $[3, 2]^t$, $[4.5, 4]^t$, $[4.75, 4.5]^t$, $[4.875, 4.75]^t$, $[4.9375, 4.875]^t$, $[4.96875, 4.9375]^t$, $[4.984375, 4.96875]^t$, $[4.9921875, 4.984375]^t$. Thus the eigenvector correct up to one decimal place is $[5, 5]^t$. Note that $\begin{pmatrix} 3 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix} + \begin{pmatrix} -2 \\ -4 \end{pmatrix}$, where $\begin{pmatrix} -2 \\ -4 \end{pmatrix}$ should be the eigenvector corresponding to the other eigenvalue. And hence other eigenvalue is 1.

Remark 5.1. Note that since $\|\cdot\|$ is a continuous function, from (5.5) we have $\lim_{k\to\infty}\|\lambda_1^{-k}B^kz\| = \|\alpha_1v_1\|$. And hence $\lim_{k\to\infty}\frac{\lambda_1^{-k}B^kz}{\|\lambda_1^{-k}B^kz\|} = \frac{\alpha_1v_1}{\|\alpha_1v_1\|}$, or $\lim_{k\to\infty}\frac{B^kz}{\|B^kz\|} = \frac{v_1}{\|v_1\|}$. Further since B represents a continuous linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$, we have $\lim_{k\to\infty} B\left(\frac{B^kz}{\|B^kz\|}\right) = B\left(\frac{v_1}{\|v_1\|}\right)$, or equivalently,

$$\lim_{k\to\infty}\frac{B^{k+1}z}{\|B^kz\|} = \lambda_1\frac{v_1}{\|v_1\|} \tag{5.6}$$

Note that by writing first few terms of the sequence $\left\{\frac{B^{k+1}z}{\|B^kz\|}\right\}$ for the Problem 5.1 w.r.t. infinity norm as $\frac{1}{8}\begin{pmatrix}18\\16\end{pmatrix} = 2.25\begin{pmatrix}1\\.88\end{pmatrix}$, $\frac{1}{18}\begin{pmatrix}38\\36\end{pmatrix} = 2.1111\begin{pmatrix}1\\.9474\end{pmatrix}$, $\frac{1}{38}\begin{pmatrix}78\\76\end{pmatrix} = 2.0526\begin{pmatrix}1\\.9744\end{pmatrix}$, $\frac{1}{78}\begin{pmatrix}158\\156\end{pmatrix} = 2.0256\begin{pmatrix}1\\.9873\end{pmatrix}$, $\frac{1}{158}\begin{pmatrix}318\\316\end{pmatrix} = 2.0127\begin{pmatrix}1\\.9937\end{pmatrix}$, $\frac{1}{318}\begin{pmatrix}638\\636\end{pmatrix} = 2.006\begin{pmatrix}1\\.9969\end{pmatrix}$, $\frac{1}{638}\begin{pmatrix}1278\\1276\end{pmatrix} = 2.003\begin{pmatrix}1\\.9948\end{pmatrix}$. This shows that the eigenvalue is 2 correct up to two decimal places and corresponding unit norm eigenvector is $\begin{pmatrix}1\\1\end{pmatrix}$ correct up to one decimal places.

5.2. QR Decomposition

Let $A = [a_1, a_2, \ldots, a_n]$ be a non-singular square matrix of order $n \times n$ such that $a_1, a_2, \ldots, a_n$ are the column vectors of A. We apply Gram-Schmidt Process to find an orthonormal basis from the basis $\{a_1, a_2, \ldots, a_n\}$. For this we first consider a unit vector in the direction of a_1 as $e_1 = \frac{a_1}{\|a_1\|}$. Next we search for a vector u_2 in the perpendicular direction of a_1 such that $span\{a_1, a_2\} = span\{e_1, u_2\}$, this can be obtained by subtracting from a_2 the projection of a_2 in the direction of a_1, that is, $\frac{\langle a_2, a_1\rangle}{\|a_1\|^2}a_1$. Thus $u_2 = a_2 - \langle a_2, e_1\rangle e_1$. Consider unit vector in the direction of u_2, that is, $e_2 = \frac{u_2}{\|u_2\|}$. Using $\langle a_2, e_2\rangle = \langle u_2, e_2\rangle$, we have $\langle a_2, e_2\rangle e_2 = \langle u_2, e_2\rangle e_2 = \|u_2\|e_2 = u_2$. Similarly we define $e_3, e_4, \ldots, e_n$ and get

$$\begin{aligned} a_1 &= \langle a_1, e_1\rangle e_1, \\ a_2 &= \langle a_2, e_1\rangle e_1 + \langle a_2, e_2\rangle e_2 \\ \text{In general ,} \qquad e_k &= \sum_{j=1}^{k} \langle a_k, e_j\rangle e_j. \end{aligned}$$

Thus if we consider

$$Q = [e_1, e_2, \ldots, e_n]^t, \qquad \text{and} \qquad R = \begin{bmatrix} \langle a_1, e_1\rangle & \langle a_2, e_1\rangle & \ldots & \langle a_n, e_1\rangle \\ 0 & \langle a_2, e_2\rangle & \ldots & \langle a_n, e_2\rangle \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \langle a_n, e_2\rangle \end{bmatrix},$$

then $A = QR$.

Note that the norm used here is $\|\cdot\|_2$, which is compatible with inner product. Moreover, it can also be shown that QR decomposition of a non-singular square matrix is unique.

Problem 5.2. Find the QR decomposition of the matrix $\begin{bmatrix}-3 & -5 & -8\\ 6 & 4 & 1\\ -6 & 2 & 5\end{bmatrix}$.

Solution. Note that $a_1 = [-3, 6, -6]^t$, $a_2 = [-5, 4, 2]^t$, and $a_3 = [-8, 1, 5]^t$. So that $\|a_1\| = \sqrt{9+36+36} = 9$ and $e_1 = [-1/3, 2/3, -2/3]^t$. Now $\langle a_1, e_1\rangle = 9$, $\langle a_2, e_1\rangle = 3$, $\langle a_3, e_1\rangle = 0$ so that $u_2 = [-5, 4, 2]^t - 3[-1/3, 2/3, -2/3]^t = [-4, 2, 4]^t$. Thus $e_2 = [-2/3, 1/3, 2/3]^t$ and $\langle a_2, e_2\rangle = 6$, $\langle a_3, e_2\rangle = 9$. Now $u_3 = [-8, 1, 5]^t - 0e_1 - 9[-2/3, 1/3, 2/3]^t = [-2, -2, -1]^t$ so that $e_3 = [-2/3, -2/3, -1/3]^t$ and $\langle a_3, e_3\rangle = 3$. Thus $Q = \frac{1}{3}\begin{bmatrix}-1 & -2 & -2\\ 2 & 1 & -2\\ -2 & 2 & -1\end{bmatrix}$ and $R = \begin{bmatrix}9 & 3 & 0\\ 0 & 6 & 9\\ 0 & 0 & 3\end{bmatrix}$.

5.3. QR Algorithm

Let A_1 be a square matrix of order n with n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$. Decompose A_1 as $A_1 = Q_1R_1$, where R_1 is an upper triangular matrix and

Q_1 is orthogonal matrix such that $Q_1^t = Q_1^{-1}$. Consider $A_2 = R_1Q_1 = Q_1^{-1}Q_1R_1Q_1 = Q_1^{-1}A_1Q_1$. Thus A_2 is similar to A_1 and hence the set of eigenvalues of A_2 is same as the set of eigenvalues of A_1. Now if A_2 has the QR decomposition as $A_2 = Q_2R_2$, we define $A_3 = R_3Q_3$, which is again similar to A_2 and hence similar to A_1. Thus we find a sequence of similar matrices $\{A_n\}$.

For the convergence following theorem is stated without proof.

Theorem 5.1. *If a matrix A of order $n \times n$ has n distinct eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$ such that $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$ and all the principal minors of the matrix of eigenvectors of A^t are non zero, then sequence $\{A_n\}$ converges to a diagonal matrix with diagonal entries as eigenvalues.*

Problem 5.3. Use QR algorithm to find the eigenvalues of the matrix $\begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$.

Solution. Let $A_1 = \begin{bmatrix} 3 & 1 \\ 2 & 2 \end{bmatrix}$. The QR factorization of $A = Q_1R_1$ where $Q_1 = \frac{1}{\sqrt{13}}\begin{bmatrix} 3 & -2 \\ 2 & 3 \end{bmatrix}$ and $R_1 = \frac{1}{\sqrt{13}}\begin{bmatrix} 13 & 7 \\ 0 & 4 \end{bmatrix}$. Thus $A_2 = R_1Q_1 = \frac{1}{13}\begin{bmatrix} 53 & -5 \\ 8 & 12 \end{bmatrix}$. This shows that the first approximation to the eigenvalues are $\frac{53}{13}$, $\frac{12}{13}$. To decompose A_2 as Q_2R_2 we follow the same procedure to get $Q_2 = \frac{1}{13\sqrt{17}}\begin{bmatrix} 53 & -8 \\ 8 & 53 \end{bmatrix}$ and $R_2 = \frac{1}{\sqrt{17}}\begin{bmatrix} 17 & -1 \\ 0 & 4 \end{bmatrix}$. Now $A_3 = R_2Q_2 = \frac{1}{221}\begin{bmatrix} 893 & 189 \\ 32 & 212 \end{bmatrix}$.Thus second approximation to the eigenvalues turns out to be $\frac{893}{221}$ and $\frac{212}{221}$. We can proceed further to find the next approximation.

5.4. Location of Eigenvalues

If λ is an eigenvalue of a square matrix B with eigenvector v, then $Bv = \lambda v$, or $\|Bv\| = |\lambda|\|v\|$ and hence

$$|\lambda| = \frac{\|Bv\|}{\|v\|} \leq \frac{\|B\| \times \|v\|}{\|v\|} = \|B\|.$$

Note that this is true for all possible matrix norms, that is, $\|\cdot\|_1, \|\cdot\|_2$ and $\|\cdot\|_\infty$. Thus

$$|\lambda| \leq \min\{\|B\|_1, \|B\|_2, \|B\|_\infty\}$$

Theorem 5.2. ***(Gershgorin's Theorem)*** *Let A be a square matrix B of order $n \times n$. Each eigenvalue λ of B satisfies*

$$|a_{ii} - \lambda| \leq \sum_{j=1, j\neq i}^{n} |a_{ij}|, \tag{5.7}$$

at least for some $1 \leq i \leq n$.

Proof. Let λ be the eigenvalue of the matrix B and v be an eigenvector corresponding to this eigenvalue so that $Av = \lambda v$. If $\|v\|_\infty = |v_k|$, then $|v_k| \geq |v_i|$ for all $1 \leq i \leq n$. Now since $\lambda v_k - a_{kk}v_k = a_{k1}v_1 + a_{k2}v_2 + \ldots + a_{k,k-1}v_{k-1} + a_{k,k+1}v_{k+1} + \ldots + a_{kn}v_n$, we have

$$\begin{aligned} |\lambda - a_{kk}| &= \left| a_{k1}\frac{v_1}{v_k} + a_{k2}\frac{v_2}{v_k} + \ldots + a_{k,k-1}\frac{v_{k-1}}{v_k} + a_{k,k+1}\frac{v_{k+1}}{v_k} + \ldots + a_{kn}\frac{v_n}{v_k} \right| \\ &\leq |a_{k1}|\frac{|v_1|}{|v_k|} + |a_{k2}|\frac{|v_2|}{|v_k|} + \ldots + |a_{k,k-1}|\frac{|v_{k-1}|}{|v_k|} + |a_{k,k+1}|\frac{|v_{k+1}|}{|v_k|} + \ldots + |a_{kn}|\frac{|v_n|}{|v_k|} \\ &\leq |a_{k1}| + |a_{k2}| + \ldots + |a_{k,k-1}| + |a_{k,k+1}| + \ldots + |a_{kn}|. \end{aligned}$$

Thus for $i = k$ the inequality (5.7) holds. □

Remark 5.2. Since the set of eigenvalues of a square matrix A is as of its transpose A^t, one can apply the Gershgorin's theorem to A^t to conclude

$$|a_{ii} - \lambda| \leq \sum_{j=1, j\neq i}^{n} |a_{ji}|.$$

Problem 5.4. Use Gerschgorin's theorem to find the location of eigenvalues of the matrix

$$\begin{pmatrix} 1 & 0 & -1 \\ 1 & -2 & 1 \\ 2 & -1 & -1 \end{pmatrix}.$$

Solution. Let λ be eigenvalue of the given matrix. According to Gerschgorin's theorem the λ has to satisfy at least one of the following conditions. $|\lambda - 1| \leq 1$, $|\lambda + 2| \leq 1 + 1$ and $|\lambda + 1| \leq 2 + 1$. Thus all the eigenvalues of the matrix lie within the union of these three disks. Further if we apply Gershgorin's theorem to transpose of the given matrix, then λ should lie within the union of the disks $|\lambda - 1| \leq 1 + 2$, $|\lambda + 2| \leq 1$, and $|\lambda + 1| \leq 1 + 1$. Thus finally we conclude that all the eigenvalues should lie within the intersection of these two unions.

6 Nonlinear Equation

In last section we learned to find the solution of a system of linear equations. But in practical problems it is also very important to find the value (or values) of an unknown satisfying certain nonlinear equation.

Question 6.1. Let f be a **nonlinear continuous** function defined on real line. Can we determine the values of x satisfying

$$f(x) = 0, \qquad x \in \mathbb{R}? \tag{6.1}$$

Remark 6.1. It might be easy to find the solution for $f(x) = 0$, if we know some how that this solution will lie in some particular interval. By using Intermediate value theorem one can find such an interval $[a, b]$ by determining two points $a, b \in \mathbb{R}$ such that $f(a)$ and $f(b)$ have different signs.

Now we will discuss some iterative methods to find such a solution.

6.1. Bisection Method

This method is based on the fact that if a continuous function f is changing the sign at two points, then f must have at least one root in the smallest interval containing these two points. And further subdivide this interval in two subintervals of equal length and search for the subinterval in which the root lies. This method can be described in following steps.

- First determine the interval $[a, b]$ in which the root lies. Which can be determined by finding two points a, b such that $f(a)f(b) < 0$. We consider middle point x_1 of $[a, b]$ as the first approximation to the root. If x_1 is the exact root then done or otherwise proceed to next step.

- Next choose the interval $[a_2, b_2]$ with one endpoint as x_1 and other endpoint as one of a or b according to $f(x_1)f(a) < 0$ or $f(x_1)f(b) < 0$ respectively. Now find second approximation x_2 as the middle point of $[a2, b_2]$ such that $|x_2 - r| < \frac{|b_2 - a_2|}{2} = \frac{|b-a|}{4}$. If x_2 is the exact root then done or otherwise proceed to next step.

- In general find the k_{th} approximation x_k to root as the middle point of $[a_k, b_k]$. If x_k is the root, stop the process. And if not, find next interval $[a_{k+1}, b_{k+1}]$ such that one endpoint of this interval is x_k and other endpoint is one of a_k or b_k according to $f(x_k)f(a_k) < 0$ or $f(x_k)f(b + k) < 0$ respectively. Note that

$$|x_k - r| \leq \frac{|b_k - a_k|}{2} = \frac{|b - a|}{2^k}. \tag{6.2}$$

6.2. Error analysis in bisection method

In bisection method we search for the location of the a root r of (6.1) lying in each interval $[a, b]$. The equation (6.2) shows that

$$|e_k| \leq \frac{(b - a)}{2^k}.$$

Using this inequality we can approach to the root as closer as we want by increasing the number of iterates.

6.3. The secant method

This method is based upon the linear approximation of the function. First we search two points near the root of the function, name these points as x_0, x_1 such that $|f(x_1)| < |f(x_0)|$. These points x_0, x_1 are first two approximations to the root. The line passing through two points $(x_0, f(x_0)), (x_1, f(x_1))$ on the graph of the function is called secant. We consider this secant as the

approximating function to find the next iterate x_2 to root as the intersection point of this secant with the real line. One can draw a figure and check that the slope of this secant can be written in two different ways using similar triangles rule (assuming without loss of generality that $x_1 < x_0$ and $0 < f(x_1) < f(x_0)$).

$$\frac{f(x_0)}{x_0 - x_2} = \frac{f(x_0) - f(x_1)}{x_0 - x_1}. \tag{6.3}$$

After solving this we get

$$x_2 = x_1 - f(x_1)\frac{x_0 - x_1}{f(x_0) - f(x_1)}. \tag{6.4}$$

Now we will use x_1, x_2 to determine the secant as the line passing through $(x_1, f(x_1)), (x_2, f(x_2))$ on the graph of the original function and consider x_3 as the intersection of this secant with the real line. In general, we consider the secant passing through the graphical points $(x_{k-1}, f(x_{k-1})), (x_k, f(x_k))$ and the next approximation x_{k+1} as the intersection point of the secant with the real line. We have

$$x_{k+1} = x_k - f(x_k)\frac{x_{k-1} - x_k}{f(x_{k-1}) - f(x_k)}. \tag{6.5}$$

Remark 6.2. The sequence of iterates does not need to converge to root of the function. In fact it might also diverse to infinity.

Question 6.2. Then, why does one use secant method instead of bisection method, which gives the security of convergence.

6.4. Order of convergence and asymptotic error

Suppose the sequence of iterates $\{x_0, x_1, \ldots\}$ converges to root r and for the sequence of errors $\{e_k = r - x_k\}$, there exists a number p and a constant $C \neq 0$ such that

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^p} = C, \tag{6.6}$$

then p is called the order of convergence and C is called the asymptotic error.

6.5. Error analysis of the secant method

Since we do not know exact positioning of the root, our treatment to error will be different from bisection method, where we knew for sure the interval in which the root lies. Let us go back to the theme of secant method: linear approximation to the function at each new iterate. In fact for linear approximation with nodes as x_{k-1}, x_k, we can use (2.35) to write the expression for the function as

$$f(x) = f(x_k) + f[x_k, x_{k-1}](x - x_k) + f[x_k, x_{k-1}, x](x - x_k)(x - x_{k-1}). \tag{6.7}$$

If r is the root for the function, that $f(r) = 0$, we have from (6.7)

$$0 = f(x_k) + f[x_k, x_{k-1}](r - x_k) + f[x_k, x_{k-1}, r](r - x_k)(r - x_{k-1}).$$

This implies (assuming $f[x_k, x_{k-1}] \neq 0$)

$$r = x_k - \frac{f(x_k)}{f[x_k, x_{k-1}]} - \frac{f[x_k, x_{k-1}, r]}{f[x_k, x_{k-1}]}(r - x_k)(r - x_{k-1}).$$

Now using (6.5) we get

$$r = x_{k+1} - \frac{f[x_k, x_{k-1}, r]}{f[x_k, x_{k-1}]}(r - x_k)(r - x_{k-1}). \tag{6.8}$$

Further if $r - x_k = e_k$ denotes the error at k_{th} iterate, then we get from (6.8) that

$$e_{k+1} = -\frac{f[x_k, x_{k-1}, r]}{f[x_k, x_{k-1}]} e_k e_{k-1}. \tag{6.9}$$

To determine the order of convergence of the secant method we note that from (6.9)

$$|e_{k+1}| = c_k |e_k e_{k-1}|, \tag{6.10}$$

where

$$c_k = \left| \frac{f[x_k, x_{k-1}, r]}{f[x_k, x_{k-1}]} \right|. \tag{6.11}$$

Now if we assume that the sequence of iterates converges to the root r and the function is twice continuously differentiable, we can write

$$\lim_{k\to\infty} c_k = \lim_{k\to\infty} |f''(\eta_k)/2f'(\xi_k)|, \tag{6.12}$$

where η_k belongs to smallest interval containing x_k, x_{k-1}, r and ξ_k belongs to smallest interval containing x_k, x_{k-1}. But because of convergence of the iterates to the root r these intervals are eventually shrinking to one point r as k tends to infinity and hence we have

$$\lim_{k\to\infty} c_k = |f''(r)/2f'(r)| := c. \tag{6.13}$$

Now using (6.10), we can write

$$\frac{|e_{k+1}|}{|e_k|^p} = c_k |e_k|^{1-p} |e_{k-1}| = c_k \left(\frac{|e_k|}{|e_{k-1}|^p} \right)^{\alpha}, \tag{6.14}$$

provided $\alpha = 1 - p$ and $\alpha p = -1$, that is, $p^2 - p - 1 = 0$ so that $p = (1 + \sqrt{5})/2$. Note that the negative value of p is not considered because it leads to $\lim_{k\to\infty} |e_{k+1}|/|e_k|^p = 0$, we don't want this situation to determine the order of convergence. If we consider $y_k = |e_k|/|e_{k-1}|^p$, we have from (6.14)

$$y_{k+1} = c_k y_k^{-1/p} \tag{6.15}$$

Now if $y_k \to y$, then $y = cy^{-1/p}$ or $y = c^{1/p}$. Thus the asymptotic error is $|f''(r)/2f'(r)|^{1/p}$ and order of convergence is $p = (1 + \sqrt{5})/2 = 1.618$.

6.6. Newton's Method

Newton's method of finding the root of the equation $f(x) = 0$ is based upon the approximation of the function with the tangent line near the root. Thus our function is assumed to be differentiable. Here we only need one point x_0 near the root as the initial approximation to begin with. To find next approximation we draw the tangent line at $(x_0, f(x_0)$ and consider the intersection point of this tangent line with the real line and name it as x_1. Thus

$$f'(x_0) = \frac{f(x_0)}{x_0 - x_1}.$$

This gives

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}. \tag{6.16}$$

In general to find $(k+1)_{th}$ iterate, we use tangent line at $(x_k, f(x_k)$ as the approximation function and consider its root, that is, the intersection point with the real line. And we have

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}. \tag{6.17}$$

Remark 6.3. Newton's method also need not to converge.

6.7. Error analysis of Newton's method

One can write the function in the Newtonian form as

$$f(x) = f(x_k) + f[x_k, x_k](x - x_k) + f[x_k, x_k, x](x - x_k)^2. \tag{6.18}$$

If r is the root of the function, we have $0 = f(x_k) + f[x_k, x_k](r - x_k) + f[x_k, x_k, r](r - x_k)^2$. Or,

$$r = x_k - \frac{f(x_k)}{f[x_k, x_k]} - (r - x_k)^2 \frac{f[x_k, x_k, r]}{f[x_k, x_k]} = x_{k+1} - (r - x_k)^2 \frac{f[x_k, x_k, r]}{f[x_k, x_k]}. \tag{6.19}$$

Further if $e_k = r - x_k$ is the error at k_{th} stage,

$$e_{k+1} = -\frac{f[x_k, x_k, r]}{f[x_k, x_k]} e_k^2 = -\frac{f''(\eta_k)}{2f'(x_k)} e_k^2, \tag{6.20}$$

where η_k belongs to smallest interval containing x_k, r. And if sequence of iterates converges to r, then this interval eventually shrinks to point r itself and hence

$$\lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^2} = \lim_{k\to\infty} \left| \frac{f''(\eta_k)}{2f'(x_k)} \right| = \left| \frac{f''(r)}{2f'(r)} \right|. \tag{6.21}$$

Above equation shows that if the Newton's method converges, then order of convergence is 2 with asymptotic error as $|f''(r)/2f'(r)|$.

6.8. Fixed point iteration method

If $f(x) = 0 \Leftrightarrow x = g(x)$, then instead of finding roots of $f(x)$ we search for the fixed points of the function $g(x)$ (both are same). Consider some initial approximation to fixed point say x_0 to determine first approximation by the equation $x_1 = g(x_0)$ and in general

$$x_{n+1} = g(x_n) \qquad \text{for all } n \in \mathbb{N}. \tag{6.22}$$

Theorem 6.1. *If $g : [a, b] \to [a, b]$ is a continuous function, then there is at least one fixed point for g in $[a, b]$. If g is differentiable in (a, b) and*

$$|g'(x)| \le \alpha < 1 \text{ for all } x \in (a, b).$$

then there is exactly one fixed point of g in (a, b). Further if $x_1 \in (a, b)$, then the sequence, defined by

$$x_n = g(x_{n-1}),$$

converges to the fixed point.

Proof. Since the function $g(x) - x$ changes sign in $[a, b]$, by Intermediate Value Property of continuous functions $g(x) - x$ has a root in $[a, b]$. Now suppose if possible, $\xi_1, \xi_2 \in (a, b)$ be such that $g(\xi_1) = \xi_1$ and $g(\xi_2) = \xi_2$. Then by mean value theorem there exists a point c in between ξ_1 and ξ_2 such that

$$g'(c) = \frac{g(\xi_2) - g(\xi_1)}{\xi_2 - \xi_1} = \frac{\xi_2 - \xi_1}{\xi_2 - \xi_1} = 1,$$

a contradiction to the assumption. Hence there is only one fixed point say ξ in (a, b).

We will use Mathematical Induction to prove that the set $\{x_n : n \in \mathbb{N}\}$ is a subset of (a, b). Clearly $x_2 = g(x_1) \in (a, b)$ since $x_1 \in (a, b)$. Next if $x_k \in (a, b)$, $x_{k+1} = g(x_k) \in (a, b)$. Note that

$$|x_{k+1} - x_k| = |g(x_k) - g(x_{k-1})| = |g'(\lambda_k)(x_k - x_{k-1})| = |g'(\lambda_k)||(x_k - x_{k-1})| \le \alpha |x_k - x_{k-1}|.$$

Now since $\alpha < 1$, it can be proved that $\{x_k\}$ is a Cauchy sequence and hence convergent. If $x_k \to x$, then by continuity of g, $g(x_k) \to g(x)$ (Note that $x \in [a, b]$). And hence taking the limit in (6.22), we get $x = g(x)$. Thus x is the fixed point of g in (a, b), or $x = \xi$. □

Theorem 6.2. *Let l_0 be a fixed point of $g(x)$. Suppose $\epsilon > 0$ be such that g is differentiable on $[l_0 - \epsilon, l_0 + \epsilon]$ and $|g'(x)| \leq \alpha < 1$ for all $x \in [l_0 - \epsilon, l_0 + \epsilon]$. Then the sequence defined by $x_n = g(x_{n-1})$ and $x_1 \in [l_0 - \epsilon, l_0 + \epsilon]$, converges to l_0.*

Proof. To prove this theorem we first prove that $g([l_0-\epsilon, l_0+\epsilon]) \subset [l_0-\epsilon, l_0+\epsilon]$. If $x \in [l_0-\epsilon, l_0+\epsilon]$, then $|l_0 - g(x)| = |g(l_0) - g(x)| = |g'(c)|\ |l_0 - x| \leq \alpha|l_0 - x| \leq \alpha\epsilon < \epsilon$, hence $g(x) \in [l_0 - \epsilon, l_0 + \epsilon]$. Now g fulfils the condition of Theorem (6.1) with interval $[a, b]$ as $[l_0 - \epsilon, l_0 + \epsilon]$ so we conclude the rest. □

Remark 6.4. Suppose if g is twice differentiable, and $g'(r) \neq 0$, where r is the fixed point of g, then by Taylor's formula

$$x_{k+1} = g(x_k) = g(r + x_k - r) = g(r) + g'(r)(x_k - r) + \frac{1}{2}g''(c)(x_k - r)^2 = r + g'(r)e_k + \frac{1}{2}g''(c)e_k^2.$$

Thus,

$$\frac{|e_{k+1}|}{|e_k|} = |g'(r) + \frac{1}{2}g''(c)e_k|.$$

Now if $\{x_k\}$ converges to r or equivalently $e_k \to 0$, we have

$$\lim_{k\to\infty} \frac{|e_{k+1|}}{|e_k|} = |g'(r)|.$$

This shows that the order of convergence of fixed point iteration method is 1 and asymptotic error is $|g'(r)|$.

Remark 6.5. Note that the iterations of Newton's method are given by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}.$$

If $g(x) = x - \frac{f(x)}{f'(x)}$, then Newton's method is also a fixed point iteration method. From the last remark order of convergence of fixed point iteration method is 1, but we know that the order of convergence of Newton;s method is 2. This ambiguity arises because in Newton's method $g'(r) = 0$, which is not the case in general for fixed point iteration method. Now let us examine the convergence behavior of the fixed point iteration method if $g'(r) = 0$ and $g''(r) \neq 0$. Suppose g is thrice differentiable, then by Taylors formula

$$\begin{aligned} x_{k+1} &= g(x_k) = g(r + x_k - r) = g(r) + g'(r)(x_k - r) + \frac{1}{2}g''(r)(x_k - r)^2 + \frac{1}{6}g'''(c)(x_k - r)^3, \\ &= r + \frac{1}{2}g''(r)e_k^2 + \frac{1}{6}g'''(c)e_k^3. \end{aligned}$$

Thus

$$\frac{|e_{k+1}|}{|e_k|^2} = |\frac{1}{2}g''(r) + \frac{1}{6}g'''(c)e_k|, \quad \text{or} \quad \lim_{k\to\infty} \frac{|e_{k+1}|}{|e_k|^2} = |\frac{1}{2}g''(r)|.$$

Hence the convergence is of order 2 with asymptotic error as $|\frac{1}{2}g''(r)|$. Since for Newton's method $g(x) = x - \frac{f(x)}{f'(x)}$, the asymptotic error $|\frac{1}{2}g''(r)| = \frac{1}{2}\frac{|f''(r)|}{|f'(r)|}$.

6.9. Multiple Roots

If $f(x) = 0$ has a root at $x = r$ with multiplicity m greater than 1, then f can be written as $f(x) = (x - r)^m h(x)$, where h is a function such that $h(r)$ is nonzero finite quantity. Now

$$g(x) = x - \frac{(x - r)^m h(x)}{m(x - r)^{m-1}h(x) + (x - r)^m h'(x)} = x - \frac{x - r}{m} + \frac{(x - r)^2 h'(x)}{m^2 h(x) + mh'(x)(x - r)}.$$

And we have $g'(r) = 1 - \frac{1}{m}$. This shows that Newton's method converges linearly when $m > 1$. We aim to modify Newton's method to ensure the quadratic convergence. For this we modify g such that $g'(r) = 0$ and $g(r) = r$. Note that if we assume modified g as $g(x) = x - \alpha\frac{f(x)}{f'(x)}$, then $g'(r) = 1 - \frac{\alpha}{m}$. Thus we assume $g(x) = x - m\frac{f(x)}{f'(x)}$, so that modified sequence of iterates is

$$x_{k+1} = x_k - m\frac{f(x_k)}{f'(x_k)}. \tag{6.23}$$

Note that this sequence of iterates can be applied if we know a priory the multiplicity of the root. But to get the quadratic convergence, without knowing the multiplicity of the roots, we assume $g(x) = x - \frac{u(x)}{u'(x)}$, where $u(x) = \frac{f(x)}{f'(x)}$. Thus $g(r) = r$ and $g'(r) = 0$, and hence iterates defined by $x_{k+1} = g(x_k)$, converges quadratically irrespective of the multiplicity of the root of $f(x) = 0$. Thus in general we use following approximations to get quadratic convergence in case of roots with multiplicity greater than 1,

$$x_{k+1} = x_k - \frac{f(x_k)f'(x_k)}{(f'(x_k))^2 - f(x_k)f''(x_k)}. \tag{6.24}$$

Note that in this case we need an extra evaluation of $f''(x)$ in each iteration.

6.10. Solution of System of Nonlinear Equations

Consider the system of two nonlinear equations in two unknowns.

$$\begin{aligned} f(x,y) &= 0,\\ g(x,y) &= 0. \end{aligned}$$

Let $F : \mathbb{R}^2 \to \mathbb{R}^2$ be defined by $F(x,y) = \begin{bmatrix} f(x,y) \\ g(x,y) \end{bmatrix}$. Using Taylor's series expansion for multi variable function we have

$$\begin{aligned} 0 &= f(x,y) = f(x_k + x - x_k, y_k + y - y_k) = f(x_k + \Delta x, y_k + \Delta y)\\ &= f(x_k, y_k) + [\Delta x_k(\partial/\partial x) + \Delta y_k(\partial/\partial y)]f|_{x_k,y_k} + \frac{1}{2}[\Delta x_k(\partial/\partial x) + \Delta y_k(\partial/\partial y)]^2 f|_{x_k,y_k} + \dots. \end{aligned}$$

Neglecting higher order terms $[\Delta x_k(\partial/\partial x) + \Delta y_k(\partial/\partial y)]f|_{x_k,y_k} \approx -f(x_k, y_k)$. Thus we have

$$\begin{aligned} [\Delta x_k f_x + \Delta y_k f_y]|_{(x_k,y_k)} &\approx -f(x_k, y_k),\\ [\Delta x_k g_x + \Delta y_k g_y]|_{(x_k,y_k)} &\approx -g(x_k, y_k). \end{aligned}$$

Or,

$$\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix}_{(x_k,y_k)} \begin{bmatrix} \Delta x_k \\ \Delta y_k \end{bmatrix} \approx -\begin{bmatrix} f \\ g \end{bmatrix}_{(x_k,y_k)}. \tag{6.25}$$

Let J denotes the Jacobian $\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix}$ and J_k denotes the Jacobian J evaluated at $X_k = [x_k, y_k]^t$. Thus we can write $J_k\Delta X_k \approx -F(X_k)$, or $\Delta X_k \approx -J_k^{-1}F(X_k)$, or $X \approx X_k - J_k^{-1}F(X_k)$, or

$$\begin{bmatrix} x \\ y \end{bmatrix} \approx \begin{bmatrix} x_k \\ y_k \end{bmatrix} - J_k^{-1}\begin{bmatrix} f(x_k, y_k) \\ g(x_k, y_k) \end{bmatrix}. \tag{6.26}$$

Thus we define sequence of iterates as

$$\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} x_k \\ y_k \end{bmatrix} - J_k^{-1}\begin{bmatrix} f(x_k, y_k) \\ g(x_k, y_k) \end{bmatrix}. \tag{6.27}$$

This method is known as Newton's iteration method of solving system of nonlinear equations. In general for system of n linear equations in n unknowns we define the iterates as

$$X_{k+1} = X_k - J_k^{-1} F(X_k), \tag{6.28}$$

where $X_k = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^t$, $F(X) = [f_1(X), f_2(X), \ldots, f_n(X)]^t$ and J_k is the Jacobian of $f_1, f_2, \ldots, f_n$ with respect to $x_1, x_2, \ldots, x_n$ evaluated at $X_k = [x_1^{(k)}, x_2^{(k)}, \ldots, x_n^{(k)}]^t$.

Remark 6.6. Suppose we have system of two linear equations say

$$ax + by = s \tag{6.29}$$
$$cx + dy = t. \tag{6.30}$$

We can view them as $f_1(x, y) = ax+by-s = 0$ and $f_2(x, y) = cx+dy-t = 0$. So that Jacobian J of f_1, f_2 with respect to x, y is $\begin{bmatrix} f_x & f_y \\ g_x & g_y \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$, which is constant. Now we assume that $[x_0, y_0]^t$ is the initial approximation to the solution and J is invertible so that we can apply Newton's method to find first approximation to the solution say $[x_1, y_1]^t$ as

$$\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} \begin{bmatrix} f(x_0, y_0) \\ g(x_0, y_0) \end{bmatrix}. \tag{6.31}$$

Or,

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - \begin{bmatrix} f_1(x_0, y_0) \\ g(x_0, y_0) \end{bmatrix} = \begin{bmatrix} ax_0 + by_0 \\ cx_0 + dy_0 \end{bmatrix} - \begin{bmatrix} ax_0 + by_0 - s \\ cx_0 + dy_0 - t \end{bmatrix} = \begin{bmatrix} s \\ t \end{bmatrix}. \tag{6.32}$$

This shows that when we apply Newton's method to linear equations, then defining sequence of iterates becomes exact.

Remark 6.7. We can find the complex roots of an equation $f(z) = 0$ by finding the real and imaginary parts of $f(z) = f(x + iy)$ as $f(x + iy) = u(x, y) + iv(x, y)$ and then solving the system of equations as

$$u(x, y) = 0; \qquad v(x, y) = 0.$$

Problem 6.1. Find the solution of the system of following equations

$$x^2 + 10x + 2y - 13 = 0; \qquad x^2 + 6y^2 - 7 = 0,$$

using initial approximation as $x_0 = 0.5$ and $y_0 = 0.5$.

Solution. Let $f(x, y) = x^2 + 10x + 2y - 13$ and $g(x, y) = x^2 + 6y^2 - 7$. Thus Jacobian

$$J = \begin{bmatrix} 2x + 10 & 2 \\ 2x & 12y \end{bmatrix}; \qquad J^{-1} = \frac{1}{24xy + 120y - 4x} \begin{bmatrix} 12y & -2 \\ -2x & 2x + 10 \end{bmatrix}$$

and $\begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = \begin{bmatrix} x_0 \\ y_0 \end{bmatrix} - J^{-1}(x_0, y_0) \begin{bmatrix} f(x_0, y_0) \\ g(x_0, y_0) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} - \frac{1}{64} \begin{bmatrix} 6 & -2 \\ -1 & 11 \end{bmatrix} \begin{bmatrix} -6.75 \\ -4.5 \end{bmatrix} = \begin{bmatrix} 0.9921875 \\ 1.16796875 \end{bmatrix}$ and $\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} - J^{-1}(x_1, y_1) \begin{bmatrix} f(x_1, y_1) \\ g(x_1, y_1) \end{bmatrix} = \begin{bmatrix} 0.9921875 \\ 1.16796875 \end{bmatrix} - \frac{1}{163.9997559} \begin{bmatrix} 1.4015625 & -2 \\ -1.984375 & 11.984375 \end{bmatrix} \begin{bmatrix} 0.2422485352 \\ 2.169342041 \end{bmatrix} = \begin{bmatrix} 1.016572644 \\ 1.021542721 \end{bmatrix}$. Now we find next iterate as $\begin{bmatrix} x_3 \\ y_3 \end{bmatrix} = \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} - J^{-1}(x_2, y_2) \begin{bmatrix} f(x_2, y_2) \\ g(x_2, y_2) \end{bmatrix} = \begin{bmatrix} 1.016572644 \\ 1.021542721 \end{bmatrix} - \frac{1}{143.4421732} \begin{bmatrix} 12.19887173 & -2 \\ -2.033145288 & 12.033145288 \end{bmatrix} \begin{bmatrix} 0.2422318225 \\ 0.2947171255 \end{bmatrix} = \begin{bmatrix} 1.00008153 \\ 1.000252737 \end{bmatrix}$. Thus solution correct up to one decimal place is $x = 1, y = 1$.