Numerical Analysis - Math 464 and 465 Notes

Brett Saiki

March 2022

This is a compilation of notes from the numerical analysis sequence, Math 464 and 465, at the University of Washington. Math 464 was taught by Kenneth P. Bube and those notes borrow heavily from *Numerical Analysis*, 2nd edition, 1982, by L. W. Johnson and R.D. Wiess.

Contents

1 Floating Point and Roundoff Error

1.1 Number Representation

Definition 1.1. Let $\beta > 1$ be an integer. We call β the *base* of a number system. Let a_k, b_k be integers such that $0 \leq a_k, b_k < \beta$. Then any real number x can be represented by

$$x = (a_n\, a_{n-1} \cdots a_1\, a_0 \,.\, b_1\, b_2\, b_3 \cdots)_\beta \,.$$

We call the dot between a_0 and b_1 the *radix point.* Alternatively, we can represent x by two summations:

$$x = a_k\beta^k + a_{k-1}\beta^{k-1} + \cdots + a_1\beta + a_0 + b_1\beta^{-1} + b_2\beta^{-2} + \cdots = \sum_{k=0}^{n} a_k\beta^k + \sum_{k=1}^{\infty} b_k\beta^{-k}$$

We call the first sum the *integral part of* x and denote it by x_I, and the second sum the *fractional part of* x and denote it by x_F. We call for formulas above the *expansion* of x.

Definition 1.2. An expansion of some real number x is said to *terminate* if there exists some $K \geq 0$ such that $b_k = 0$ for all $k \geq K$.

Theorem 1.3. A real number x has a terminating expansion in base β if and only if x is rational and when x is expressed in simplest form, the only prime factors of the denominator of x are factors of β.

Theorem 1.4. Let x be a real number. If x does not have a terminating expansion in base β, then the expansion of x in base β is unique. If $x \neq 0$, has a terminating expansion in base β, then it has exactly on terminating expansion (ending in zeros) and exactly one nonterminating expansion (ending in $(\beta - 1)$'s).

Remark.

(i) The expansions of negative numbers are just prefixed by a minus sign, e.g. $-1/8 = -(0.12500\cdots)_{10}$.

(ii) There are algorithms for converting expansions from one case to another.

1.2 Normalized Scientific Notation in Base β

Lemma 1.5. Let $\beta > 1$ be an integer. For any real number $x > 0$, there is a unique integer c and a unique number $r \in [1/\beta, 1)$ so that $x = r\beta^c$. The number r can be expressed as an expansion in base β,

$$r = (.\, d_1\, d_2\, d_3 \cdots)_\beta$$

with $d_1 \neq 0$.

Theorem 1.6. Let $x \neq 0$ be any real number. Then x has an expansion in base β,

$$x = \pm\, (.\, d_1\, d_2\, d_3 \cdots)_\beta\, \beta^c$$

with $d_1 \neq 0$.

Definition 1.7. The representation of x in Theorem 1.6 is called the *normalized scientific notation* for x in base β. It is unique, except for real numbers x with terminating expansions (which have two expansions); we always choose the terminating expansion.

1.3 Floating Point Arithmetic

Definition 1.8. An *m-digit floating-point number in base* β is denoted by

$$x = \pm\,(.\,d_1\,d_2\cdots d_m)_\beta\,\beta^c$$

where $(.\,d_1\,d_2\cdots d_m)_\beta$ is called the *mantissa* and c is called the exponent. If $d_1 \neq 0$ (or $x = 0$), called a *normalized floating-point number.*

Remark. In computers, the base is usually $\beta = 2$ and mantissa lengths usually comes in two sizes: single (23) and double (52). Additionally, the exponent c has a limited range $-M \leq c \leq M$.

Definition 1.9. Any real number can be represented approximately by floating-point numbers. For every real number x, the floating-point value $\mathrm{fl}(x)$ is the approximate value of x. Generally, fl is only well defined for some domain $\{x : \beta^{\mu-1} \leq |x| < \beta^M\}$. Otherwise, *underflow* or *overflow* occurs.

Definition 1.10. The function fl is commonly defined in two different ways:

(i) *Rounding* - $\mathrm{fl}(x)$ is the normalized floating-point number closest fo x. In case of a tie, round to an even digit (symmetric rounding about 0).

(ii) *Truncating* - $\mathrm{fl}(x)$ is the nearest normalized floating-point number between x and 0.

Remark. A more precise definition of the fl functions exists for even β. Let $x = \pm r\beta^c$ be a real number in normalized scientific notation where

$$r = (0\,.\,d_1\,d_2\,d_3\cdots)$$

Then $\mathrm{fl}(x)$ for an m-digit floating-point representation with a maximum M exponent is

$$\mathrm{fl}(x) = \begin{cases} 0, & x = 0 \\ \text{underflow}, & 0 < |x| < \beta^{\mu-1} \text{ (possibly extended to } \beta^{\mu-m} \leq |x| < \beta^{\mu-1}) \\ \text{overflow}, & |x| \geq \beta^M \\ \pm(.\,d_1\,d_2\cdots d_m)_\beta\beta^c, & \text{truncating} \\ \pm(.\,d_1\,d_2\cdots d_m)_\beta\beta^c, & \text{rounding, } (\,d_{m+1}\,d_{m+2}\cdots) < 1/2 \\ \pm\left[(.\,d_1\,d_2\cdots d_m)_\beta + (.00\cdots 1)_\beta\right]\beta^c, & \text{rounding, } (\,d_{m+1}\,d_{m+2}\cdots) > 1/2 \\ \pm\left[(.\,d_1\,d_2\cdots d_m)_\beta + (.00\cdots 1)_\beta\right]\beta^c, & \text{rounding, } (\,d_{m+1}\,d_{m+2}\cdots) = 1/2, d_m \text{ is odd} \\ \pm\left[(.\,d_1\,d_2\cdots d_m)_\beta - (.00\cdots 1)_\beta\right]\beta^c, & \text{rounding, } (\,d_{m+1}\,d_{m+2}\cdots) = 1/2, d_m \text{ is even} \end{cases}$$

1.4 Absolute and Relative Error

Definition 1.11. Suppose that x' is an approximation to a real number x. Then the *absolute error in* x' is $x - x'$ and the *relative error in* x' (if $x \neq 0$) is $(x - x')/x$.

Definition 1.12. The *roundoff error* is the error in $\mathrm{fl}(x)$ as an approximation to x. Usually it is absolute error $x - \mathrm{fl}(x)$.

Theorem 1.13. Suppose $\beta^{\mu-1} \leq |x| < \beta^M$. Define $\delta = \delta(x) = (\mathrm{fl}(x) - x)/x$ to be the relative error of $\mathrm{fl}(x)$.

(i) For rounding, $|\delta| \leq \beta^{1-m}/2$.

(ii) For truncating, $-\beta^{1-m} < \delta \leq 0$.

Definition 1.14. The maximum possible value for $|\delta|$ when there is no underflow or overflow is called the *unit roundoff*, denoted by u. In rounding, $u = \beta^{1-m}/2$. In truncating, $u = \beta^{1-m}$.

Remark. The value $\delta = (\mathrm{fl}(x) - x)/x$ can be rearranged to form $\mathrm{fl}(x) = x(1+\delta)$. This is useful in error analysis. If we define $\varepsilon(x) = (\mathrm{fl}(x) - x)/\mathrm{fl}(x)$, then $|\varepsilon| < \beta^{1-m}/2$ for rounding and $|\varepsilon| < \beta^{1-m}$ for truncating. Here, $\mathrm{fl}(x) = x/(1+\epsilon)$.

Definition 1.15. The *machine epsilon* is defined to be $\varepsilon = \sup\{y > 0 : \mathrm{fl}(1+y) = 1\}$.

Remark. The machine epsilon can also be defined to be $\varepsilon = \inf\{y > 0 : \mathrm{fl}(1+y) > 1\}$. The machine epsilon is exactly the same as the unit roundoff.

1.5 Arithmetic Operations with Floating-Point Numbers

Definition 1.16. With β, m fixed, the set of floating-point numbers is not closed under the usual operations $+$, $-$, $\times$, and $\div$. Machines are usually constructed so that

$$x \circ^* y = \mathrm{fl}(x \circ y).$$

where $\circ$ is $+$, $-$, $\times$, or $\div$, and $\circ^*$ is the corresponding *floating-point operation.* Unless underflow or overflow occurs

$$x \circ^* y = (x \circ y)(1+\delta)$$

for some δ where $\delta \leq u$ where x, y are floating-point numbers. Alternatively,

$$x \circ^* y = (x \circ y)/(1+\varepsilon)$$

for some ε where $|\varepsilon| \leq \mu$.

Theorem 1.17. Suppose $0 < u < 1$ and $|\delta_j| \leq u$ for $j = 1, \ldots, r$. Then there exists a δ with $|\delta| \leq u$ such that

$$(1+\delta_1)\cdots(1+\delta_r) = (1+\delta)^r$$

Corollary 1.18. For the theorem above, if $ru \ll 1$, then $(1+\delta)^r \approx 1 + r\delta$.

Remark. For two real number p, q, the operation $\mathrm{fl}(p) \times \mathrm{fl}(q)$ is

$$\mathrm{fl}(p) \times \mathrm{fl}(q) = pq(1+\delta_1)(1+\delta_2)(1+\delta_3) = pq(1+\delta)^3.$$

This kind of analysis is called backward error analysis.

Definition 1.19. Suppose x is written in normalized scientific notation in base β,

$$x = (.\, d_1\, d_2\, d_3 \cdots)_\beta \beta^c$$

where $d_1 \neq 0$. The digit d_j is called the *j-th significant digit* of x; d_j is the coefficient of β^{c-j}.

Definition 1.20. Suppose x' is an approximation to x. If $|x - x'| \leq \beta^{c-r}/2$, we say x' *approximates x to r significant digits.* Very approximately, the number of significant digits in x' is $-\log_\beta |(x - x')/x|$.

Theorem 1.21. Very approximately, if x and y have t significant digits, have the same sign, and agree to s significant digits, then the computed value of $x - y$ will have only $t - s$ significant digits.

Theorem 1.22. Let $x_1, x_2, \ldots, x_{n+1}$ be positive normalized floating-point numbers, $+$ be true addition, $\bigoplus$ be machine addition, u be the unit roundoff with $0 < u < 1$, and assume no overflow when we add $x_1, \ldots, x_{n+1}$. Then there are numbers $\delta_1, \ldots, \delta_n$ with $|\delta_j| \leq u$ for which

(i) $x_1 \bigoplus x_2 = (x_1 + x_2)(1 + \delta_1)$

(ii) $(x_1 \bigoplus x_2) \bigoplus x_3 = (x_1 + x_2)(1 + \delta_1)(1 + \delta_2) + x_3(1 + \delta_2)$

(iii) $x_1 \bigoplus x_2 \bigoplus \cdots \bigoplus x_{n+1} = (x_1 + x_2)(1 + \delta_1) \cdots (1 + \delta_n) + x_3(1 + \delta_2) \cdots (1 + \delta_n) + \cdots + x_{n+1}(1 + \delta_n)$

Remark. Consider solving $ax^2 + bx + c = 0$ by the quadratic formula when $ac \neq 0$, $b \neq 0$, and $b^2 - 4ac > 0$. The two solutions can be each written in two ways:

$$\frac{-b + \sqrt{b^2 - 4ac}}{2a} = \left(\frac{-b + \sqrt{b^2 - 4ac}}{2a}\right)\left(\frac{-b - \sqrt{b^2 - 4ac}}{-b - \sqrt{b^2 - 4ac}}\right) = \frac{4ac}{2a(-b - \sqrt{b^2 - 4ac})} = \frac{2c}{-b - \sqrt{b^2 - 4ac}},$$

and similarly,

$$\frac{-b - \sqrt{b^2 - 4ac}}{2a} = \frac{2c}{-b + \sqrt{b^2 - 4ac}}.$$

When $b > 0$, $-b + \sqrt{b^2 - 4ac}$ could have cancellation, and when $b < 0$, $-b - \sqrt{b^2 - 4ac}$ could have cancellation. Thus a better implementation of the quadratic formula is when $b > 0$, the two roots are $2c/(-b - \sqrt{b^2 - 4ac})$ and $(-b - \sqrt{b^2 - 4ac})/2a$, and when $b < 0$, the two roots are $(-b + \sqrt{b^2 - 4ac})/2a$ and $2c/(-b + \sqrt{b^2 - 4ac})$.

1.6 Converting Between Bases

Theorem 1.23. Suppose $N = (a_n\, a_{n-1} \cdots a_0)_\alpha$ is represented in base α. The expansion of N in base β can be found using two different methods:

(i) Express $\alpha, a_0, a_1, \cdots, a_n$ in base β. Then N is

$$N = (((a_n \cdot \alpha + a_{n-1}) \cdot \alpha + \cdots) \cdot \alpha + a_1) \cdots \alpha + a_0$$

where each operation is in base β arithmetic.

(ii) Suppose $N = (c_m\, c_{m-1} \cdots c_0)_\beta$. Then

$$N = c_0 + \beta \cdot (c_1 + \beta \cdot (c_2 + \cdots)).$$

Theorem 1.24. Suppose $x = (.\, b_1\, b_2 \cdots b_m)_\alpha$ is represented in base α. The expansion of x in base β can be found using two different methods:

(i) Express $\alpha, b_1, b_2 \cdots, b_m$ in base β. Then N is

$$N = (((b_m/\alpha + b_{m-1})/\alpha + \cdots + b_2)/\alpha + b_1)/\alpha$$

where each operation is in base β arithmetic.

(ii) Suppose $N = (c_m\, c_{m-1} \cdots c_0)_\beta$. The expansion of x can be found by successively solving for each coefficient in base β. Let $x = (.\, c_1\, c_2 \cdots)_\beta$ for unknown coefficients $c_1, c_2, \ldots$.

$$\begin{aligned} \beta x &= (c_1 \,.\, c_2\, c_3 \cdots)_\beta, \quad \text{so} \quad c_1 = (\beta x)_I \\ \beta(\beta x)_F &= (c_2 \,.\, c_3\, c_4 \cdots)_\beta, \quad \text{so} \quad c_2 = (\beta(\beta x)F)_I \\ &\vdots \end{aligned}$$

2 Solutions of Linear Systems

2.1 Solutions of Linear Systems using Elimination

Definition 2.1. Consider the matrix equation $A\mathbf{x} = \mathbf{b}$ where A is an upper triangular matrix whose diagonal entires are all non-zero, that is,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

To solve for $\mathbf{x}$, begin with x_n: $x_n = b_n/a_{nn}$. Then solve for x_{n-1}: $x_{n-1} = (b_{n-1} - a_{n-1,n}x_n)/a_{n-1,n-1}$. In general,

$$x_k = \frac{b_k - \sum_{j=k+1}^{n} a_{kj}x_j}{a_{kk}}.$$

This method of solving is called *back subtitution.*

Theorem 2.2. An upper triangular matrix A is invertible if and only if all diagonal entries are non-zero.

Definition 2.3. For any matrix equation $A\mathbf{x} = \mathbf{b}$ where A is a square matrix, the method of solving for $\mathbf{x}$ by transforming the equation into an equivalent equation where the matrix is an upper triangular matrix is called *Gaussian elimination.* This transformation requires finding a sequence of equivalent linear systems

$$A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}, \quad 0 \le k \le n-1$$

where $A^{(0)} = A, \mathbf{b}^{(0)} = \mathbf{b}$ and $A^{(n-1)}$ is an upper triangular matrix. The i-th equation and $(i+1)$-th equation is separated by a single row operation.

Remark. Fix $k > 1$ (the case $k-1 = 0$ is trivial). If $a_{kk}^{(k-1)} \neq 0$, add a multiple $-a_{ik}^{(k-1)}/a_{kk}^{k-1}$ of k-th row to the i-th row for $i = k+1, \ldots, n$. Then $a_{ik}^{k} = 0$ for $i = k+1, \ldots, n$.

Remark. The value $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ gets stored in the ik-position (if no pivoting).

Definition 2.4. Assuming no pivoting is necessary, Gaussian elimination reduces to

$$A^{n-1} = M_{n-1} \cdots M_1 A^{(0)}.$$

where $m_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$ and

$$M_k = \begin{bmatrix} 1 & & & & 0 \\ & 1 & & & \\ & & 1 & & \\ & & -m_{k+1,k} & 1 & \\ & & \vdots & & \ddots \\ 0 & & -m_{n,k} & 0 & 1 \end{bmatrix}.$$

Let $U = A^{(n-1)}$. U is upper triangular with non-zero diagonal elements. Then

$$A = M_1^{-1}M_2^{-1} \cdots M_{n-1}^{-1}U.$$

Now,

$$M_k^{-1} = \begin{bmatrix} 1 & & & & & 0 \\ & 1 & & & & \\ & & 1 & & & \\ & & m_{k+1,k} & 1 & & \\ & & \vdots & & \ddots & \\ 0 & & m_{n,k} & 0 & & 1 \end{bmatrix}.$$

Let $L = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$. Then

$$L = \begin{bmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ m_{31} & m_{32} & 1 & & \\ \vdots & \vdots & & \ddots & \\ m_{n1} & m_{n2} & \cdots & \cdots & 1 \end{bmatrix}.$$

and $A = LU$. The product LU the *LU factorization of A*. The matrix L is a unit lower-triangular matrix.

Remark. Let $\mathbf{y}$ be the solution of $L\mathbf{y} = \mathbf{b}$. Since $L = M_1^{-1} M_2^{-1} \cdots M_{n-1}^{-1}$,

$$y = M_{n-1} \cdots M_1 \mathbf{b}.$$

Solving for $\mathbf{y}$ is equivalent to performing elimination steps on $\mathbf{b}$. Then we only need to solve $U\mathbf{x} = \mathbf{y}$ to obtain $\mathbf{x}$. Since $\mathbf{x}$ is upper-triangular we only need to perform back subtitution.

Consider solving $A\mathbf{x} = \mathbf{b}$ for an $n \times n$ matrix using Gaussian elimination.

Step	Multiplies (Scaling)	Multiplies (Elimination)	Additions (Eliminations)
$A^{(0)} \to A^{(1)}$	$n-1$	$(n-1)^2$	$(n-1)^2$
$A^{(1)} \to A^{(2)}$	$n-2$	$(n-2)^2$	$(n-2)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$A^{(n-3)} \to A^{(n-2)}$	2	4	4
$A^{(n-2)} \to A^{(n-1)}$	1	1	1

The total number of multiplication operations is

$$\sum_{j=1}^{n-1} j + \sum_{j=1}^{n-1} j^2 = \frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{6} \approx \frac{1}{3}n^3$$

while the total number of additions is

$$\sum_{j=1}^{n-1} j^2 = \frac{n(n-1)(2n-1)}{6} \approx \frac{1}{3}n^3.$$

Thus the total number of operations is $2n^3/3$.

Consider instead using the LU-factorization of A. For the forward elimination step ($L\mathbf{y} = \mathbf{b}$),

Solving	Multiplies	Additions
$\mathbf{y}_2$	1	1
$\mathbf{y}_3$	2	2
$\vdots$	$\vdots$	$\vdots$
$\mathbf{y}_{n-1}$	$n-2$	$n-2$
$\mathbf{y}_n$	$n-1$	$n-1$

the total number of operations is

$$\sum_{j=1}^{n-1} j + \sum_{j=1}^{n-1} j = \frac{n(n-1)}{2} + \frac{n(n-1)}{2} \approx n^2.$$

For the back substitution step,

Solving	Multiplies	Additions
$\mathbf{x}_n$	1	0
$\mathbf{x}_{n-1}$	2	1
$\vdots$	$\vdots$	$\vdots$
$\mathbf{x}_2$	$n-1$	$n-2$
$\mathbf{x}_1$	n	$n-1$

the total number of operations is

$$\sum_{j=1}^{n} j + \sum_{j=0}^{n-1} j = \frac{n(n+1)}{2} + \frac{n(n-1)}{2} \approx n^2.$$

Therefore, the LU-factorization method requires $2n^2$ operations.

2.2 Interchanging

Theorem 2.5. Let U be an equivalent, upper-triangular form of A, that is,

$$U = (M_{n-1}P_{n-1}) \cdots (M_1 P_1)A,$$

where P_k is either the identity matrix if no interchanging occurs in the k-th step or P_k just interchanges row k with row I for some $I > k$.

Theorem 2.6. Suppose $k > l$ and P_k interchanges rows k and I where $I > k$. Then $P_k M_l = \widetilde{M}_l P_k$ where $\widetilde{M}_l P$ is the same as M_l except hte multiplies m_{kl} and m_{Il} have been interchanged.

$$P_k = \begin{bmatrix} 1 & & & \\ & 0 & 1 & \\ & 1 & 0 & \\ & & & 1 \end{bmatrix} \qquad P_k M_l = \begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & m_{Il} & 0 & 1 & \\ & m_{kl} & 1 & 0 & \\ & & & & 1 \end{bmatrix}$$

Definition 2.7. Let the matrix $\hat{M}_l$ be the same as M_l, except all the multiplies in the i-th columns have been interchanged by the P_k's for $k > l$. Then, $U = (\hat{M}_{n-1} \cdots \hat{M}_1)(P_{n-1} \cdots P_1)A = L^{-1}P^\top A$. Then, $A = PLU$. This is called the *PLU factorization* of A. Note that $P^\top A = LU$, so it also encodes the LU factorization of $(P_{n-1} \cdots P_1)A$ which is just A wiht its rows permuted.

2.3 Pivoting

Definition 2.8. In elimination, a *pivotal equation* is the equation used to elimination an unknown from the other equations. At the start of the k-th elimination step, a pivotal equation is the equation with a non-zero coefficient for x_k in the k-th, $k+1$-th, …, n-th equations.

Theorem 2.9. A is invertible if and only if there is at least one pivotal equation at every elimination step.

Remark. Pivoting can be viewed as multiplying A by a permutation matrix $P^\top$, and then finding the LU-factorization of $P^\top A$. Then, $A = PLU$.

Theorem 2.10. Every invertible matrix A can be written as a product PLU where P is a permutation matrix, L is a unit lower-triangular matrix and U is an (invertible) upper triangular matrix.

Theorem 2.11. An invertible matrix A has an LU-factorization if and only if each of the upper left hand submatrices

$$A_k = \begin{bmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{bmatrix}$$

for $k = 1, \dots, n$ are invertible.

Remark. In practice, not every pivot equation is good for numerical calculations

(i) Do not choose near-zero pivots.

(ii) Cannot just use absolute comparison of $a_{ik}^{(k-1)}$.

(iii) The best pivot maximizes the ratio of the size of pivot entry to the size of the row.

Remark. Suppose we are on the k-th step of Gaussian Elimination (where $1 \leq k \leq n-1$). The current matrix looks like

$$A^{(k-1)} = \begin{bmatrix} a_{11}^{(k-1)} & & \cdots & & a_{1n}^{(k-1)} \\ & \ddots & & & \\ & & a_{kk}^{(k-1)} & & \vdots \\ & & \vdots & \ddots & \\ & & a_{nk}^{(k-1)} & \cdots & a_{nn}^{(k-1)}. \end{bmatrix}$$

Which entries $a_{kk}^{(k-1)}, \cdots, a_{nk}^{(k-1)}$ should we use as the k-th pivot element?

Definition 2.12. The technique of *simple pivoting* involves choosing the pivot row with the smallest $I \geq k$ for which $A_{Ik}^{(k-1)} \neq 0$, and interchanging the k-th row and the I-th row.

Definition 2.13. The technique of *partial pivoting* involves choosing the pivot row with the entry $\left|a_{Ik}^{(k-1)}\right|$ that is the largest of $\left|a_{kk}^{(k-1)}\right|, \left|a_{k+1,k}^{(k-1)}\right|, \cdots, \left|a_{nk}^{(k-1)}\right|$, and interchanging the k-th row and the I-th row.

Definition 2.14. The technique of *scaled partial pivoting* involves computing scale factors for each row:

$$d_i = \max_{1 \leq j \leq n} |a_{ij}| \quad \text{for} \quad i = 1, \dots, n$$

before elimination procedure begins and interchanging them when rows are interchanged. At the k-th step, the pivot row for which $a_{Ik}^{(k-1)}/d_I$ is the maximized for all $I \geq k$, is chosen, and the k-th and I-th row are interchanged. Alternatively, the scale factors can be recomputed at every step.

Definition 2.15. In *total pivoting*, the columns are also interchanged. At the k-th step, choose $I \geq k$ and $J \geq k$ for which $|a_{IJ}^{(k-1)}|$ is the maximum of $|a_{ij}|$ for $i = k, \dots, n$ and $j = k, \dots, n$. Interchange the k-th row and the I-th row and interchange the k-th column and the J-th column.

Lemma 2.16. The operation counts of each pivoting strategy are as follows:

(i) partial pivoting: $\sum_{k=1}^{(n-1)}(n-k) \approx n^2/2$,

(ii) scaled pivoting (without updating scale factors): $n(n-1)+\sum_{k=1}^{(n-1)}[(n-k+1)+(n-k)] \approx 2n^2$,

(iii) scaled pivoting (updating scale factors): $\sum_{k=1}^{(n-1)}[(n-k+1)(n-k)+(n-k+1)+(n-k)] \approx n^3/3$,

(iv) total pivoting: $\sum_{k=1}^{n-1}\left[(n-k+1)^2-1\right] \approx n^3/3$.

2.4 Vector Norms on $\mathbb{R}^n$ and $\mathbb{C}^n$

Definition 2.17. A *norm* on a vector space is a function that maps a vector, $\mathbf{x} \in \mathcal{V}$, to a number and is denoted by $||\mathbf{x}||$. A norm must satisfy the following properties for all $\mathbf{x}, \mathbf{y} \in \mathcal{F}^n$ and $\alpha \in \mathcal{F}$ where $\mathcal{F}$ is $\mathbb{R}$ or $\mathbb{C}$:

(i) $||\mathbf{x}|| \geq 0$; $||\mathbf{x}|| = 0$ if and only if $\mathbf{x} = \mathbf{0}$,

(ii) $||\alpha\mathbf{x}|| = |\alpha| \cdot ||\mathbf{x}||$,

(iii) $||\mathbf{x}+\mathbf{y}|| \leq ||\mathbf{x}|| + ||\mathbf{y}||$ (triangle inequality).

Remark. Common examples of vector norms include:

(i) $||\mathbf{x}||_1 = \sum_{1\leq j\leq n} |x_j|$,

(ii) $||\mathbf{x}||_2 = \left(\sum_{j=1}^{n} |a_j|^2\right)^{1/2}$,

(iii) $||\mathbf{x}||_\infty = \max_{1\leq j\leq n} |a_j|$.

Definition 2.18. The set of $n \times n$ matrices is itself a vector space. A norm on this vector space satisfies for matrices $A, B \in \mathcal{F}^{n\times n}$ and $\alpha \in \mathcal{F}$ where $\mathcal{F}$ is $\mathbb{R}$ or $\mathbb{C}$:

(i) $||A|| \geq 0$ and $||A|| = 0$ if and only if A is the 0 matrix,

(ii) $||\alpha A|| = |\alpha| \cdot ||A||$,

(iii) $||A+B|| \leq ||A|| + ||B||$.

We call the norm a *matrix norm* if in addition we have

$$||AB|| \leq ||A|| \cdot ||B||.$$

Definition 2.19. Given a vector norm on $\mathbb{R}^n$ (or $\mathbb{C}^n$), the *operator norm induced by vector norm*, or just *operator norm*, on $n \times n$ matrices is

$$||A|| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{||A\mathbf{x}||}{||\mathbf{x}||}.$$

Informally, this norm gives the maximum stretch factor when $\mathbf{x}$ is mapped through A. For $p = 1, 2, \infty$, we call the operator norm induced by $||\cdot||_p$ also $||A||_p$.

Theorem 2.20. For $p = 1$ and $p = \infty$, there are explicit expressions for $||A||_1$ and $||A_\infty||$.

$$||A||_1 = \max_{1\leq j\leq n} \sum_{j=1}^{n} |a_{ij}| \qquad ||A||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|$$

Definition 2.21. Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$ where $\mathbf{x} = (x_1, x_2, \ldots, x_n)^\top$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^\top$. We recall the familiar *scalar product*, or dot product given by

$$\mathbf{x}^\top\mathbf{y} = x_1y_1 + x_2y_2 + \cdots + x_ny_n.$$

Lemma 2.22. For all vectors $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ in $\mathbb{R}^n$ and for all scalars α:

(i) $\mathbf{x}^\top\mathbf{y} = \mathbf{y}^\top\mathbf{x}$,

(ii) $(\alpha \mathbf{x})^\top \mathbf{y} = \alpha(\mathbf{x}^\top y)$,

(iii) $(\mathbf{x} + \mathbf{y})^\top \mathbf{z} = \mathbf{x}^\top z + \mathbf{y}^\top z$,

(iv) $\mathbf{x}^\top \mathbf{x} \geq 0$ where $\mathbf{x}^\top \mathbf{x} = 0$ if and only if $\mathbf{x} = 0$

Theorem 2.23 (The Cauchy-Schwarz Inequality). Given any $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$, $|\mathbf{x}^\top \mathbf{y}| \leq ||\mathbf{x}||_2 ||\mathbf{y}||_2$.

Theorem 2.24. The operator norm $||A||_2$ is the square root of the largest eigenvalue of $A^H A$.

Definition 2.25. We say a matrix norm $||\cdot||_m$ is *compatible* with a vector norm $||\cdot||_v$ if for all $A \in \mathcal{F}^{m \times n}$ and $\mathbf{x} \in \mathcal{F}^n$, $||A\mathbf{x}||_v \leq ||A||_m \cdot ||\mathbf{x}||_v$.

Definition 2.26. Define the Frobenius norm of A to be

$$||A||_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}.$$

Theorem 2.27. The Frobenius norm of A is compatible with $||\mathbf{x}||_2$.

2.5 Residual Error

Definition 2.28. Consider $A\mathbf{x} = \mathbf{b}$. Let $\mathbf{x}$ be the true solution and let $\hat{\mathbf{x}}$ be the approximate solution. Define $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ be the *error vector* and let $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}} = A\mathbf{x} - A\hat{\mathbf{x}} = A\mathbf{e}$ be the *residual vector.*

Theorem 2.29. For all n-vector $\mathbf{y}$ for an invertible matrix A such that $A\mathbf{x} = \mathbf{b}$,

$$\frac{||\mathbf{y}||}{||A^{-1}||} \leq ||A\mathbf{y}|| \leq ||A|| \cdot ||\mathbf{y}||.$$

Definition 2.30. Define $\kappa(A) = ||A|| \cdot ||A^{-1}||$ to be the *condition number of* A when $\kappa(A) \geq 1$.

Theorem 2.31. The relative error of $||\mathbf{e}||/||\mathbf{x}||$ is as large as $\kappa(A) \cdot ||\mathbf{r}||/||\mathbf{b}||$.

Remark. Method for iteratively solving for the solution of a linear system. Consider the origin matrix A. To find $A\hat{\mathbf{x}}$ set $\mathbf{r} = \mathbf{b} - A\hat{\mathbf{x}}$ and solve $A\mathbf{e} = \mathbf{r}$. Call the computed solution $\hat{\mathbf{e}}$. Then $||\hat{\mathbf{e}}||/||\hat{\mathbf{x}}||$ is approximately $||\mathbf{e}||/||\mathbf{x}||$, e.g. if $||\hat{\mathbf{e}}||/||\hat{\mathbf{x}}|| \approx 10^{-s}$, then we expect $\hat{\mathbf{x}}$ has approximately s significant digits as an approximation to $\hat{\mathbf{x}}$. Also expect that $\hat{\mathbf{e}}$ has s significant digits as an approximation to $\mathbf{e}$, but the absolute error in $\hat{\mathbf{e}}$ is much smaller than the absolute error in $\hat{\mathbf{x}}$. If $||\hat{\mathbf{e}}||/||\hat{\mathbf{x}}||$ sufficienty small, then $\hat{\mathbf{x}} + \hat{\mathbf{e}}$ is the approximate solution. Else set $\hat{\mathbf{x}}' = \hat{\mathbf{x}} + \hat{\mathbf{e}}$ and repeat the procedure. Solving successive systems is not very expensive since elimination required $2/3n^3$ and each solve requires $2n^2$.

Definition 2.32. The method of *backward error analysis* involves considering the approximation to be the exact solution of a perturbed system. Let $\hat{\mathbf{x}}$ be the approximate solution of $A\mathbf{x} = \mathbf{b}$ and consier $\hat{\mathbf{x}}$ to be the exact solution of $\hat{A}\mathbf{x} = \mathbf{b}$ where $\hat{A} = A - E$ for some matrix E. Then a bound on E can be found to analyze its effect on $\hat{\mathbf{x}}$ as an approximation to $\mathbf{x}$.

Theorem 2.33. In general, the bound on the error in $\hat{\mathbf{x}}$ relative to $\hat{\mathbf{x}}$ is

$$\frac{||\mathbf{x} - \hat{\mathbf{x}}||}{||\hat{\mathbf{x}}||} \leq \kappa(A) \cdot \frac{||E||}{||A||}.$$

Theorem 2.34. Let $\hat{\mathbf{x}}$ be the computed PLU solution of a linear system and the exact solution of $(A + PE)\hat{\mathbf{x}} = \mathbf{b}$ for some $n \times n$ matrix E. Let $u = n \cdot 1.01 \cdot u$ where u is the unit roundoff. If

$$|e_{ij}| \leq u_n |(P^\top A)_{ij}| + u_n(3 + u_n) \sum_{k=1}^{n} |\hat{l}_{ik} \cdot |\hat{u}_{kj}$$

then the following is usually true,

$$||E|| \leq n \cdot u \cdot ||A|| \quad \text{and} \quad \frac{||\mathbf{x} - \hat{\mathbf{x}}||}{||\hat{\mathbf{x}}||} \leq \kappa(A) \cdot n \cdot u.$$

Remark. If $\kappa(A)$ is large in the above formula, the system is ill-conditioned, although we must compare to u since this definition changes with precision. Let $s = -\log_\beta(\kappa(A) \cdot n \cdot u)$. Then this method gets us approximately s significant digits in $\hat{\mathbf{x}}$ and each successive iteration gets about s more significant digits.

2.6 General Iterative Methods

Definition 2.35 (General Iterative Method). Let M be a real $n \times n$ matrix, and let $\mathbf{x}^{(0)}$ be a vector in $\mathbb{R}^n$. Generate a sequence of vector $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ by setting

$$\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{g} \quad \text{for} \quad k = 0, 1, 2, \ldots$$

where $\mathbf{g}$ is a given fixed vector in $\mathbb{R}^n$.

Lemma 2.36. If $\mathbf{x}^{(k)} \to \hat{\mathbf{x}}$ as $k \to \infty$, then $\hat{\mathbf{x}} = M\hat{\mathbf{x}} + \mathbf{g}$, so $\hat{\mathbf{x}}$ is a solution of the linear system $(I - M)\hat{\mathbf{x}} = \mathbf{g}$.

Theorem 2.37. Let $||\cdot||$ be a vector norm on $\mathbb{R}^n$, and let $\alpha = ||M||$, the matrix norm of M subordinate to the vector norm $||\cdot||$. Suppose $\alpha = ||M|| < 1$. Then

(i) $I - M$ is invertible,

(ii) For any choice of $\mathbf{x}^{(0)}$, the sequence $\mathbf{x}^{(k)}$ generated by $\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{g}$ converges to $\hat{\mathbf{x}}$, i.e. $\mathbf{x}^{(k)} \to \hat{\mathbf{x}}$ as $k \to \infty$.

(iii) If $\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \hat{\mathbf{x}}$, then $||\mathbf{e}^{(k)}|| \leq \alpha^k ||\mathbf{e}^{(0)}||$.

This theorem is a special case of the Contraction Mapping Fixed Point Theorem.

Definition 2.38 (Splitting Methods). Choose matrices N and P for which $A = N - P$, and consider the iteration

$$N\mathbf{x}^{(k+1)} = P\mathbf{x}^{(k)} + \mathbf{b} \quad \text{for} \quad k = 0, 1, 2, \ldots.$$

We want to choose N and P so that (i) N is invertible, (ii) $N\mathbf{x} = \mathbf{b}$ is easy to solve, and (iii) $||N^{-1}P|| < 1$ in some norm. Analytically, the iteration is the same as $\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{g}$ where $M = N^{-1}P$ and $\mathbf{g} = N^{-1}\mathbf{b}$ (multiply original iteration by N^{-1}). Each iteration is solving the linear system $N\mathbf{x} = \mathbf{w}$ for $\mathbf{x}^{(k+1)}$ where $\mathbf{w} = P\mathbf{x}^{(k)} + \mathbf{b}$.

Lemma 2.39. For the methods descibed above,

(i) if the iteration converges, i.e. $x^{(k)}$ converges, it converges to a solution of $A\mathbf{x} = \mathbf{b}$,

(ii) if N is invertible and $||N^{-1}P|| < 1$ (in some matrix norm subordinate to a vector norm on $\mathbb{R}^n$), the iteration converges to the unique solution of $A\mathbf{x} = \mathbf{b}$.

Definition 2.40 (Jacobi's Method). Given an $n \times n$ matrix A, let

$$L = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{bmatrix}, \qquad U = \begin{bmatrix} 0 & a_{12} & \cdots & a_{1n} \\ 0 & 0 & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Then $A = L + D + U$. Choose $N = D$ and $P = -(L+U)$. Jacobi's method involves iteratively applying the following

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

This is equivalent to the equation:

$$x_i^{(k+1)} = \left(b_i - \sum_{j<i} a_{ij} x_j^{(k)} - \sum_{j>i} a_{ij} x_j^{(k)} \right) / a_{ii}$$

for $1 \le i \le n$ and $k = 0, 1, \ldots$.

Definition 2.41. A matrix is called *(strictly row) diagonally dominant* if

$$|a_{ii}| > \sum_{j \neq i} |a_{ij}| \qquad \text{for} \qquad 1 \le i \le n.$$

Theorem 2.42. If A is diagonally dominant, then Jacobi's Method converges.

Definition 2.43 (Gauss-Seidel). From the decompositon in Jacobi's method, choose $N = D + L$ and $P = -U$ and iteratively compute:

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

In the kth iteration (computing $\mathbf{x}^{(k+1)}$ from $\mathbf{x}^{(k)}$), this system for $\mathbf{x}^{(k+1)}$ is solved by forward substitution.

$$x_i^{(k+1)} = \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j>i} a_{ij} x_j^{(k)} \right) / a_{ii}$$

for $1 \le i \le n$ and $k = 0, 1, \ldots$.

Remark. For Gauss-Seidel, only one vector is needed to store $\mathbf{x}^{(k)}$ and $\mathbf{x}^{(k+1)}$ since $\mathbf{x}$ can be overwritten in-place.

Theorem 2.44. If A is diagonally dominant, then Gauss-Seidel converges, that is, for any choice of $\mathbf{x}^{(0)}$, the sequence $\mathbf{x}^{(k)}$ generated by $(D+L)\mathbf{x}^{(k+1)} = -U\mathbf{x}^{(k)} + b$ converges to the unique solution of $A\mathbf{x} = \mathbf{b}$.

Definition 2.45. A real $n \times n$ matrix is called *symmetric positive definite*, or just positive definite, if A is symmetric, i.e. $A^\top A$ and for all $\mathbf{x} \neq \mathbf{0}$, $\mathbf{x}^\top A\mathbf{x} > 0$.

Theorem 2.46. A real symmetric $n \times n$ matrix is positive definite if and only if all of its eigenvalues are positive.

Theorem 2.47. If A is symmetic positive definite, then Gauss-Seidel converges.

Remark. Usually Gauss-Seidel converges to the true solution faster than Jacobi's method.

Definition 2.48 (Successive Over-Relaxation (SOR)). This is a variant of Gauss-Seidel. Rewrite the Gauss-Seidel iteration as

$$x_i^{(k+1)} = x_i^{(k)} + \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j \geq i} a_{ij} x_j^{(k)} \right) / a_{ii}.$$

Fix an ω where $0 < \omega < 2$. The SOR iteration is

$$x_i^{(k+1)} = x_i^{(k)} + \omega \left(b_i - \sum_{j<i} a_{ij} x_j^{(k+1)} - \sum_{j \geq i} a_{ij} x_j^{(k)} \right) / a_{ii}.$$

When $0 < \omega < 1$, it is called under-relaxation; when $\omega = 1$, it is Gauss-Seidel; when $1 < \omega < 2$, it is called over-relaxation. In matrix form,

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} + \omega D^{-1} \left(\mathbf{b} - L\mathbf{x}^{(k+1)} - (D+U)\mathbf{x}^{(k)} \right) \\
(D + \omega L)\mathbf{x}^{(k+1)} &= D\mathbf{x}^{(k)} + \omega(\mathbf{b} - (D+U)\mathbf{x}^{(k)}) \\
\mathbf{x}^{(k+1)} &= (D + \omega L)^{-1}((1-\omega)D - \omega U)\mathbf{x}^{(k)} + \omega(D + \omega L)^{-1}\mathbf{b} \\
\mathbf{x}^{(k+1)} &= M_\omega \mathbf{x}^{(k)} + \mathbf{g}_\omega
\end{aligned}$$

2.7 Linear Least Squares

Definition 2.49 (Linear Least Squares). Often times the linear system $A\mathbf{x} = \mathbf{b}$ where A is an $m \times n$ real matrix and $\mathbf{b} \in \mathbb{R}^m$ has no solution since $m > n$. The range of A has dimension less than or equal to $n < m$ so it is a proper subspace of $\mathbb{R}^m$ and there are many $\mathbf{b} \in \mathbb{R}^m$ for which no solution exists. Instead, we find a vector $\mathbf{x} \in \mathbb{R}^n$ that minimizes

$$||\mathbf{e}||_2^2 = ||A\mathbf{x} - \mathbf{b}||_2^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j - b_i \right)^2,$$

the sum of the squares of the error terms.

Theorem 2.50. Let Y be a subspace of $\mathbb{R}^m$ and let $\mathbf{b} \in \mathbb{R}^m$. Then there is a unqiue closest element $\hat{\mathbf{y}}$ of Y to $\mathbf{b}$ in the 2-norm $|| \cdot ||_2$, i.e. $||\mathbf{b} - \hat{\mathbf{y}}||_2 \leq ||\mathbf{b} - \mathbf{y}||_2$ for all $\mathbf{y} \in Y$ and $||\mathbf{b} - \hat{\mathbf{y}}||_2 < ||\mathbf{b} - \mathbf{y}||_2$ for $y \neq \hat{\mathbf{y}}$. Moreover, $\mathbf{b} - \hat{\mathbf{y}}$ is orthogonal to Y i.e. $(\mathbf{b} - \hat{\mathbf{y}})^\top \mathbf{y} = 0$ for all $\mathbf{y} \in Y$.

Theorem 2.51 (The Normal Equations). Given a real $m \times n$ matrix A, vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^m$ and $\mathbf{x} \in \mathbb{R}^n$ minimizes $||A\mathbf{x} - \mathbf{b}||_2^2$ if and only if $\mathbf{x}$ is a solution of the normal equations

$$A^\top A\mathbf{x} = A^\top \mathbf{b}.$$

Remark. Computation concerns with linear least squares:

(i) The normal equations are often very ill-conditioned in the 2-norm, $\kappa(A^\top A) = \kappa(A)^2$, so it is not always best to use the normal equations.

(ii) Better numerical methods for linear least squares problems: QR factorization (closely related to Gram-Schmidt), Singular Value Decomposition (for ill-conditioned problems).

3 Solutions of Non-Linear Systems

3.1 Methods for Solving Non-Linear Systems

Definition 3.1. A real number x for which $f(x) = 0$ is called a *root* of that equation; x is called a *emph* of f.

Definition 3.2 (General Methods of Solving Linear Equations). To solve a non-linear equation, write it in the form $f(x) = 0$, assuming that f is a continuous real-valued function that is defined on some interval $I \in \mathbb{R}$. In practice, locate approximately a zero s of the given function f. We want to find an x such that $|x - s|$ is small or $|f(x)|$ is small.

Theorem 3.3. If f is continuous on $[a, b]$ and $f(a)f(b) < 0$, then there exists an $s \in (a, b)$ for which $f(s) = 0$.

Definition 3.4 (Bisection Method). The *bisection method* is a bracketing method where at each step in the iteration, we have an interval $[a, b]$ in which f as a 0. Start with an interval $[a, b]$ that brackets a zero of f, i.e. $f(a)f(b) < 0$. For each step, shrink the length of the interval by a fator of 2 while still bracketing a zero of f. The bisection method is guaranteed to converge, but it has a slow convergence rate, approximately 3 iterations per decimal digit of accuracy.

Definition 3.5 (Newton's Method). Start with an approximation x_0 to s. Iteratively, find the zero of the tangent line to the graph of f at $(x_n, f(x_n))$ to get x_{n+1}. Converges rapidly if it converges, so it needs to start sufficiently close to the zero and we ened to be able to evaluate f', i.e. computable and $f'(s) \neq 0$.

Definition 3.6 (Secant Method). Start with two approximations x_{n-1} and x_n to s. Find the zero of the secant line joining the two previous points $(x_{n-1}, f(x_{n-1}))$, $(x_n, f(x_n))$. Similar to Newton's method with a slower convergence, but f' is not required to evaluate the derivative f'.

Remark. Ideally, we would like the dependability of the bisection method and the speed of Newton. For example, *Regular Falsi* (see text) is a bracketing method similar to the secant method. Often, one endpoint converges quickly to a zero of f. *Brent's Method* (also called the *Brent-Dekker method*) is a combination of bisection, secant, and inverse quadratic interpolation that converges rapidly.

3.2 Fixed-Point Iteration

Remark. Many iterative methods, e.g. Newton's method, can be viewed as $x_{n+1} = g(x_n)$ where g is some particular function.

Definition 3.7. For a function g, a *fixed point* of g is a point x where $g(x) = x$.

Theorem 3.8. If $x_{n+1} = g(x_n)$ where g is continuous and x_n converges to a number ζ in the domain of g, then $g(\zeta) = \zeta$, i.e. ζ is a fixed point.

Theorem 3.9. Let g be a continuous function on a closed bounded interval $I = [a, b]$, and suppose for all $x \in I$, $g(I) \in I$, i.e. g maps I to itself. Then g has at least one fixed point in I.

Theorem 3.10 (Contraction Mapping Fixed-Point Theorem, Differentiable Functions). Suppose g is differentiable on a closed, bounded interval $I = [a, b]$, that g maps I to itself, and for some $L < 1$, $|g'(x)| \leq L < 1$ for all $x \in I$. Then the following are true:

(i) g has a unique fixed point in I; call it ζ,

(ii) for any $x_0 \in I$, $x_{n+1} = g(x_n)$ generates a sequence such that $x_n \to \zeta$,

(iii) if $e_n = x_n - \zeta$, then

$$|e_n| \leq \frac{L^n}{1-L}|x_1 - x_0|.$$

Corollary 3.11 (Local Convergence Theorem). Suppose g is continuously differentiable in an open interval I containing a fixed point ζ, and suppose $|g'(\zeta)| < 1$. Then there exists an $\epsilon > 0$, so that when $|x_0 - \zeta| \leq \epsilon$, the fixed-point iteration $x_{n+1} = g(x_n)$ yields a sequence x_n with $x_n \to \zeta$.

Definition 3.12. Let $x_0, x_1, x_2, \ldots$ be a sequence which converges to a number ζ. Let $e_n = x_n - \zeta$. If there is a number $p \geq 1$ and a constant $C \neq 0$ for which

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C$$

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

Definition 3.13. For specific values of p and C we assign specific names to the order of convergence:

(i) if $p = 1$ and $C = 1$, convergence is called *sub-linear*;

(ii) if $p = 1$ and $0 < C < 1$, convergence is called *linear*;

(iii) if $\lim_{n\to\infty} |e_{n+1}|/|e_n| = 0$, convergence is called *super-linear*;

(iv) if $p = 2$, convergence is called *quadratic*.

Definition 3.14. A function $f \in C^k$ on an interval $[a, b]$ where k is a non-negative integer when $f, f', f'', \ldots, f^{(k)}$ are all defined and continuous on $[a, b]$. In the case of $k = 0$, f is continuous. In the case of $k = 1$, f is continuously differentiable.

Theorem 3.15 (Taylor's Theorem with Remainder). If $f \in C^{k+1}$ then for each x, there exists a ζ between a and x for which

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(k+1)}(\zeta)}{(k+1)!}(x-a)^{k+1}.$$

Theorem 3.16. Suppose $g \in C^{k+1}$, $g(s) = s$, x_n is generated by $x_{n+1} = g(x_n)$ and $x_n \to s$, and $g'(s) = g''(s) = \cdots = g^{(k)}(s) = 0$ and $g^{(k+1)}(s) \neq 0$. Then $x_n \to s$ to order $k+1$ with an asymptotic error constant of $|g^{(k+1)}(s)|/(k+1)!$.

Theorem 3.17. Suppose $f \in C^3$, $f(s) = 0$, $f'(s) \neq 0$, and x_n is generated by Newton's method $x_{n+1} = x_n - f(x_n)/f'(x_n)$. Then

(i) if $x_n \to s$, convergence is at least quadratic,

(ii) if x_0 is close enough to s, then $x_n \to s$.

4 Approximation Theory and Interpolation

4.1 Polynomials

Definition 4.1. A (real) *polynomial* is a function $p : \mathbb{R} \to \mathbb{R}$ of the form

$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

If $a_n \neq 0$, we define the *degree* of $p(x)$ to be n. If $p(x) = 0$, $\deg(p) = -\infty$.

Lemma 4.2. Let $p(x)$ and $q(x)$ be polynomials. Then $p(x)q(x)$ is also a polynomial and

$$\deg(pq) = \deg(p) + \deg(q).$$

Theorem 4.3 (Euclidean Algorithm). Suppose $p(x)$ and $d(x)$ are polynomials of degree at least 0. Then there exist polynomials $q(x)$ and $r(x)$ such that

$$p(x) = q(x)d(x) + r(x)$$

where $\deg(r) < \deg(d)$. The polynomials $q(x)$, $d(x)$, and $r(x)$ are called the *quotient*, *divisor*, and *remainder*.

Corollary 4.4. If $\deg(p) \geq 1$ and $p(x_1) = 0$, then there exists a polynomial $q(x)$ such that $p(x) = q(x)(x-x_1)$ where $\deg(q) = \deg(p) - 1$.

Definition 4.5. A number x_1 is called a *zero of* p *with multiplicity* m if

$$p(x_1) = p'(x_1) = \cdots = p^{(m-1)}(x_1) = 0 \neq p^{(m)}(x_1).$$

Theorem 4.6. If x_1 is a zero of multiplicity m, then there exists a polynomial $q(x)$ such that $p(x) = q(x)(x - x_1)^m$ and $q(x_1) \neq 0$.

Corollary 4.7. If $x_1, \ldots, x_k$ are zeros of p with multiplicities $m_1, \cdots, m_k$, then there exists a polynomial $q(x)$ such that

$$p(x) = q(x)(x - x_1)^{m_1}(x - x_2)^{m_2} \cdots (x - x_k)^{m_k}.$$

Corollary 4.8. If $p(x)$ is a polynomial of degree less than or equal to n and $p(x)$ has at least $n + 1$ zeroes (counting multiplicites), then $p = 0$.

Theorem 4.9. Given a real polynomial $p(x)$ with degree $n \geq 1$, there exists at least one value r (possibly complex) such that $p(r) = 0$.

Theorem 4.10 (Fundamental Theorem of Algebra). Given a real polynomial $p(x)$ with degree $n \geq 1$, $p(x)$ can be written as

$$p(x) = a_0(x - r_1)(x - r_2) \cdots (x - r_n)$$

where $r_1, \ldots, r_n$ are the zeros of $p(x)$. Moreover, the set of zeros is unique.

Definition 4.11 (Synthetic Division). Let $p(x)$ be a polynomial given by

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

where $a_n \neq 0$, and let α be constant. If we let $b_0 = a_0$ and generate $\{b_j\}_{j=1}^n$ by

$$b_j = \alpha b_{j-1} + a_j, \qquad 1 \leq j \leq n,$$

then $p(\alpha) = b_n$.

4.2 Interpolation by Polynomials

Definition 4.12. Given a set of points $(x_0, y_0), (x_1, y_1), \ldots$, the method of *interpolation* involves finding a function $p(x)$ for which $p(x_i) = y_i$. The function $p(x)$ is called an *interpolant*. Often $y_i = f(x_i)$ for some unknown function $f(x)$, so we say that the interpolant is used as an approximation to f.

Lemma 4.13. If $f(x)$ is a function such that $f(x_i) = y_i$, $0 \leq i \leq n$, then it has in $\mathbb{P}_n$ an interpolating polynomial of the form

$$p(x) = \sum_{j=0}^{n} f(x_j)\ell_j(x)$$

where $\ell_j(x)$ is

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

The form of $p(x)$ above is called the *Lagrange form* of $p(x)$.

Lemma 4.14. If $f(x)$ is a function such that $f(x_i) = y_i$, $0 \leq i \leq n$, then has in $\mathbb{P}_n$ an interpolating polynomial

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

whose cofficients a_j can be computed by solving

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & & \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

or $V\mathbf{a} = \mathbf{y}$ where V is called a *Vandermonde matrix*.

Theorem 4.15 (Polynomial Interpolation). If $x_0, x_1, \ldots, x_n$ are distinct, then for arbitrary real $y_0, y_1, \ldots, y_n$, there exists a unique polynomial $p(x)$ of degree less than or equal to n such that $p(x_i) = y_i$.

Definition 4.16. Suppose $x_0, x_1, \ldots, x_k$ are distinct and $f(x_0), f(x_1), \ldots, f(x_k)$ are given. Define the *k-th divided difference* $f[x_0, x_1, \ldots, x_k]$ to be the coefficient of x^k in the unique polynomial $p_k(x)$ of degree less than or equal to k which interpolates f at $x_0, x_1, \ldots, x_k$.

Theorem 4.17. For $k \geq 1$, we have a recursive formula for k-th divided difference

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}.$$

4.3 Approximation Theory

Definition 4.18 (Approximation Theory). Suppose $f(x)$ is a function defined on $[a, b]$ that we wish to approximate (perhaps f is unknown or it method of computation is exhaustive). We would prefer finding a function $g(x)$ so that $g(x) \approx f(x)$ (based on some measure of closeness) such that $g(x)$ is easy to compute (at least for $x \in [a, b]$).

Definition 4.19. Given functions f and g that are continuous and real-valued on some closed finite interval $[a, b]$, we can define *function norms* to measure "closeness" between these two functions. Common norms include extension of the ℓ_p vector norms:

$$||f - g||_1 = \int_a^b |f(x) - p(x)|w(x)\,dx,$$

$$||f - g||_2 = \left(\int_a^b (f(x) - p(x))^2 w(x)\,dx\right)^{1/2},$$

$$||f - g||_\infty = \max_{a\leq x\leq b} |f(x) - g(x)|.$$

For the 1- and 2-norm, we can define a weighting function $w(x)$ that provides some flexibility in measuring closeness. The weighting function must be continuous and nonnegative on (a, b). It is common to let $w(x) = 1$ so that no region on $[a, b]$ is weighted more than the other.

Remark. In the case of functions, $||\cdot||_1$, $||\cdot||_2$, and $||\cdot||_\infty$ are norms on the ∞-dimensional vector space $C[a, b]$ (the set of continuous real-valued functions on $[a, b]$).

Remark. Typical application of approximation theory: given a continuous function $f \in C[a, b]$ and some finite dimensional subspace M of $C[a, b]$ (e.g. $M = \mathbb{P}_n$ for some fixed n), find the closest function $\hat{g} \in M$ for which $||f - \hat{g}|| \leq ||f - g||$ for all $g \in M$ in some norm on $C[a, b]$. Often, we would like to minimize $||f - g||_\infty$ over all $g \in \mathbb{P}_n$.

Theorem 4.20 (Weierstrass). Let $f \in C[a, b]$. For each $\epsilon > 0$ there exists a polynomial $p(x)$ of degree N_ϵ (N_ϵ depends on ϵ) such that $||f - p||_\infty < \epsilon$.

Theorem 4.21. Given $f \in C[a, b]$ and given an integer $n \geq 0$, there exists a unique polynomial $p \hat{+} n \in \mathbb{P}_n$ for which $||f - \hat{p_n}||_\infty \leq ||f - p_n||_\infty$ for all $p_n \in \mathbb{P}_n$.

Definition 4.22. We call $\hat{p}$ in Theorem 4.21, the *best n-th degree uniform approximation to* $f(x)$ and call $E_n(f) = ||f - \hat{p}_n||_\infty$ the *degree of approximation to* $f(x)$.

Remark. Theorem 4.20 and Theorem 4.21 state that any continuous function on an interval $[a, b]$ can be approximated uniformly by a polynomial and for any fixed degree k, there exists a unique, closest polynomial approximation to f.

4.4 Error of Polynomial Interpolation

Lemma 4.23. Suppose f has k continuous derivatives. Let $x_0, \ldots, x_k \in \mathbb{R}$ be distinct. Then there exists some ξ between $\min\{x_1, \ldots, x_k\}$ and $\max\{x_1, \ldots, x_k\}$ such that $f[x_0, \ldots, x_k] = f^{(k)}(\xi)/k!$.

Lemma 4.24. Suppose f has k continuous derivatives. Let $x_0, \ldots, x_k \in \mathbb{R}$ be distinct and let $x \neq x_i$ $(0 \leq i \leq n)$. If p is an approximation to f defined by

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}),$$

then

$$f(x) = p_n(x) + f[x_0, \ldots, x_n, x](x - x_0) \cdots (x - x_n).$$

Theorem 4.25. Suppose $f \in C^{n+1}[a, b]$ and $x_0, \ldots, x_n \in \mathbb{R}$ are distinct in $[a, b]$. If p is an approximation to f defined by

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}),$$

then for each $x \in [a, b]$, there exists a $\xi \in [a, b]$ such that

$$f(x) = p_n(x) + f^{(n+1)}(\xi)/(n + 1)!(x - x_0) \cdots (x - x_n).$$

Corollary 4.26. If $f(x) = p(x) + f^{(n+1)}(\xi)/(n + 1)!(x - x_0) \cdots (x - x_n)$, then

$$|f(t) - p_n(t)| \leq \frac{M_{n+1}}{(n + 1)!}|W(t)|$$

where $M_{n+1} = \max_{x \in [a,b]} |f^{(n+1)(x)}$ and $W(t) = (t - x_0) \cdots (t - x_n)$.

4.5 Taylor Polynomials

Definition 4.27. Suppose $f(x) \in \mathbb{C}^{n+1}[a,b]$, that is, $f(x)$ and its first $n+1$ derivatives are continuous on $[a,b]$ and uppose for some $c \in [a,b]$ we know the values $f(c), f'(c), \ldots, f^{(n)}(c)$. Then we can approximate f on $[a,b]$ by an n-th degree Taylor polynomial centered at c:

$$p_n(x) = f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n.$$

Definition 4.28. Under the assumptions on f above, Taylor's Theorem with remainder states that for any $x \in [a,b]$, there exists a ξ between x and c such that

$$p_n(x) = f(c) + f'(c)(x-c) + \frac{f''(c)}{2!}(x-c)^2 + \cdots + \frac{f^{(n)}(c)}{n!}(x-c)^n + \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}.$$

Subtracting by $p_n(x)$ above, we get the *error equation* for the Taylor polynomial p_n:

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x-c)^{n+1}.$$

Using the infinity-norm we can get a maximum value of the error equation,

$$||f - p_n||_\infty = \max_{a \le x \le b} |f(x) - p_n(x)|.$$

If we let $M_{n+1} = \max_{a \le x \le b} |f^{(n+1)}(x)|$, then for any $x \in [a,b]$,

$$|f(x) - p_n(x)| \le \frac{M_{n+1}}{(n+1)!}|(x-c)^{n+1}|.$$

This is called a *pointwise upper bound* for the error function $f(x) - p_n(x)$. To get an upper bound for $||f - p_n||_\infty$, let $d = \max(c-a, b-c)$, that is, d is the largest distance $|x - c|$ from a point $x \in [a,b]$ to c, so

$$||f - p_n||_\infty = \max_{a \le x \le b} |f(x) - p_n(x)| \le \frac{M_{n+1} d^{n+1}}{(n+1)!}.$$

4.6 Chebyshev Polynomials

Definition 4.29 (Chebyshev Polynomials of the First Kind). For $k = 0, 1, 2, \ldots$ define $T_k(x) = \cos(k \cos^{-1} x)$ for $-1 \le x \le 1$ (using the principal branch of $\cos^{-1} x$). Then $T_0(x) = \cos 0 = 1$, $T_1(x) = \cos(\cos^{-1} x) = x$, and so on. These polynomials are called *Chebyshev polynomials of the first kind*. They can be computed by a recursion formula:

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x).$$

Clearly $T_k(x)$ has degree k for $k \ge 0$, so by induction on the recusion formula, the coefficient of x^k in $T_k(x)$ is 2^{k-1} for $k \ge 1$. Because cosine of odd multiples is $\pi/2$, we can find (for $k \ge 1$), k distinct zeros of $T_k(x)$ in the interval $(-1, 1)$, and by the Fundamental Theorem of Algebra,

$$T_k(x) = 2^{k-1}(x - x_0)(x - x_1) \cdots (x - x_{k-1}).$$

Lemma 4.30. On the interval $-1 \le x \le 1$, $|T_k(x)| \le 1$.

Lemma 4.31. For a fixed $k \ge 1$, let $y_j = \cos(j\pi/k)$ for $j = 0, 1, \ldots, k$. Then $1 = y_0 > y_1 > \cdots > y_k = -1$ and

$$T_k(y_j) = \cos(k(j\pi/k)) = \cos(j\pi) = (-1)^j.$$

Then there are $k+1$ points where $|T_k(x)|$ takes on its maximum and the sign of T_k alternates at these $k+1$ points. These points are called the *Chebyshev nodes*.

Theorem 4.32. Let $W(x) = (x-x_0)\cdots(x-x_n)$ be the function described in Corollary 4.26, fixing the interval $[a,b]$ to be $[-1,1]$. Then the set of points $x_0,\ldots,x_n \in [-1,1]$ that minimizes $||W||_\infty = \max_{-1\le x\le 1}|W(x)|$ are the zeroes of $T_{n+1}(x)$:

$$x_j = \cos\left(\frac{j+1/2}{n+1}\pi\right), \qquad j = 0,1,\ldots,n.$$

Then $W(x) = T_{n+1}(x)/2^n$ and $||W||_\infty = 1/2^n$.

Corollary 4.33. If $f \in C^{n+1}[-1,1]$ and we interpolate f at the Chebyshev nodes (the zeroes of T_{n+1}), then

$$||f-p_n||_\infty \le \frac{M_{n+1}}{(n+1)!}||W||_\infty \le \frac{M_{n+1}}{2^n(n+1)!}.$$

Corollary 4.34. Let $f \in \mathbb{C}^{n+1}[a,b]$ and let t be a variable in $[-1,1]$, and let x be a variable in $[a,b]$ related by

$$x = \frac{b-a}{2}t + \frac{a+b}{2}, \qquad t = 2\frac{x-a}{b-a} - 1.$$

Define a shifted Chebyshev polynomial $\hat{T}_k(x)$ on $[a,b]$ by

$$\hat{T}_k(x) = T_k(t) = T_k\left(2\cdot\frac{x-a}{b-a} - 1\right).$$

For $k \ge 1$, the coefficient of x^k in $\hat{T}_k(x)$ is $2^{k-1}(2/(b-a))^k$, and the Chebyshev nodes for $\hat{T}_{n+1}$ are

$$x_j = \frac{b-a}{2}\cos\left(\frac{j+1/2}{n+1}\pi\right) + \frac{a+b}{2}, j = 0,1,\ldots,n.$$

Then

$$W(x) = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}\hat{T}_{n+1}(x),$$

so

$$||W||_\infty = \max_{a\le x\le b}|W(x)| = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}.$$

If a polynomial $p_n(x)$ of degree n interpolates f at the Chebyshev nodes $x_0,\ldots,x_n$, then

$$||f-p_n||_\infty \le \frac{M_{n+1}}{(n+1)!}\frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}$$

where $M_{n+1} = ||f^{(n+1)}||_\infty$.

4.7 Equal-Spaced and Osculatory Interpolation

Definition 4.35 (Equal-Spaced Interpolation). Suppose $f(x)$ is defined on $[a,b]$ and n is a positive integer. Let $h = (b-a)/n$ and $x_i = a + ih$ where $i = 0,1,\ldots,n$. Then $x_0 = a, x_1 = a+h, \ldots, x_2 = a+2h, \cdots, x_n = a+nh = b$ are equally spaced. For fixed h, define $\Delta f(x) = f(x+h) - f(x)$ which we call the *forward difference of* f. Define $\Delta^2 f(x) = (\Delta(\Delta f))(x) = f(x+2h) - 2f(x+h) + f(x)$. Recursively define

$$\Delta^k f(x) = (\Delta(\Delta^{k-1}f))(x) = \Delta^{k-1}f(x+h) - \Delta^{k-1}f(x).$$

By induction, we can write $\Delta^k f(x)$ as

$$\Delta^k f(x) = \sum_{j=0}^{k}\binom{k}{j}(-1)^{k-j}f(x+jh).$$

By induction it can be shown that

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{\Delta^k f(x_i)}{k!h^k}.$$

Remark. Often, a *forward difference table* is used instead of a divided difference table to interpolate f.

Definition 4.36. Let $x_0, x_1, \ldots, x_n$ be not necessarily distinct points in $[a, b]$. To say that a polynomial $p(x)$ *interpolates* f *at* $x_0, \ldots, x_n$ means for each distinct number α in $x_0, \ldots, x_n$, let k_α be the number of times α appears in the list, then

$$p^{(j)}(\alpha) = f^{(j)}(\alpha) \quad \text{for} \quad j = 0, 1, \ldots, k_{\alpha-1}$$

Theorem 4.37 (Osculatory Interpolation). Let $x_0, x_1, \ldots, x_n$ be not necessarily distinct points in $[a, b]$, and suppose for each distinct α in the list, $f^{(j)}(\alpha)$ is defined for $j = 0, \ldots, k_{\alpha-1}$ (where k_α is defined above). Then there exists a unique polynomial $p_n(x)$ of degree $d \leq n$ which interpolates f at $x_0, \ldots, x_n$.

Definition 4.38. The value $f[x_0, \ldots, x_k]$ is defined to be the coefficient of x^k in the unique polynomial $p_k(x)$ which interpolates f at $x_0, \ldots, x_k$.

Theorem 4.39. Let $p(x)$ be a polynomial that interpolates f at $x_0, \ldots, x_k$. Then the following are true

(i) when $x_0 \neq x_k$, the recursive formula still holds:

$$f[x_0, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0},$$

(ii) if $f \in C^k$, then $f[x_0, \ldots, x_k]$ is a continuous function of the $k+1$ variables $x_0, \ldots, x_k$,

(iii) $f[c, c, \ldots, c] = f^{(k)}(c)/k!$ for some $c \in [a, b]$,

(iv) the formula for $p_n(x)$ still holds:

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \cdots (x - x_{n-1}),$$

(v) the error formula still holds: if $f \in C^{n+1}[a, b]$, and $x_0, \ldots, x_n \in [a, b]$, then for each $x \in [a, b]$, then there exists ξ such that

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W(x)$$

where $W(x) = (x - x_0) \cdots (x - x_n)$.

Definition 4.40 (Cubic Hermite Interpolation). Let $f \in \mathbb{C}^1[a, b]$ be some function, α, β be distinct points, and consider finding $p_3(x)$ which interpolates f and f' at α and β, i.e. p_3 interpolates f at $\alpha, \alpha, \beta, \beta$. Then the formula for $p_3(x)$ is given by

$$\begin{aligned} p_3(x) &= f[\alpha] + f[\alpha, \alpha](x - \alpha) + f[\alpha, \alpha, \beta](x - \alpha)^2 + f[\alpha, \alpha, \beta, \beta](x - \alpha)^2(x - \beta) \\ &= f(\alpha) + f'(\alpha)(x - \alpha) + f[\alpha, \alpha, \beta](x - \alpha)^2 + f[\alpha, \alpha, \beta, \beta](x - \alpha)^2(x - \beta). \end{aligned}$$

4.8 Piecewise Polynomial Interpolation and Approximation

Definition 4.41. A *piecewsise-polynomial* function of order k on $[a, b]$ with interior breakpoints at $x_1, \ldots, x_{n-1}$ is a function of the form

$$S(x) = \begin{cases} S_0(x), & x \in [x_0, x_1), \\ S_1(x), & x \in [x_1, x_2), \\ \vdots & \\ S_{n-1}(x), & x \in [x_{n-1}, x_n] \end{cases}$$

where each $S_j(x)$ is a polynomial of degree at most k, that is,

$$S_j(x) = c_{0j} + c_{1j}x + \cdots + c_{kj}x^k.$$

For integers $0 \leq m \leq k$, define $\mathbb{PP}_k^m$ to be the set of all piecewise polynomial functions of order k which are in $C^m[a, b]$.

Lemma 4.42. If $m \geq k$, then $\mathbb{PP}_k^m = \mathbb{P}_k$.

Definition 4.43. The points $x_0, x_1, \ldots, x_n$ (endpoints and interior breakpoints) are called *knots*. Elements of $\mathbb{PP}_k^{k-1}$ (here $m = k - 1$) are called *splines* of order k. Splines are the smoothest (most continuous derivatives) piecewise polynomials which are not just single polynomial functions.

Theorem 4.44. Given $S \in \mathbb{PP}_k^m$, S is in C^m, if at each of the $n-1$ interior breakpoints x_j

$$S_{j-1}^{(n)}(x_j) = S_j^{(n)}(x_j) \qquad n = 0, 1, \ldots, m$$

for $j = 1, \ldots, n-1$.

Remark. The dimension of $\mathbb{PP}_k^m$ tells us how many "free parameters" there are in S. To determine S, we need a number of conditions equal to the number of free parameters; moreover, these conditions need to be linearly independent.

Definition 4.45. For $\mathbb{PP}_1^c$, *piecewise-linear interpolation* involves solving $S(x_j) = f(x_j)$ for $0 \leq j \leq n$ where each point is connected by a line.

Remark. The following describes piecewise-linear interpolation. Fix j with $0 \leq j \leq n-1$. Then $S_j(x) = c_{0j} + c_{1j}x$. Use the boundary conditions $S_j(x_j) = f(x_j)$ and $S_{j+1}(x_{j+1}) = f(x_{j+1})$ to solve the for the unknowns. In Newton's form, $S_j(x) = f(x_j) + f[x_j, x_{j+1}](x - x_j)$ where

$$c_{0j} = f(x_j) - x_j f[x_j, x_{j+1}], \qquad c_{1j} = f[x_j, x_{j=1}].$$

Theorem 4.46. Given that the conditions for piecewise-polynomial interpolation hold true, for piecewise-linear interpolation, there exists a unique interpolant for arbitrary f.

Lemma 4.47. For piecewise-linear interpolation, if S is the interpolant to f then for $x_j \leq x \leq x_{j+1}$,

$$|f(x) - S(x)| = |f(x) - S_j(x)| \leq \frac{||f''||_\infty}{2}|(x - x_j)(x - x_{j+1})|.$$

In general, along $[x_0, x_n]$,

$$||f - S||_\infty \leq \frac{M_2}{8}h^2.$$

where $M_2 = ||f''||_\infty$.

Definition 4.48. For $\mathbb{PP}_3^1$, *piecewise-cubic Hermite interpolation* involves solving $S(x_j) = f(x_j)$ and $S'(x_j) = f'(x_j)$ for $0 \leq j \leq n$.

Remark. The following describes piecewise-cubic Hermite interpolation. Fix j with $0 \leq j \leq n-1$. Then $S_j(x)$ is the polynomial of degree at most 3 which interpolates f at x_j, x_j, x_{j+1} and x_{j+1} using $f(x_j)$, $f'(x_j)$, $f(x_{j+1})$, and $f'(x_{j+1})$. Use the boundary conditions $S_j(x_j) = f(x_j)$, $S'_{j+1}(x_{j+1}) = f'(x_{j+1})$, $S_j(x_j) = S_{j+1}(x_{j+1}) = f(x_j)$, and $S'_{j+1}(x_{j+1}) = f'(x_{j+1})$ to solve the for the unknowns.

Theorem 4.49. Given that the conditions for piecewise-polynomial interpolation hold true, for piecewise-cubic Hermite interpolation, there exists a unique interpolant for arbitrary f.

Lemma 4.50. For piecewise-cubic Hermite interpolation, if S is the interpolant to f then for $x_j \leq x \leq x_{j+1}$,

$$|f(x) - S(x)| = |f(x) - S_j(x)| \leq \frac{||f^{(4)}||_\infty}{4!}|(x-x_j)^2(x-x_{j+1})^2|.$$

In general, along $[x_0, x_n]$,

$$||f - S||_\infty \leq \frac{M_4}{384}h^4.$$

where $M_4 = ||f^{(4)}||_\infty$.

Definition 4.51. For $\mathbb{PP}_3^2$, *natural cubic spline interpolation* involves boundary conditions. If the function f is interpolated on $[a, b]$, then $S''(a) = 0$, $S''(b) = 0$ and $S(x_j) = f(x_j)$ for $0 \leq j \leq n$.

Remark. The following describes natural cubic spline interpolation. Let $y_0'', y_1'', \ldots, y_n''$ represent $S''(x_0), S''(x_1), \ldots, S''(x_n)$. Let $f_j = f(x_j)$ for $0 \leq j \leq n$ and $\Delta f_j = f_{j+1} - f_j$ and $0 \leq j \leq n-1$. For $j = 0, \ldots, n-1$, we will express S_j in terms of f_j, $fj+1$ and y_j'', y_{j+1}''. Let $h_j = \Delta x_j = x_{j+1} - x_j$ for $0 \leq j \leq n$.

Fix j with $0 \leq j \leq n-1$ and use $S_j(x_j) = f_j$, $S_j''(x_j) = y_j''$, $S_j(x_{j+1}) = S_{j+1}(x_{j+1}) = f_{j+1}$, and $S_j''(x_{j+1}) = S_{j+1}''(x_{j+1}) = y_{j+1}''$ to uniquely determine S_j. Each polynomial $S_j(x)$ is given by

$$S_j(x) = \frac{y_j''}{6h_j}(x_{j+1}-x)^3 + \frac{y_{j+1}''}{6h_j}(x-x_j)^3 + \left(\frac{f_{j+1}}{h_j} - \frac{y_{j+1}''h_j}{6}\right)(x-x_j) + \left(\frac{f_j}{h_j} - \frac{y_j''h_j}{6}\right)(x_{j+1}-x).$$

The piecwise polynomial function $S(x)$ built from each $S_j(x)$ is piecewise-cubic. To solve for the unknowns $y_0'', y_1'', \ldots, y_n''$, use the derivatives of $S_j(x)$,

$$\begin{aligned}
S_{j-1}'(x) &= -\frac{y_{j-1}''}{2h_{j-1}}(x_j - x)^2 + \frac{y_j''}{2h_{j-1}}(x-x_{j-1})^2 + \left(\frac{f_j}{j_{j-1}} - \frac{y_j''h_{j-1}}{6}\right) - \left(\frac{f_{j-1}}{h_{j-1}} - \frac{y_{j-1}''h_{j-1}}{6}\right) \\
S_j'(x) &= -\frac{y_j''}{2h_j}(x_{j+1}-x)^2 + \frac{y_{j+1}''}{2h_j}(x-x_j)^2 + \left(\frac{f_{j+1}}{h_j} - \frac{y_{j+1}''h_j}{6}\right) - \left(\frac{f_j}{h_j} - \frac{y_j''h_j}{6}\right)
\end{aligned}$$

Evaluating both at x_j we get

$$h_{j-1}y_{j-1}'' + 2(h_j + h_{j-1})y_j'' + h_j y_{j+1}'' = b_j$$

where $b_j = 6(\Delta f_j/h_j - \Delta f_{j-1}/h_{j-1})$ which holds for $1 \leq j \leq n-1$. The boundary conditions $S''(x_0) = S''(x_n) = 0$ imples that $y_0'' = y_n'' = 0$. Thus solving for $y_1'', \ldots, y_{n-1}''$ involves solving the linear system

$$\begin{bmatrix} \gamma_1 & h_1 & 0 & \cdots & 0 & 0 \\ h_1 & \gamma_2 & h_2 & \cdots & 0 & 0 \\ 0 & h_2 & \gamma_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \gamma_{n-2} & h_{n-2} \\ 0 & 0 & 0 & \cdots & h_{n-2} & \gamma_{n-1} \end{bmatrix} \begin{bmatrix} y_1'' \\ y_2'' \\ y_3'' \\ \vdots \\ v_{n-2}'' \\ v_{n-1}'' \end{bmatrix} = \begin{bmatrix} b_1 - h_0 y_0'' \\ b_2 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} - h_{n-1}y_n'' \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-2} \\ b_{n-1} \end{bmatrix}$$

where $\gamma_j = 2(h_j + h_{j-1})$. Since $\gamma_j > h_{j-1} + h_j$ the matrix is diagonally dominant which implies that the matrix is invertible. Moreover the system does not require interchanges to solve by Gaussian elimination.

Theorem 4.52. Given that the conditions for piecewise-polynomial interpolation hold true, for natural cubic spline interpolation, there exists a unique interpolant for arbitrary f.

Lemma 4.53. For natural cubic spline interpolation, if S is the interpolant to f then

$$||f - S||_\infty \leq C_0||f''||_\infty h^2$$

for some constant C_0 that is independent of f and h where $h = \max_{0\leq j\leq n-1} \Delta x_j$. Additionally,

$$||f' - S'||_\infty \leq C_1||f''||_\infty h$$

for some constant C_1. If $f''(a) =''(b) = 0$ and $f \in C^4[a, b]$, then

$$||f - S||_\infty \leq C_2||f^{(4)}||_\infty h^4$$

and

$$||f' - S''||_\infty \leq C_3||f^{(4)}||_\infty h^3$$

for some constants C_2, C_3.

Definition 4.54. For $\mathbb{PP}_3^2$, *complete cubic spline interpolation* is similar to natural cubic spline interpolation, except $S'(a) = f'(a)$ and $S'(b) = f'(b)$ are used as the boundary conditions.

Remark. Modifying the steps from natural cubic spline interpolation, set $S_0'(x_0) = f_0$ and $S_{n-1}'(x_n) = f_n'$ which leads to

$$\begin{cases} 2h_0y_0'' + h_0y_1'' = b_0, & \text{where } b_c = 6\left(\dfrac{\Delta f_0}{h_0} - f_0'\right), \\ h_{n-1}y_{n-1}' + 2h_{n-1}y_n'' = b_n, & \text{where } b_n = 6\left(f_n' - \dfrac{\Delta f_{n-1}}{h_{n-1}}\right). \end{cases}$$

The new system is diagonally dominan, so the matrix is invertible.

Theorem 4.55. Given that the conditions for piecewise-polynomial interpolation hold true, for complexte cubic spline interpolation, there exists a unique interpolant for arbitrary f.

Lemma 4.56. For complete cubic spline interpolation, if S is the interpolant to f and $f \in \mathbb{C}^4[a, b]$ then

$$||f - S||_\infty \leq \frac{5}{384}||f^{(4)}||_\infty h^4$$

where $h = \max_{0\leq j\leq n-1} \Delta x_j$. Additionally,

$$||f' - S'||_\infty \leq \frac{1}{24}||f^{(4)}||_\infty h^3.$$

Definition 4.57. For $\mathbb{PP}_3^2$, *"not a knot" cubic spline interpolation* is similar to natural cubic spline interpolation, except $S_0'''(x_1) = S_1'''(x_1)$ and $S_{n-2}'''(x_{n-1}) = S_{n-1}'''(x_{n-1})$ are used as the boundary conditions. The error bound is similar to that for complete cubic spline interpolation.

Theorem 4.58. Among the class of all functions $g(x) \in \mathbb{C}^2[a, b]$ which interpolates f at $x_0, x_1, \ldots, x_n$, the unique one which minimizes $\int_a^b (g''(x))^2\, dx$ is the natural cubic spline $S(x)$.

Theorem 4.59. Among the class of all functions $g(x) \in \mathbb{C}^2[a, b]$ which interpolates f at $x_0, x_0, x_1, \ldots, x_{n-2}, x_{n-1}, x_n, x_n$, the unique one which minimizes $\int_a^b (g''(x))^2\, dx$ is the complete cubic spline $S(x)$.

5 Numerical Integration

5.1 Overview

Definition 5.1. The following describes *numerical integration.* Suppose $f \in C[a,b]$ and we know $f(x_0), \cdots, f(x_n)$ for some points $x_0, \ldots, x_n \in [a,b]$ where $a \leq x_0 < x_1 < \ldots < x_n \leq b$. Certain choices of points x_j can lead to very accurate approximations to $\int_a^b f(x)\,dx$. If $g(x)$ is an approximation to $f(x)$ on $[a,b]$, we can consider $\int_a^b g(x)\,dx$ as an approximation to $\int_a^b f(x)\,dx$. Often $g(x)$ is a polynomial interpolant of $f(x)$.

Definition 5.2. If the interval $[a,b]$ is known and fixed, let $I(f) = \int_a^b f(x)\,dx$. Then I is a function whose domain is $C[a,b]$, a set consisting itself of function. We call I and *operator* or a *mapping.* The domain of I is $C[a,b]$ and the range of I is $\mathbb{R}$.

Definition 5.3. Any formula which approximates $I(f)$ using values of f is called a *numerical integration formula* (or a *quadrature formula*). Any quadrature formula can also be thought of as a *mapping* $Q(f)$ which assigns to each function $f \in C[a,b]$ a real number $Q(f)$. Quadrature formulas obtained by an interpolating polynomial are called *interpolatory quadrature.*

Definition 5.4. Let $a \leq x_0 < x_1 < \cdots < x_n \leq b$ be all fixed, and let $Q_n(f)$ be the interpolatory quadrature given by $Q_n(f) = I(p_n)$ where $p_n(x)$ is the unique polynomial of degree $d \leq n$ which interpolates f at $x_0, \ldots, x_n$. If we write $p_n(x)$ in Lagrange form, we obtain

$$Q_n(f) = I(p_n) = \int_a^b p_n(x)\,dx = \int_a^b \sum_{j=0}^{n} f(x_j)\ell_j(x)\,dx = \sum_{j=0}^{n} \left(\int_a^b \ell_j(x)\,dx \right) f(x_j) = \sum_{j=0}^{n} A_j f(x_j).$$

where $A_j = \int_a^b \ell_j(x)\,dx$. We call the A_j's the *weights* and the x_j's the *nodes.*

Definition 5.5. Let Q be some quadature formula on $[a,b]$. If for some integer $k \geq 0$, $Q(p) = I(p)$ for all $p \in \mathbb{P}_k$ (i.e. for all polynomials of degree $d \leq k$). Then we say Q has *precision* (at least) k.

Theorem 5.6. Every $(n+1)$-point interpolatory quadrature has precision at least n.

Theorem 5.7. Let Q_n be the $(n+1)$-point interpolatory quadrature on $[a,b]$ with nodes $x_0, x_1, \ldots, x_n$. For $k = 0, 1, \ldots, n$, let $f_k(x) = x^k$. Since Q_n has precision at least n, $Q(n)(f_k) = I(f_k)$ for $k = 0, 1, \ldots, n$. Then we have a linear system where $A_0, A_1, \ldots, A_n$ are the unknowns:

$$\sum_{j=0}^{n} x_j^k A_j = \int_A^b x^k\,dx, \qquad 0 \leq k \leq n.$$

The matrix in this system is a Vandermonde matrix, so the system can be solved.

Definition 5.8. The *closed Newton-Cotes Formulas* are obtained using interpolatory quadrature with equally spaced nodes $x_0, x_1, \ldots, x_n$ with $x_0 = a$ and $x_n = b$ where

$$x_j = a + jh, \qquad j = 0, 1, \ldots, n, \qquad h = \frac{b-a}{n}.$$

The *open Newton-Cotes Formulas* are obtained using interpolatory quadrature with equally spaced nodes $y_1, y_2, \ldots, y_{n+1}$ with $a < y_1$ and $y_{n+1} < b$ where

$$x_j = a + jh, \qquad j = 1, 2, \ldots, n+1, \qquad h = \frac{b-a}{n+2}.$$

Remark. Some Newton-Cotes Formulas on $[-1,1]$:

Closed	n	x_j	A_j	Formula
Trapezoid Rule	1	$x_0=-1, x_1=1$	$A_0=1, A_1=1$	$Q_1(f)=f(-1)+f(1)$
Simpson's Rule	2	$x_0=-1, x_1=0, x_2=1$	$A_0=\frac{1}{3}, A_1=\frac{4}{3}, A_2=\frac{1}{3}$	$Q_2(f)=\frac{1}{3}f(-1)+\frac{4}{3}f(0)+\frac{1}{3}f(1)$
	3	$x_0=-1, x_1=-\frac{1}{3},$	$A_0=\frac{1}{4}, A_1=\frac{3}{4},$	$Q_3(f)=\frac{1}{4}f(-1)+\frac{3}{4}f(-1/3)$
		$x_2=\frac{1}{3}, x_3=1$	$A_2=\frac{3}{4}, A_3=\frac{1}{4}$	$+\frac{3}{4}f(1/3)+\frac{1}{4}f(1)$
Open	n	x_j	A_j	Formula
Midpoint Rule	0	$y_1=0$	$A_1=2$	$Q_1(f)=2f(0)$
	1	$y_1=-\frac{1}{3}, y_2=\frac{1}{3}$	$A_1=1, A_2=1$	$Q_1(f)=f(-1/3)+f(1/3)$
	2	$y_1=-\frac{1}{2}, y_2=0, y_3=\frac{1}{2}$	$A_1=\frac{4}{3}, A_2=-\frac{2}{3}, A_3=\frac{4}{3}$	$Q_2(f)=\frac{4}{3}f(-1/2)-\frac{2}{3}f(0)+\frac{4}{3}f(1/2)$

Each of these formulas can be transformed to a formula on an arbitrary interval $[a,b]$. Using $t\in[-1,1]$ as the variable on $[-1,1]$ and $x\in[a,b]$ on $[a,b]$, let $x=\alpha t+\beta$ where $\alpha=(b-a)/2$ and $\beta=(b+a)/2$. Then we have

$$x_j=\frac{b-a}{2}t_j+\frac{a+b}{2} \qquad \text{and} \qquad y_j=\frac{b-a}{2}u_j+\frac{a+b}{2}.$$

Additionally, the A_j's get multiplied by a factor of α since

$$\int_a^b f(x)\,dx=\int_{-1}^{-1}\alpha f(\alpha t+\beta)\,dt.$$

For example,

$$\begin{aligned}
&\text{Trapezoid Rule} && T(f)=\frac{b-a}{2}(f(a)+f(b)),\\
&\text{Simpson's Rule} && S(f)=\frac{b-a}{2}\left[\frac{1}{3}f(a)+\frac{4}{3}f\left(\frac{a+b}{2}\right)+\frac{1}{3}f(b)\right],\\
&\text{Midpoint Rule} && M(f)=\frac{b-a}{2}\left[2f\left(\frac{a+b}{2}\right)\right].
\end{aligned}$$

Remark. We can also apply different rules within a given interval $[a,b]$. Partition $[a,b]$ into N subintervals $a=x_0<x_1<\cdots<x_N=b$, and one of these rules is applied in each subinterval $[x_j,x_{j+1}]$ for $j=0,\cdots N-1$. Then

$$\begin{aligned}
&\text{Trapezoid Rule} && T_{x_j}^{x_{j+1}}(f)=\frac{h_j}{2}(f(x_j)+f(x_{j+1})),\\
&\text{Simpson's Rule} && S_{x_j}^{x_{j+1}}(f)=\frac{h_j}{2}\left[\frac{1}{3}f(x_j)+\frac{4}{3}f\left(\frac{x_j+x_{j+1}}{2}\right)+\frac{1}{3}f(x_{j+1})\right],\\
&\text{Midpoint Rule} && M_{x_j}^{x_{j+1}}(f)=\frac{h_j}{2}\left[2f\left(\frac{x_j+x_{j+1}}{2}\right)\right]
\end{aligned}$$

where $h_j=x_{j+1}-x_j$.

Theorem 5.9. Let Q_n be the $(n+1)$-point interpolating quadrature on $[a,b]$ with nodes $x_0,x_1,\dots,x_n$. Let $f\in C^{n+1}[a,b]$ and let $e_n(f)=I(f)-Q_n(f)$, the error in $Q_n(f)$. Since for each $x\in[a,b]$ there is a ξ for which

$$f(x)-p_n(x)=\frac{f^{(n+1)}(\xi)}{(n+1)!}W(x)$$

where $W(x)=(x-x_0)\cdots(x-x_n)$. Then

$$e_n(f)=I(f)-I(p_n)=\int_a^b f(x)-p_n(x)\,dx=\int_a^b\frac{f^{(n+1)}(\xi)}{(n+1)!}W(x)\,dx$$

and thus

$$|e_n(f)| \leq \frac{M_{n+1}}{(n+1)!} \int_a^b |W(x)|\, dx$$

where $M_{n+1} = \max_{a \leq x \leq b} |f^{(n+1)}(x)|$.

Remark. More useful forms for $e_n(f)$ can be derived for many quadrature formulas. For the Trapezoid Rule $T(f) = (b-a)(f(a) - f(b))/2$, then

$$e^T(f) = I(f) - T(f) = -\frac{f''(\eta)}{12}(b-a)^3$$

for some $\eta \in [a,b]$.

Definition 5.10. The method of *composite numerical integration* involves subdividing an interval $[a,b]$ into N subintervals by choosing $x_0, \ldots, x_N$ with $a = x_0 < \cdots < x_N = b$, and applying a quadrature formula in each subinterval $[x_j, x_{j+1}]$ for $j = 0, \ldots, N-1$.

Remark. Examples:

$$\text{Composite Trapezoid Rule} \qquad T_N(f) = \sum_{j=0}^{N-1} T_{x_j}^{x_{j+1}}(f) = \sum_{j=0}^{N-1} \frac{h_j}{2}(f(x_j) + f(x_{j+1})),$$

$$\text{Composite Simpson's Rule} \qquad S_N(f) = \sum_{j=0}^{N-1} S_{x_j}^{x_{j+1}}(f) = \sum_{j=0}^{N-1} \frac{h_j}{6}\left[f(x_j) + 4\left(\frac{x_j + x_{j+1}}{2}\right) + f(x_{j+1})\right].$$

with $h_j = x_{j+1} - x_j$. With equally spaced points with $h = (b-a)/N$, $x_j = a + jh$, $0 \leq j \leq N$, then

$$T_N(f) = \sum_{j=0}^{N-1} \frac{h_j}{2}(f(x_j) + f(x_{j+1})) = \frac{h}{2}(f(x_0) + f(x_N)) + h\sum_{j=1}^{N-1} f(x_j),$$

$$S_N(f) = \sum_{j=0}^{N-1} \frac{h_j}{6}\left[f(x_j) + 4\left(\frac{x_j + x_{j+1}}{2}\right) + f(x_{j+1})\right] = \sum_{j=0}^{N-1} \frac{h}{6}\left[(f(x_0) + f(x_N)) + 2\sum_{j=1}^{N-1} f(x_j) + 4\sum_{j=0}^{N-1} f\left(\frac{x_j + x_{j+1}}{2}\right)\right].$$

Theorem 5.11. Let $f \in \mathbb{C}^2[a,b]$ and consider the composite quadrature obtained via the composite Trapezoid Rule with equally spaced points. Then the error $e_N^T(f) = I(f) - T_N(f)$ is

$$e_N^T(f) = \sum_{j=0}^{N-1} \left(I_{x_j}^{x_{j+1}}(f) - T_{x_j}^{x_{j+1}}(f)\right) = \sum_{j=0}^{N-1} \left(-\frac{f''(\eta_j)}{12}h^3\right)$$

for each $\eta \in [x_j, x_{j+1}]$. By the Intermediate Value Theorem, it can be shown that

$$h\sum_{j=0}^{N-1} f''(\eta_j) = (b-a)f''(\eta)$$

for some $\eta \in [a,b]$ and thus

$$e_N^T(f) = -\frac{f''(\eta)(b-a)h^2}{12}.$$

Definition 5.12. Suppose we have approximations $A(h)$ (one for each $h > 0$ in some sequence of h's tending to 0) to an unknown quantity, and suppose

$$a_0 = A(h) + a_k h^k + C_k(h) h^{k+1}$$

where k is a known positive integer, a_k is an unknown constant, and $C_k(h)$ is an unknown bounded function of h. Let r be some constant with $0 < r < 1$ (usually we take $r = 1/2$). Then

$$a_0 = A(rh) + a_k(rh)^k + C_k(rh)(rh)^{k+1}.$$

Combining the two equations,

$$a_0 = \frac{r^k A(h) - A(rh)}{r^k - 1} + \tilde{C}_k(h)h^{k+1}$$

where $\tilde{C}_k(h) = r^k/(r^k - 1)(C_k(r) - rC_k(rh))$ is another unknown bounded function of h. The error in $A(h)$ is $\mathcal{O}(h^k)$ but the error in

$$\frac{A(rh) - r^k A(h)}{1 - r^k}$$

is $\mathcal{O}(h^{k+1})$. This method is called *Richardson Extrapolation*.

Definition 5.13. Suppose we know more about the form of the error in $A(h)$, the better approximation we can get. Suppose we know that

$$a_0 = A(h) + a_1 h^{k_1} + a_2 h^{k_2} + \cdots + a_m h^{k_m} + C_m(h)h^{k_{m+1}}$$

where $k_1 < k_2 < \cdots < k_{m+1}$ are known, $a_1, \ldots, a_m$ are unknown, and $C_m(h)$ is unknown and bounded. Let $A_0(h) = A(h)$. Its leading error term is $\mathcal{O}(h^{k_1})$. Apply Richardson extrapolation to get $A_1(h) = \frac{A_0(rh) - r^{k_1} A_0(h)}{1 - r^{k_1}}$. Its leading error term is then $\mathcal{O}(h^{k_2})$. Apply Richardson extrapolation again to get

$$A_2(h) = \frac{A_1(rh) - r^{k_2} A_1(h)}{1 - r^{k_2}}.$$

Its leading error term is then $\mathcal{O}(h^{k_3})$. Repeating this method, we get $A_m(h)$ with error $\mathcal{O}(h^{k_{m+1}})$. This method is called *Repeated Richardson extrapolation*.

Definition 5.14. *Romberg Integration* is the application of repeated Richardson Extrapolation to the Trapezoid Rule. It can be shown that if $f \in C^{\nu}[a, b]$, and we apply the Composite Trapezoid Rule to f with equally spaced points, then

$$I(f) = T_N(f) + c_2 h^2 + c_4 h^4 + \cdots + C_{\nu}(h)h^{\nu}$$

where $h = (b - a)/N$, $c_2, c_4, \ldots$ are unknown constants, and $C_{\nu}(h)$ is bounded and unknown. The constants can be computed by

$$\begin{aligned} c_2 &= -\frac{1}{12} \int_a^b f''(x)\, dx \\ c_4 &= \frac{1}{720} \int_a^b f^{(4)}(x)\, dx \\ &\vdots \end{aligned}$$

Use $r = 1/2$ and define for $m = 0, 1, 2, \ldots$ $T_{0,m} = T_{2^m}(f)$, i.e. split $[a, b]$ into $N = 2^m$ equal subintervals, so $h = (b-a)/2^m$. Fix m and let $h = (b-a)/2^m$ be fixed too. Then $T_{0,m}$ is the value of the composite Trapezoid Rule approximation where the subintervals have length h and $T_{0,m+1}$ is the value when the subintervals have length $(b - a)/2^{m+1} = h/2$. Applying Richardson Extrapolation, define

$$T_{1,m} = \frac{T_{0,m+1} - \frac{1}{4}T_{0,m}}{1 - \frac{1}{4}}.$$

The error in $T_{0,m}$ is $\mathcal{O}(h^2)$; the error in $T_{1,m}$ is $\mathcal{O}(h^4)$. Repeated Richardson extrapolation leads to

$$T_{i,m} = \frac{T_{i-1,m+1} - \left(\frac{1}{4}\right)^i T_{i-1,m}}{1 - \left(\frac{1}{4}\right)^i}$$

for $i = 1, 2, \ldots$.

Theorem 5.15. The function $T_{1,m}$ in Romberg Integration is $S_N(f)$, the composite Simpson's Rule with $N = 2^m$ and $h = (b-a)/2^m$.