Numerical Analysis 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

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

where $\left(. \, d_1 \, d_2 \cdots d_m\right)_\beta$ 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 $\text{fl}(x)$ 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* - $\text{fl}(x)$ 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* - $\text{fl}(x)$ 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 $\text{fl}(x)$ for an m-digit floating-point representation with a maximum M exponent is

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

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 $\text{fl}(x)$ as an approximation to x. Usually it is absolute error $x - \text{fl}(x)$.

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

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

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

Remark. The machine epsilon can also be defined to be $\varepsilon = \inf\{y > 0 : \text{fl}(1+y) > 1\}$. 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

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

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 $\text{fl}(p) \times \text{fl}(q)$ is

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

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 $1 \le k \le n-1$). 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 $I \ge k$ 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:

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

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 $I \ge k$, 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 $I \ge k$ and $J \ge k$ 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

$$\mathbf{x}^\top \mathbf{y} = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n.$$

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) $(\mathbf{x}+\mathbf{y})^\top\mathbf{z} = \mathbf{x}^\top\mathbf{z} + \mathbf{y}^\top\mathbf{z}$,
(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 $1 \leq i \leq n$ and $k = 0, 1, \ldots$.

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

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

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 $1 \leq i \leq n$ 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 *zero* 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 has a zero. 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:

$$\begin{aligned} ||f - g||_1 &= \int_a^b |f(x) - p(x)| w(x)\, dx, \\ ||f - g||_2 &= \left(\int_a^b (f(x) - p(x))^2 w(x)\, dx \right)^{1/2}, \\ ||f - g||_\infty &= \max_{a \leq x \leq b} |f(x) - g(x)|. \end{aligned}$$

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 $\hat{p_n} \in \mathbb{P}_n$ 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 *piecewise-polynomial* function of order k on $[a, b]$ with interior breakpoints at $x_1, \ldots, x_{n-1}$ is a function of the form

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

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 $S_j(x)$,

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

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

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

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

$$\begin{cases} 2h_0 y_0'' + h_0 y_1'' = b_0, & \text{where } b_c = 6\left(\dfrac{\Delta f_0}{h_0} - f_0'\right), \\ h_{n-1} y_{n-1}' + 2h_{n-1} y_n'' = b_n, & \text{where } b_n = 6\left(f_n' - \dfrac{\Delta f_{n-1}}{h_{n-1}}\right). \end{cases}$$

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 $g(x) \in C^2[a, b]$ 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 $g(x) \in C^2[a, b]$ which interpolates f at $x_0, x_0, x_1, x_1, \ldots, x_{n-1}, x_{n-1}, x_n, x_n$, 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 $a \le x_0 < x_1 < \ldots < x_n \le b$. 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 an *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 $a \le x_0 < x_1 < \cdots < x_n \le b$ 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 $d \le n$ 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 $k \ge 0$, $Q(p) = I(p)$ for all $p \in \mathbb{P}_k$ (i.e. for all polynomials of degree $d \le k$). 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:

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

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	$x_0=-1, x_1=-\frac{1}{3},$ $x_2=\frac{1}{3}, x_3=1$	$A_0=\frac{1}{4}, A_1=\frac{3}{4},$ $A_2=\frac{3}{4}, A_3=\frac{1}{4}$	$Q_3(f)=\frac{1}{4}f(-1)+\frac{3}{4}f(-1/3)$ $+\frac{3}{4}f(1/3)+\frac{1}{4}f(1)$
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 $x_0,x_1,\ldots,x_n$. 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

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

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

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 $h = (b-a)/N$, $x_j = a + jh$, $0 \le j \le N$, 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 $h = (b-a)/2^m$.

6 Eigenvalues and Eigenvectors

6.1 Review of Eigenvalues and Eigenvectors

Definition 6.1. Let A be an $n \times n$ matrix. A (complex) number λ is an *eigenvalue* of A if there exists a vector $\mathbf{x} \neq 0$ such that $A\mathbf{x} = \lambda\mathbf{x}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. The vector $\mathbf{x}$ is called an *eigenvector* of A associated with the eigenvalue λ.

Theorem 6.2. Let A be an $n \times n$ matrix. The following are equivalent:

(i) λ is an eigenvalue of A;

(ii) $\lambda I - A$ is not invertible;

(iii) $\det(\lambda I - A) = 0$.

Definition 6.3. From (iii) above, the expression $p(\lambda) = \det(\lambda I - A)$ is a polynomial of degree n in λ and it has a leading coefficient of 1. We call $p(\lambda)$ the *characteristic polynomial* of A. The eigenvalues of A are the *zeros* of $p(\lambda)$: $\lambda_1, \lambda_2, \ldots, \lambda_n$.

Definition 6.4. Two $n \times n$ matrices A and C are called *similar* if there exists an invertible matrix S for which $C = S^{-1}AS$ (Note: $A = SCS^{-1}$).

Theorem 6.5. Similar matrices have the same characteristic polynomial, and hence have the same eigenvalues; their eigenvectors transform using the transition matrix S.

Definition 6.6. The matrix A is said to have a complete set of eigenvectors if there exists a basis of $\mathbb{R}^n$ (or $\mathbb{C}^n$) consisting of eigenvectors of A. The matrix A is said to be diagonalizable if A is similar to a diagonal matrix, i.e. if there exists an invertible matrix S and a diagonal matrix Λ for which $S^{-1}AS = \Lambda$.

Theorem 6.7. An $n \times n$ matrix A is diagonalizable if and only if A has a complete set of eigenvectors.

Definition 6.8. The *algebraic multiplicity* of an eigenvalue λ of A is the number of times it appears as a zero of the characteristic polynomial $p_A(\lambda)$. The *geometric multiplicity* of an eigenvalue λ of A is $\dim(\mathrm{Col}(\lambda I - A))$, i.e. the largest number of linearly independent eigenvectors associated with λ.

Theorem 6.9. For any eigenvalue λ of A, the geometric multiplicity of λ is less than or equal to the algebraic multiplicity of λ.

Theorem 6.10. An $n \times n$ matrix A is diagonalizable if and only if for every eigenvalue λ of A, the geometric multiplicity of λ is exactly the algebraic multiplicity of A.

6.2 Power Method

Definition 6.11. Suppose $A \in \mathbb{R}^{n\times n}$, and suppose that A had n linearly independent real eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$ corresponding to real eigenvalues $\lambda_1, \ldots, \lambda_n$, and in addition, that $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$. The eigenvalue λ_1 is called the *dominant eigenvalue* of A.

Definition 6.12. Let $B = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ be the $n \times n$ matrix whose columns are the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. The *basic power method* involves starting with an initial nonzero vector $\mathbf{x}_0$, and for $k = 0, 1, \ldots$, setting $\mathbf{x}_{k+1} = A\mathbf{x}_k$. By induction, $\mathbf{x}_k = A^k\mathbf{x}_0$.

Remark. Computationally, we never actually compute A^k in the basic power method for $k \geq 2$. Computing $A(A(\ldots(Ax_0)))$ requires k matrix-vector multiplies while $(A\cdots A\cdots A)x_0$ requires $(k-1)n+1$.

Remark. Since $\mathbf{u}_1, \ldots, \mathbf{u}_n$ are linearly independent, they form a basis of $\mathbb{R}^n$, so the initial vector is a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_n$, say

$$\mathbf{x}_0 = \alpha_1\mathbf{u}_1 + \alpha_2\mathbf{u}_2 + \cdots + \alpha_n\mathbf{u}_n.$$

Since $A\mathbf{u}_j = \lambda_j\mathbf{u}_j$, we have

$$\begin{aligned}
\mathbf{x}_1 &= A\mathbf{x}_0 = \alpha_1\lambda_1\mathbf{u}_1 + \alpha_2\lambda_2\mathbf{u}_2 + \cdots + \alpha_n\lambda_n\mathbf{u}_n,\\
\mathbf{x}_2 &= A\mathbf{x}_1 = \alpha_1\lambda_1^2\mathbf{u}_1 + \alpha_2\lambda_2^2\mathbf{u}_2 + \cdots + \alpha_n\lambda_n^2\mathbf{u}_n,\\
&\vdots\\
\mathbf{x}_k &= A\mathbf{x}_k = \alpha_1\lambda_1^k\mathbf{u}_1 + \alpha_2\lambda_2^k\mathbf{u}_2 + \cdots + \alpha_n\lambda_n^k\mathbf{u}_n\\
&= \lambda_1^k(\alpha_1\mathbf{u}_1 + \alpha_2(\lambda_2/\lambda_k)^k\mathbf{u}_2 + \cdots + \alpha_n(\lambda_n/\lambda_k)^k\mathbf{u}_n).
\end{aligned}$$

Since $|\lambda_1| > |\lambda_j|$, as $k \to \infty$, $\mathbf{x}_k \to \lambda_1^k\alpha_1\mathbf{u}_1$. More precisely, $\mathbf{x}_k/\lambda_1^k \to \alpha_1\mathbf{u}_1$ as $k \to \infty$.

Remark. Given $\mathbf{x}_k$ and $x_{k+1} = A_k$, how is λ_1 estimated? Typically, we take inner products with a vector $\mathbf{v}_k$. Let

$$\beta_{k+1} = \frac{\mathbf{v}_k^\top\mathbf{x}_{k+1}}{\mathbf{v}_k^\top\mathbf{x}_k}$$

where $\mathbf{v}_k$ is chosen to be either

(i) $\mathbf{x}_k$ itself;

(ii) the standard basis vector $\mathbf{e}_r$ where r is the index of the largest component of $\mathbf{x}_k$; or

(iii) some fixed vector $\mathbf{v}$.

Case (i) is used most commonly. Case (iii) is easiest to analyze.

If $|\lambda_i| > 1$, then $|\lambda_i^k| \to \infty$, and if $|\lambda_i| < 1$, then $|\lambda_i^k| \to 0$, so this method is likely to overflow or underflow.

Definition 6.13. The *scaled power method* is similar to the basic power method except we choose a vector $\mathbf{x}_0$ for which $\mathbf{x}_0^\top\mathbf{x}_0 = 1$. For $k = 0, 1, 2, \ldots$, we set

$$\begin{aligned}
\mathbf{y}_{k+1} &= A\mathbf{x}_k,\\
\beta_{k+1} &= \mathbf{x}_k^\top\mathbf{y}_{k+1},\\
n_{k+1} &= \sqrt{\mathbf{y}_{k+1}^\top\mathbf{y}_{k+1}},
\end{aligned}$$

so then $\mathbf{x}_{k+1} = \mathbf{y}_{k+1}/n_{k+1}$.

Lemma 6.14. Given an eigenvector $\mathbf{x} \neq 0$, the "best estimate" for the corresponding eigenvalue could be chosen to be the value of α that minimizes $||A\mathbf{x} - \alpha\mathbf{x}||_2^2$. The value of α that minimizes $g(\alpha) = ||A\mathbf{x} - \alpha\mathbf{x}||_2^2$ is $\alpha = \mathbf{x}^\top A\mathbf{x}/\mathbf{x}^\top\mathbf{x}$.

Definition 6.15. For $\mathbf{x} \neq 0$, $\mu_A(x) = \mathbf{x}^\top A\mathbf{x}/\mathbf{x}^\top\mathbf{x}$ is called the *Rayleigh Quotient* of $\mathbf{x}$ for the matrix A. Note that β_{k+1} above is $\beta_{k+1} = \mu_A(x)$.

Theorem 6.16 (Spectral Mapping Theorem, special case). Suppose A has eigenvalues $\lambda_1, \ldots, \lambda_n$ with corresponding eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. Then the eigenvalues of $A - \alpha I$ are $\lambda_1 - \alpha, \ldots, \lambda_n - \alpha$ with the eigenvectors $\mathbf{u}_1, \ldots, \mathbf{u}_n$. If for all $1 \leq j \leq n$, $\alpha \neq \lambda_j$, then the eigenvalues of $(A - \alpha I)^{-1}$ are $1/(\lambda_1 - \alpha_1), \ldots, 1/(\lambda_n - \alpha)$ with the same eigenvectors.

Definition 6.17. The *inverse power method* is similar to the scaled power method: Start with an α (usually close to an eigenvalue) and $\mathbf{x}_0$ with $\mathbf{x}_0^\top\mathbf{x}_0 = 1$. Then for $k = 0, 1, 2, \ldots$, we solve

$$(A - \alpha I)\mathbf{y}_{k+1} = \mathbf{x}_k$$

to get $\mathbf{y}_{k+1}$ and compute

$$\begin{aligned}
\beta_{k+1} &= \mathbf{x}_k^\top \mathbf{y}_{k+1}, \\
n_{k+1} &= \sqrt{\mathbf{y}_{k+1}^\top \mathbf{y}_{k+1}}, \\
\mathbf{x}_{k+1} &= \mathbf{y}_{k+1}/n_{k+1}.
\end{aligned}$$

Remark. A few remarks on the inverse power method:

(i) Analyticially, $\mathbf{y}_{k+1} = (A - \alpha I)^{-1}\mathbf{x}_k$, so this just the power method for $(A - \alpha I)^{-1}$, hence "inverse".

(ii) If α stays the same, we only need a PLU factorization of $A - \alpha I$ once.

(iii) If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A, then $1/(\lambda_1 - \alpha), \ldots, 1/(\lambda_n - \alpha)$ are the evaluations of $(A - \alpha I)^{-1}$.

(iv) The dominant eigenvalue of $(A - \alpha I)^{-1}$ is $1/(\lambda_j - \alpha)$ where λ_j is the closest eigenvalue of A to α.

(v) If λ_i is the second closest eigenvalue of A to α, then $\beta_{k+1} \to 1/(\lambda_j - \alpha)$ with asymptotic error constant $|\lambda_j - \alpha|/|\lambda_i - \alpha|$.

(vi) The closer α is to λ_j, the faster the rate of convergence. Howerver, when α is too close to λ_j, then $(A - \alpha I)$ can be poorly conditioned, but these errors tend to be in the direction of $\mathbf{u}_j$.

(vii) If $\beta_{k+1} \to 1/(\lambda_j - \alpha)$, then $1/\beta_{k+1} + \alpha \to \lambda_j$.

Definition 6.18. The *Rayleigh Quotient Iteration* is similar to the inverse power method, except we adjust α each time to accelerate the rate of convergence. Start with $\mathbf{x}_0$ with $\mathbf{x}_0^\top\mathbf{x}_0 = 1$. Then for $k = 0, 1, 2, \ldots$, we let

$$\alpha_k = \mathbf{x}_k^\top A\mathbf{x}_k$$

and solve

$$(A - \alpha_k I)\mathbf{y}_{k+1} = \mathbf{x}_k$$

to get $\mathbf{y}_{k+1}$ and compute

$$\begin{aligned}
\beta_{k+1} &= \mathbf{x}_k^\top \mathbf{y}_{k+1}, \\
n_{k+1} &= \sqrt{\mathbf{y}_{k+1}^\top \mathbf{y}_{k+1}}, \\
\mathbf{x}_{k+1} &= \mathbf{y}_{k+1}/n_{k+1}.
\end{aligned}$$

Remark. A few remarks on the inverse power method:

(i) Since $\mathbf{x}_k^\top \mathbf{x}_k = 1$, $\alpha_k = \mu_A(\mathbf{x}_k)$ is the Rayleigh Quotient.

(ii) We do not form $(A - \alpha_k I)^{-1}$

(iii) Each iteration requires $2/3n^3$ operations

(iv) It is prefereable to do several iterations of the power method or inverse method with fixed α to get close to an eigenvalue before switching to Rayleigh Quotient iteration.

(v) RQI converges cubically.

Definition 6.19. The *deflation* method allows us to compute additional eigenvalue/eigenvector pair. Suppose we have found an eigenvalue λ and a normalized eigenvector $\mathbf{u}$. Let S be any intevertible $n \times n$ matrix whose first column is $\mathbf{u}$. Then $B = S^{-1}AS$ is similar to A, so it has the same eigenvalues and its eigenvectors are S^{-1} times the eigenvectors of A. In particular, the matrix B is of the form

$$B = \begin{bmatrix} \lambda_1 & b_{12} & \cdots & b_{1n} \\ 0 & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & b_{n2} & \cdots & b_{nn} \end{bmatrix}.$$

Let C be the matrix

$$C = \begin{bmatrix} b_{22} & \cdots & b_{2n} \\ \vdots & \ddots & \vdots \\ b_{n2} & \cdots & b_{nn} \end{bmatrix}.$$

Then

$$p_B(\lambda) = (\lambda - \lambda_1)\det(\lambda I - C) = (\lambda - \lambda_1)p_C(\lambda).$$

So the eigenvalues of C are the other $n-1$ eigenvalues of B.

Remark. Deflation introduces round-off error. If we find an eigenvalue of C, say $\overline{\lambda_2}$, we should use the inverse power method with the original matrix A, to refine the estimate.

Definition 6.20. Let $A \in \mathbb{R}^{n\times n}$ be a matrix. Then A is *symmetric* if $A^\top = A$.

Definition 6.21. Let $\mathbf{u}_1, \ldots, \mathbf{u}_k$ be vectors in $\mathbb{R}^n$. The set of vectors are called *orthonormal* if $\mathbf{u}_i^\top \mathbf{u}_j = 0$ for $i \neq j$ and $\mathbf{u}_i^\top \mathbf{u}_i = 1$ for $1 \leq i \leq n$.

Definition 6.22. A matrix U is called *orthonormal* if its columns are orthonormal

Theorem 6.23. A symmetric matrix has a complete set of eigenvectors and the eigenvectors can be chosen to be orthonormal.

Remark. Power Method for Symmetric Matrices: The value β_{k+1} converges to λ_1 with asymptotic error constant $|\lambda_2/\lambda_1|^2$ which is twice as fast as the general case.

Deflation for Symmetric Matrices: We can factor a matrix A into the form

$$A = U\Lambda U^{-1} = U\Lambda U^\top$$

where U is the matrix whose columns are an orthonormal set of eigenvectors of A and Λ is a diagonal matrix whose entries are the eigenvalues. Then

$$A = U\Lambda U^\top = \lambda_1 \mathbf{u}_1 \mathbf{u}_1^\top + \cdots \lambda_n \mathbf{u}_n \mathbf{u}_n^\top$$

If we know $\mathbf{u}_1$ and λ_1, then $A - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^\top$ has eigenvalues $0, \lambda_2, \ldots, \lambda_n$ and eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$.

Theorem 6.24. If $A \in \mathbb{R}^{n\times n}$ is symmetric, i.e., $A^\top = A$, then the eigenvalues $\lambda_1, \ldots, \lambda_n$ of the A are all real.

Theorem 6.25. If $\lambda \in \mathbb{C}$ (or $\mathbb{R}$) is an eigenvalue of $A \in \mathbb{R}^{n\times n}$, and $||\cdot||_m$ is any matrix norm compatbile with a vector norm $||\cdot||_v$, then $|\lambda| \leq ||A_||_m$.

Corollary 6.26. Since $|\lambda| \leq ||A||_1$ and $|\lambda| \leq ||A||_\infty$. Thus

$$|\lambda| \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \qquad \text{and} \qquad |\lambda| \leq \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|.$$

In particular,

$$|\lambda| \leq n \max_{1\leq i,j\leq n} |a_{ij}|.$$

Remark. The previous corollary states that every eigenvalue λ of A is in the circle centered at 0 in the complex plane of radius $||A||_1$ and $||A||_\infty$.

Theorem 6.27 (Gershgorin Circle Theorem). Let $A \in \mathbb{R}^{n\times n}$ (or $\mathbb{C}^{n\times n}$), and define the absolute off-diagonal row and column sums to be

$$r_k = \sum_{j=1, j\neq k}^{n} |a_{kj}| \qquad \text{and} \qquad c_k = \sum_{i=1, i\neq k}^{n} |a_{ik}| \qquad \text{for } 1 \leq k \leq n.$$

For $1 \leq k \leq n$, let R_k and C_k be the circles in $\mathbb{C}$ centered at a_{kk} with radius r_k and c_k, respectively:

$$R_k = \{z \in \mathbb{C} : |z - a_{kk}| \leq r_k\} \qquad \text{and} \qquad C_k = \{z \in \mathbb{C} : |z - a_{kk}| \leq c_k\}.$$

If λ is an eigenvalue of A, then $\lambda \in \bigcup_{k=1}^{n} R_k$, and also $\lambda \in \bigcup_{k=1}^{n} C_k$.

Corollary 6.28. If m of the row Gershgorin disks of A are disjoint from the other row disks, then exactly m eigenvalues of A lie in the union of these m disks.

6.3 Transformation Methods

Definition 6.29. For any nonzero vector $\mathbf{u} \in \mathbb{R}^n$, define $Q \in \mathbb{R}^{n\times n}$ by

$$Q = I - \frac{2}{||\mathbf{u}||_2^2} \mathbf{u}\mathbf{u}^\top.$$

We call Q a *Householder transformation.* (Note that $||\mathbf{u}||_2^2 = \mathbf{u}^\top \mathbf{u}$.)

Theorem 6.30. Let $U \in \mathbb{R}^{n\times n}$ be a $n \times n$ matrix. The following are equivalent:

(i) The columns of U are orthonormal;
(ii) The rows of U are orthonormal;
(iii) $U^\top U = I$;
(iv) $UU^\top = I$;
(v) For all $\mathbf{x} \in \mathbb{R}^n$, $||U\mathbf{x}||_2^2 = ||\mathbf{x}||_2^2$.

Definition 6.31. Matrices satisfying the conditions of Theorem 6.30 are called *orthogonal matrices.*

Lemma 6.32. Let Q be a Householder transformation.

(i) Q is symmetic;

(ii) Q is orthogonal;

(iii) Geometrically, Q represents a reflection.

Remark. In general, we should never compute Q. Instead, compute $\mathbf{u}^\top\mathbf{u}$ and $\beta = 2/(\mathbf{u}^\top\mathbf{u})$.

(i) To compute $Q\mathbf{x}$, we compute $Q\mathbf{x} = \mathbf{x} - \beta(\mathbf{u}^\top\mathbf{x})\mathbf{u}$ (requires $2n$ multiplications).

(ii) To compute QA for a matrix A, we compute

$$QA = \begin{bmatrix} Q\mathbf{a}_1 & Q\mathbf{a}_2 & \cdots & Q\mathbf{a}_n \end{bmatrix}.$$

(iii) To compute AQ for a matrix A, we compute

$$AQ = \begin{bmatrix} (Q\mathbf{b}_1)^\top \\ (Q\mathbf{b}_2)^\top \\ \vdots \\ (Q\mathbf{b}_n)^\top \end{bmatrix}$$

where $\mathbf{b}_1^\top, \mathbf{b}_2^\top, \ldots, \mathbf{b}_n^\top$ are the rows of A.

Lemma 6.33. Suppose $Q = I - 2\mathbf{u}\mathbf{u}^\top/(\mathbf{u}^\top\mathbf{u})$ is a Householder transformation where for some $k \leq n$, $\mathbf{u}_1 = \mathbf{u}_2 = \cdots = \mathbf{u}_{k-1} = 0$.

(i) For any $\mathbf{x} \in \mathbb{R}^n$ the first $k-1$ elements of $Q\mathbf{x}$ are the same as the first $k-1$ elemtns of $\mathbf{x}$, that is, the first $k-1$ elements of $\mathbf{x}$ are fixed.

(ii) If in addition $x_k = x_{k+1} = \cdots = x_n = 0$, then $Q\mathbf{x} = \mathbf{x}$.

Remark. One use of Householder transformation includes creating zeros. Starting with a vector $\mathbf{v} \neq 0$ (often the $k+1$th or kth column of a matrix) and a $k \leq n$, we want to find a Householder transformation $Q = I - 2\mathbf{u}\mathbf{u}^\top/(\mathbf{u}^\top\mathbf{u})$ such that

(i) $\mathbf{u}_1 = \mathbf{u}_2 = \cdots = \mathbf{u}_{k-1} = 0$;

(ii) if $\mathbf{w} = Q\mathbf{v}$, then $\mathbf{w}_{k+1} = \mathbf{w}_{k+2} = \cdots = \mathbf{w}_n = 0$.

By (i) of the previous lemma, $w_1 = v_1, \ldots, w_{k-1} = v_{k-1}$. To make $||\mathbf{w}||_2^2 = ||\mathbf{v}||_2^2$, it must be the case that $\mathbf{w}_k = \pm\sqrt{|v_k|^2 + |v_{k+1}|^2 + \cdots + |v_n|^2}$. We choose the sign of $\mathbf{w}_k$ to avoid cancellation error. Putting everything together

$$Q\mathbf{v} = \mathbf{w} = \begin{bmatrix} v_1 & \cdots & v_{k-1} & \gamma & 0 & \cdots & 0 \end{bmatrix}^\top$$

where $\gamma = \pm\sqrt{|v_k|^2 + |v_{k+1}|^2 + \cdots + |v_n|^2}$.

Definition 6.34 (QR Factorization, invertible real matrices). Let $A \in \mathbb{R}^{n\times n}$ be invertible. QR factorization of A is similar to LU factorization except we use Householder transformations instead of row operations. We will compute Q, L where Q is an orthogonal matrix and R is an upper triangular matrix. In particular, Q will be a product of Householder transformations. We will use $Q^{-1} = Q^\top$ instead of forward substitution.

Start with $A_0 = A$, and choose the Householder transformation Q_1 mapping the first column of A_0 into a multiple of $\mathbf{e}_1$. Then $A_1 = Q_1A_0$ will have zeros below the diagonal in the first column. Suppose by induction that $A_{k-1} = Q_{k-1}\cdots Q_2Q_1A_0$ has zeros below the diagonal in columns $1, \ldots, k-1$. Let $\mathbf{v}$ be the kth column of A_{k-1}, and find the Householder transformation, Q_k that creates zeros below the diagonal in columns k. Multipying A_{k-1} by Q_k does not change the first $k-1$ columns. Since the $k+1, \ldots, n$ entries of Q_k are zero, $A_k = Q_kA_{k-1} = Q_k\cdots Q_2Q_1A_0$ has zeros below the diagonal in columns $1, \ldots, k$.

Then $R = A_{n-1} = Q_{n-1} \cdots Q_1 A$ is upper triangular. Define $Q^\top = Q_{n-1} \cdots Q_1$. Then Q is orthogonal and $R = Q^\top A$, so $A = QR$.

To solve any equation $A\mathbf{x} = \mathbf{b}$, we save the u vectors for each Householder transformation $Q_1, \ldots, Q_{n-1}$. We then solve $Q\mathbf{y} = \mathbf{b}$, $\mathbf{y} = Q_{n-1} \cdots Q_1 \mathbf{b}$ by applying the Householder transformations. Then we solve $R\mathbf{x} = \mathbf{y}$ by back substitution.

6.4 Reduction to Hessenberg Form

Definition 6.35. An $n \times n$ matrix H is said to be in *upper Hessenberg form* (or just Hessenberg form) if $h_{ij} = 0$ for $i > j + 1$. Hessenberg form is similar to upper diagonal matrices except the first subdiagonal is allowed to have non-zero elements.

Remark. Reduction to Hessenberg form is similar to QR factorization. Start with $A^0 = A$ and choose a Householder transformation $Q^{(i)}$ mapping the first column of A into

$$\begin{bmatrix} a_{11} & * & 0 & \cdots & 0 \end{bmatrix}^\top.$$

Then $A^{(1)} = Q^{(1)} A^{(0)} Q^{(1)}$ will also have this same first column. Suppose by induction that $A^{(k-1)} = Q^{(k-1)} \cdots Q^{(1)} A^0 Q^{(1)} \cdots Q^{(k-1)}$ has zeros below the first subdiagonal in columns $1, \ldots, k-1$. Let $\mathbf{v}$ be the kth column of $A^{(k-1)}$, and choose a Householder transformation $Q^{(k)}$ such that the the $k+2, \ldots, n$ entries of $Q^{(k)}\mathbf{v}$ are zero. Then $Q^{(k)} A^{(k-1)}$ has zeros below the first subdiagonal in columns $1, \ldots, k$. Multiplying by $Q^{(k)}$ from the first leaves the first columns fixed, so $A^{(k)} = Q^{(k)} A^{(k-1)} Q^{(k)}$ also zeros below the first subdiagonal in columns $1, \ldots, k$; completing the induction step. Then $A^{(n-2)} = Q^{(n-2)} \cdots Q^{(1)} A^{(0)} Q^{(1)} \cdots Q^{(n-2)}$ is Hessenberg.

Theorem 6.36. If $A \in \mathbb{R}^{n \times n}$ is symmetric, and S is an orthogonal matrix, then $S^{-1}AS = S^\top AS$ is also symmetric. Thus, if $S^{-1}AS$ is upper Hessenberg, it is symmetric and tridiagonal (the (ij) elements are 0 for $i > j + 1$ and for $i < j - 1$).

Remark. Methods for estimating eigenvalues once reduced to Hessenberg form: Krylov's method, Givens method (sturm sequences), QR algorithm.

Definition 6.37 (Krylov's Method). We set up and solve a linear system for the coefficients $a_1, \ldots, a_n$ of the characteristic polynomial

$$p_A(\lambda) = \det(\lambda I - A) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n.$$

Start with an initial vector $\mathbf{y}_0$ and apply n steps of the basic power method: $\mathbf{y}_k = A^k \mathbf{y}_0$ for $0 \le k \le n$. The values $\mathbf{y}_0, A\mathbf{y}_0, A^2\mathbf{y}_0$ are called a *Krylov sequence*. By the Cayley-Hamilton Theorem, $p_A(A) = A^n + a_1 A^{n-1} + \cdots + a_{n-1}A + a_n I = 0$, so applying this to $\mathbf{y}_0$, we get $a_1\mathbf{y}_{n-1} + a_2\mathbf{y}_{n-2} + \cdots + a_n\mathbf{y}_0 = -\mathbf{y}_n$. So $\mathbf{a} = \begin{bmatrix} a_1 & \cdots & a_n \end{bmatrix}$ is the solution of the linear system

$$\begin{bmatrix} \mathbf{y}_{n-1} & \cdots & \mathbf{y}_0 \end{bmatrix} \mathbf{a} = \mathbf{y}_n.$$

If the matrix $Y = \begin{bmatrix} \mathbf{y}_{n-1} & \cdots & \mathbf{y}_0 \end{bmatrix}$ is invertible, we can solve for $a_1, \ldots, a_n$.

Once we find $a_1, \ldots, a_n$, we still need to find the zeros of $p_A(\lambda)$. This does not work well computationally since errors can be magnified.

Theorem 6.38. If A is upper Hessenberg with $a_{j+1,j} \neq 0$ for $j = 1, \ldots, n-1$, and we choose $\mathbf{y}_0 = \mathbf{e}_1$ in Krylov's method, then Y is invertible.

Definition 6.39. A *tridiagonal matrix* is a matrix whose nonzero elements only lie on the main diagonal, the lower diagonal, and the upper diagonal. Therefore,

$$M = \begin{bmatrix} m_{11} & m_{12} & 0 & 0 & \cdots \\ m_{21} & m_{22} & m_{23} & 0 & \cdots \\ 0 & m_{23} & m_{33} & m_{34} & \cdots \\ 0 & 0 & m_{43} & m_{44} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

is a tridiagonal matrix.

Lemma 6.40. Let H be a real, $n \times n$, symmetric tridiagonal matrix. If the subdiagonal entires all the entires on the upper (lower) diagonal are nonzero, then H has n distinct simple eigenvalues $\lambda_1, \ldots, \lambda_n$.

Remark. In the case that H has a zero on the upper (lower) diagonal, we can break H up into smaller diagonal block and consider each small block separately.

Definition 6.41. Let H be a real, $n \times n$, symmetric tridiagonal matrix. For $k = 1, \ldots, n$ let

$$H_k = \begin{bmatrix} d_1 & b_1 & \cdots & 0 \\ b_1 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n. \end{bmatrix}$$

be the upper-left $k \times k$ sublock of H and let $p_k(t)$ is a polynomial of degree k, and define $p_0(t) = 1$. We call $\{p_k(t)\}$ a *Sturm sequence*. Expanding each $p_k(t)$, we have

$$\begin{aligned} p_0(t) &= 1, \\ p_1(t) &= d_1 - t, \\ p_2(t) &= (d_2 - t)p_1(t) - b_1^2 p_0(t), \\ &\vdots \\ p_n(t) &= (d_n - t)p_{n-1}(t) - b_{n-1}^2 p_{n-2}(t). \end{aligned}$$

Lemma 6.42. The polynomial $p_k(t)$ defined above is $(-1)^k$ times the characteristic polynomial of H_k.

Theorem 6.43. Let $c \in \mathbb{R}$, $P_0 = p_0(c), P_1 = p_1(c), \ldots, P_n = p_n(c)$, and $N(c)$ be the number of sign agreements in adjacent terms $P_0, \ldots, P_n$. [P_0 is always 1. If $P_k = 0$, give it the same sign as P_{k-1}.] Then $N(c)$ is the number of eigenvalues of H that are $\geq c$.

Corollary 6.44. For $r < s$, $N(r) - N(s)$ is the number of eigenvalues of H in $[r, s)$.

Remark. We do not compute the polynomials $p_1(t), \ldots, p_k(t)$. Instead we use recursion to compute $p_0(c), p_1(c), \ldots, p_n(c)$.

Definition 6.45 (Givens Method). Let H be a tridiagonal matrix. Using Theorem 6.43, we find intervals with exactly one eigenvalue and successively refine them to get an estimate for the eigenvalue.

Remark. Since we know that $|\lambda| \leq ||H||$ for any matrix norm consistent with a vector norm on $\mathbb{R}^n$, we can start Givens method with an interval like $[-||H_\infty||, ||H||_\infty]$. When $c = -||H||_\infty$, $N(c) = n$. When $c = -||H||_\infty$, $N(c) = 0$ or $N(c) = 1$. One method of finding an eigenvalue is bisection.

Definition 6.46 (Unshifted QR Algorithm). Given a real matrix $A \in \mathbb{R}^{n\times n}$ with real eigenvalues, we start with $A^{(0)} = A$ (usually A is already upper Hessenberg) and fr $m = 0, 1, 2, \ldots$, we compute the QR factorization of $A^{(m)}$: $A^{(m)} = Q^{(m)}R^{(m)}$ and set $A^{(m+1)} = R^{(m)}Q^{(m)}$.

Remark. Suppose $A^{(m)}$ is upper Hessenberg, and we use a Householder transformation to do the QR factorization. Then ${Q^{(m)}}^\top = Q_{n-1}^{(m)} \cdots Q_1^{(m)}$ is a product of Householder transformation., ${Q^{(m)}}^\top A^{(m)} = R^{(m)}$ is upper triangular, and each $Q_k^{(m)}$ (for $1 \le k \le n-1$) comes from a $\mathbf{u}_k$, $Q_k^{(m)} = I - (2/\mathbf{u}_k^\top \mathbf{u}_k)\mathbf{u}_k\mathbf{u}_k^\top$ whose only nonzero entries are the kth and $(k+1)$st. Then $Q^{(m)} = Q_1^{(m)}Q_2^{(m)} \cdots Q_{n-1}^{(m)}$, and it is easily verified that $A^{(m+1)} = R^{(m)}Q^{(m)}$ is also upper Hessenberg.

Lemma 6.47. The QR algorithm iteration $A^{(0)} \to A^{(m+1)}$ preserves upper Hessenberg form.

Definition 6.48. Define the $\widehat{Q^{(m)}}$ and $\widehat{R^{(m)}}$ to be the following matrices

$$\widehat{Q^{(m)}} = Q^{(0)}Q^{(1)} \cdots Q^{(m)}, \quad \text{and} \quad \widehat{R^{(m)}} = R^{(m)}R^{(m-1)} \cdots R^{(0)}.$$

By induction on $A^{(m+1)} = {Q^{(m)}}^\top A^{(m)}Q^{(m)}$, we see that $A^{(m+1)} = {\widehat{Q^{(m)}}}^\top A\widehat{Q^{(m)}}$. So $\widehat{Q^{(m)}}$ is the accumulation of the orthogonal matrices used in the similarity transformations of the first $m+1$ iterations of the QR algorithm.

Lemma 6.49. $\widehat{Q^{(m)}}\widehat{R^{(m)}} = A^{m+1}$.

Remark. Relation to the power method: For $i \le j \le n$, let $\widehat{q_j^{(m)}}$ denote the jth column of $\widehat{Q^{(m)}}$, and let $\widehat{r_{ij}^{(m)}}$ be the elements of $\widehat{R^{(m)}}$. Since $\widehat{R^{(m)}}$ is upper triangular,

$$A^{m+1}\mathbf{e}_1 = \widehat{Q^{(m)}}\widehat{R^{(m)}}\mathbf{e}_1 = \widehat{r_{11}^{(m)}}\widehat{q_1^{(m)}}.$$

The first column of $\widehat{Q^{(m)}}$ is in the direction of the vector obtained by applying the power method for A with $\mathbf{x}_0 = \mathbf{e}_1$. If λ_1 is the dominant eigenvalue of A with eigenvector $\mathbf{u}_1$, then $\widehat{q_1^{(m)}}$ converges to the direction of $\mathbf{u}_1$ as $m \to \infty$, and the first column of $A^{(m+1)} = {\widehat{Q^{(m)}}}^\top A\widehat{Q^{(m)}}$ converges to $[\lambda_1, 0, \ldots, 0]^\top$ as $m \to \infty$. In particular, the $(2,1)$ element of $A^{(m+1)} \to 0$ as $m \to \infty$.

Relation to the inverse power method: We reach a similar conclusion as above for the matrix $A^{(m+1)}$. If $1/\lambda_n$ is the dominant eigenvalues of $(A^\top)^{-1}$ with eigenvector $\mathbf{y}_n$ of $A^\top$, then the direction of $q_n^{(m)}$ converges in the direction of $\mathbf{y}_n$ as $m \to \infty$, and the last column of ${A^{(m+1)}}^\top = {\widehat{Q^{(m)}}}^\top A^\top\widehat{Q^{(m)}}$ converges to $[0, \ldots, 0, \lambda_n]^\top$ as $m \to \infty$. Hence the nth row of $A^{(m+1)}$ converges to $[0, \ldots, 0, \lambda_n]$.

Theorem 6.50. Suppose $A \in \mathbb{R}^{n\times n}$ has real eigenvalues $\lambda_1, \ldots, \lambda_n$ with $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$. [If S is the invertible matrix whose columns are the eigenvectors, we assume that S^{-1} has an LU factorization without pivtoing; this assumption is related to not having zero components in the direction of the dominant eigenvectors in the power method.] Then $a_{ij}^{(m)} \to 0$ for $1 \le j < i \le n$ and $a_{jj}^{(m)} \to \lambda_j$ for $1 \le j \le n$.

6.5 Schur's Decomposition

Definition 6.51. Let $A \in \mathbb{C}^{m\times n}$ be a complex matrix. Then its *Hermitian transpose* is the *conjugate transpose* $A^H = \overline{A}^\top$.

Definition 6.52. The complex inner product of two vectors $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$ is

$$\mathbf{w}^H\mathbf{z} = \sum_{j=1}^{n} \overline{w}_j z_j.$$

Note that $\mathbf{z}^H\mathbf{z} = \sum_{j=1}^n \overline{z}_j\overline{z}_j = ||\mathbf{z}||_2^2$.

Definition 6.53. Two vectors $\mathbf{w}, \mathbf{z} \in \mathbb{C}^n$ are called *orthogonal* if $\mathbf{w}^H\mathbf{z} = 0$. The analogue of orthogonal matrices in $\mathbb{C}^{n\times n}$ is *unitary matrices*.

Theorem 6.54. Let $U \in \mathbb{C}^{n\times n}$ be a matrix. The following are equivalent:

(i) the columns of U are orthonormal;

(ii) the rows of U are orthonormal;

(iii) $U^HU = I$;

(iv) $UU^H = I$;

(v) for all $\mathbf{z} \in \mathbb{C}^n$, $||U\mathbf{z}||_2^2 = ||\mathbf{z}||_2^2$.

Definition 6.55. A matrix $A \in \mathbb{C}^{n\times n}$ is called *Hermitian* if $A^H = A$.

Theorem 6.56. If $A \in \mathbb{C}^{n\times n}$ is Hermitian, its eigenvalues $\lambda_1, \ldots, \lambda_n$ are all real.

Definition 6.57. For a nonzero $U \in \mathbb{C}^n$, then $Q = I = 2/(\mathbf{u}^H\mathbf{u})\mathbf{u}\mathbf{u}^H \in \mathbb{C}^{n\times n}$ is a complex Householder transformation. Complex Householder transformations are Hermitian and unitary, so $Q^{-1} = Q^H = Q$. Given two nonzero vectors $\mathbf{v} \neq \mathbf{w} \in \mathbb{C}^n$ with $||\mathbf{v}||_2 = ||\mathbf{w}||_2$, setting $\mathbf{u} = \mathbf{v} - \mathbf{w}$ gives a complex Householder transformation for which $Q\mathbf{v} = \mathbf{w}$. provided $\mathbf{v}^H\mathbf{w}$ is real. (if not, replace $\mathbf{w}$ by $S\mathbf{w}$ for $S \in \mathbb{C}$ where $|S| = 1$.)

Definition 6.58. Two matices $A, B \in \mathbb{C}^{n\times n}$ are called *unitarily similar* if there exists a unitary matrix $U \in \mathbb{C}^{n\times n}$ for which $U^HAU = B$.

Theorem 6.59 (Schur's Theorem). Let $A \in \mathbb{C}^{n\times n}$ be a matrix. Then there exists a unitary $U \in \mathbb{C}^{n\times n}$ and an upper triangular $T \in \mathbb{C}^{n\times n}$ for which $U^HAU = T$.

Corollary 6.60. Let $A \in \mathbb{R}^{n\times n}$ be an real matrix, and suppose its eigenvalues are all real. Then there exists an orthogonal $U \in \mathbb{R}^{n\times n}$ and an upper triangular $T \in \mathbb{R}^{n\times n}$ for which $U^\top AU = T$.

Remark. The eigenvalues of an upper triangular matrix T are its diagonal elements. So if $U^HAU = T$ (with unitary U), we can read off the eigenvalues of A from the diagonal of T. Suppose $A \in \mathbb{R}^{n\times n}$ has a complex eigenvalue $\lambda = \alpha + i\beta$ where $\beta \neq 0$ with corresponding eigenvectors $\mathbf{z} = \mathbf{v} + i\mathbf{w}$. Then $\overline{\lambda} = \alpha - i\beta$ is an eigenvalue with a corresponding eigenvector $\mathbf{z} = \mathbf{v} - i\mathbf{w}$. Since $\mathbf{v}, \mathbf{w}$ are linearly independent over $\mathbb{C}$, they are linearly independent over $\mathbb{R}$, so

$$A\begin{bmatrix}\mathbf{v} & \mathbf{w}\end{bmatrix} = \begin{bmatrix}\mathbf{v} & \mathbf{w}\end{bmatrix}\begin{bmatrix}\alpha & \beta \\ -\beta & \alpha\end{bmatrix}.$$

We can reduce to almost triangular form if we allow 2 x 2 diagonal blocks for the complex eigenvalues.

Remark. Suppose $A \in \mathbb{R}^{n\times n}$ has complex eigenvalue $\lambda = \alpha + i\beta$ ($\alpha, \beta \in \mathbb{R}$, $\beta \neq 0$) with a corresponding eigenvector $\mathbf{z} = \mathbf{v} + i\mathbf{w}$ ($\mathbf{z} \neq 0, \mathbf{z} \in \mathbb{C}^n, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$). The vectors $\mathbf{v}$ and $\mathbf{w}$ (in $\mathbb{C}$) span a two-dimensional subspace spanned by $\mathbf{z}$ and $\overline{\mathbf{z}}$, but over $\mathbb{R}$ they span a two-dimensional subspace in $\mathbb{R}^n$ that is invariant under

A. Then

$$S^{-1}AS = \begin{bmatrix} \alpha & \beta & \cdots \\ -\beta & \alpha & \cdots \\ 0 & 0 & \\ \vdots & \vdots & C \\ 0 & 0 & \end{bmatrix}$$

where $C \in \mathbb{R}^{(n-2)\times(n-2)}$ and the eigenvalues of C are the rest of the eigenvalues of A.

Definition 6.61. An $n \times n$ matrix W is in quasi-upper triangular form if it is block upper triangular with only 1x1 or 2x2 blocks.

Theorem 6.62. Let $A \in \mathbb{R}^{n\times n}$ be a real matrix. Then there exists an orthogonal matrix $U \in \mathbb{R}^n$ and a quasi-upper triangular matrix $W \in \mathbb{R}^{n\times n}$ for which $U^\top AU = W$.

Theorem 6.63 (Spectral Theorem for Hermitian matrices). $A \in \mathbb{C}^{n\times n}$ is Hermitian $A^H = A$, if and only if A is unitarily similar to real diagonal matrix Λ, i.e., there is a unitary $U \in \mathbb{C}^{n\times n}$ with $U^H AU = \Lambda$.

Corollary 6.64. $A \in \mathbb{R}^{n\times n}$ is symmetric ($A^\top = A$), if and only if A is orthogonally similar to a real diagonal Λ, i.e., there is an orthogonal $U \in \mathbb{R}^{n\times n}$ with $U^\top AU = \Lambda$.

Definition 6.65. A matrix $A \in \mathbb{C}^{n\times n}$ is called a **normal matrix** if A and A^H compute, i.e., $A^H A = AA^H$.

Lemma 6.66. If $A \in \mathbb{C}^{n\times n}$ is normal and $U \in \mathbb{C}^{n\times n}$ is unitary, then $U^H AU$ is normal.

Lemma 6.67. If $T \in \mathbb{C}^{n\times n}$ is normal and upper triangular, then T is diagonal.

Theorem 6.68 (Spectral Theorem for Normal Matrices). $A \in \mathbb{C}^{n\times n}$ is normal, if and only if A is unitarily similar to a diagonal matrix $\Lambda \in \mathbb{C}^{n\times n}$, i.e., there exists a unitary $U \in \mathbb{C}^{n\times n}$ with $U^H AU = \Lambda$.

Remark. Special cases:

(i) $A \in \mathbb{C}^{n\times n}$ is Hermitian if and only if Λ is Hermitian.

(ii) $A \in \mathbb{C}^{n\times n}$ is skew Hermitian ($A^H = -A$) if and only if Λ is skew Hermitian. In this case, Λ has pure imaginary eigenvalues.

(iii) $A \in \mathbb{C}^{n\times n}$ is unitary if and only if Λ is unitary.

Definition 6.69. Let $A \in \mathbb{C}^{n\times n}$ be a matrix. The spectral radius of A, often denoted by $p(A)$, is $p(A) = \max_{1\le j\le n} |\lambda_j|$, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A.

Theorem 6.70. Let $A \in \mathbb{C}^{n\times n}$ be a matrix. Then $p(A) < 1$ if and only if $\lim_{m\to\infty} A^m = 0$.

Corollary 6.71. An iterative method $\mathbf{x}^{(k+1)} = M\mathbf{x}^{(k)} + \mathbf{g}$ is convergent for all $\mathbf{g}$ if and only if $p(M) < 1$.

Remark. It can be shown that if $p(M) < 1$, then there is a vector norm on $\mathbb{R}^n$ or $\mathbb{C}^n$ for which the operator norm of M satisfies $||M|| < 1$.

6.6 Inverse Power Method for Polynomials

Definition 6.72. Given a monic polynomial $p(x) = x^n + a_1 x^{n-1} + \cdots + a_{n-1}x + a_n$ of degree n ($a_1, \ldots, a_n$ could be real or complex), the *companion matrix* of p is the $n \times n$ matrix

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_n & a_{n-1} & -a_{n-2} & \cdots & -a_n \end{bmatrix}$$

with 1s along the first super diagonal, and $-a_n, \ldots, -a_1$ in the last row and zeros elsewhere. The transpose of A is upper Hessenberg.

Lemma 6.73. The characteristic polynomial of the companion matrix A of the polynomial p is the polynomial itself.

Theorem 6.74. Let $p(x) = x^n + a_1 x^{n-1} + \cdots + a_n$. Let $C_1, C_2, \ldots, C_n$ be the circular disks in the complex plane

$$\begin{aligned} C_1 &= \{z : |z + a_1| \leq 1\}, \\ C_k &= \{z : |z| \leq 1 + |a_k|\} \text{ for } 2 \leq k \leq n-1, \\ C_n &= \{z : |z| \leq |a_n|\}. \end{aligned}$$

Then all the zeros of p lie in $\bigcup_{k-1}^{n} C_k$.

Corollary 6.75. Let $r = 1 + \max_{1 \leq j \leq n} |a_j|$, and $C = \{z \in \mathbb{C} : |z| \leq r\}$. Then all of the zeros of p lie in C.

Remark. We could use Gershgorin's theorem on the rows of the companion matrix, but the columns give better estimates.

Theorem 6.76. The companion matrix A of $p(x) = x^n + a_1 x^{n-1} + \cdots + a_n$ is diagonalizable if and only if the zeros of p are all simple.

6.7 Shifted QR Algorithm

Definition 6.77 (QR Algorithm, Explicitly Shifted). Given $A \in \mathbb{C}^{n \times n}$, generate a sequence $A = A_0, A_1, A_2, \ldots$ of matrices unitarily similar to A as follows: given A_k, choose a scalar κ_k and choose a QR factorization of $A_k - \kappa_k I = Q_k R_k$ and use Q_k for the next similar transformation. Define $\widehat{Q_k} = Q_0 \cdots Q_k$, $\widehat{R_k} = R_k \cdots R_0$. Then $A_{k+1} = \widehat{Q_k}^H A \widehat{Q_k}$, and $\widehat{Q_k}\widehat{R_k} = (A - \kappa_0 I) \cdots (A - \kappa_k I)$.

Remark. Suppose A is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. If $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n$ and $|\kappa_k| \leq C < |\lambda_1| - |\lambda_2|$ for $k = 0, 1, 2, \ldots$ where C is a constant, then the first column of $A_{k+1} = \widehat{Q_k}^H A \widehat{Q_k}$ converges to $[\lambda_1, 0, \ldots, 0]^\top$ as $k \to \infty$.

On the other hand, if we shift by $\overline{\kappa}_\nu$ at each step, $\nu = 0, 1, \ldots, k$, and the κ_ν's for sufficiently large ν are all closest to some eigenvalue μ of A, then the best column of $A_{k+1} = \widehat{Q_k}^H A \widehat{Q_k}$ converges to $[0, \ldots, 0, \mu]^\top$ as $k \to \infty$, and the last column converges to $[0, \ldots, 0, \mu]$.

Remark. To speed up convergence, we would like shifts κ_k so that the algorithm actually applies Raleigh quotient iteration to A^H starting with $\mathbf{e}_n$. Choose $\overline{\kappa_0} = \mathbf{e}_n^H A^H \mathbf{e}_n$, i.e., choose $\kappa_0 = \alpha_{nn}$. Suppose for some $k \geq 1$, we have obtained A_k. Then we choose $\overline{\kappa_k} = (\widehat{q_n}^{(k-1)})^H A^H \widehat{q_n}^{(k-1)}$,

$$\kappa_k = (\widehat{q_n}^{(k-1)})^H A^{\widehat{q_n}^{(k-1)}} = \mathbf{e}_n^H \widehat{Q_{k-1}}^H A Q_{k-1} \mathbf{e}_n = \mathbf{e}_n^H A_k \mathbf{e}_n = \alpha_{nn}^{(k)}.$$

So choosing $\kappa_k = \alpha_{nn}^{(k)}$ gives Rayleigh quotient iteration.

Definition 6.78. A *plane rotation* of the (i,j)-plane is a matrix obtained the following way: starting with the identity matrix, set the (i,i) entry to be γ, the (i,j) entry to be σ, the (j,i) entry to be $-\sigma$, and the (j,j) entry to be γ, that is,

$$P_{ij} = I_n + \begin{bmatrix} \ddots & & & & \\ & \gamma - 1 & \cdots & \sigma - 1 & \\ & \vdots & & \vdots & \\ & -\sigma - 1 & \cdots & \gamma - 1 & \\ & & & & \ddots \end{bmatrix}.$$

Remark. Plane rotations can be used to create zeros. Given $[\alpha, \beta]^\top \in \mathbb{R}^2$, we want to choose γ, δ so that

$$\begin{bmatrix} \gamma & \sigma \\ -\sigma & \gamma \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \nu \\ 0 \end{bmatrix}.$$

Choose $\nu = \sqrt{\alpha^2 + \beta^2}$, so $\gamma = \alpha/\nu$ and $\sigma = \beta/\nu$.

Remark. Suppose A_0 is upper Hessenberg, and we apply the QR algorithm using plane rotations as shown above. Then $\alpha_{n,n-1}^{(k)} \to 0$ quadratically (cubically if A_0 is also Hermitian). After $\alpha_{n,n-1}^{(k)}$ is sufficiently small, we should working with the leading principal submatrices rather than the entire matrix, and repeat the process. If any of the subdiagonal elements are sufficiently small, we should break the matrix into submatrices with no non-zero elements on the subdiagonal.

Remark. Suppose we started with A, that U is unitary. Then there exists $A_0 = U^H A U$ is upper Hessenberg, and $A_{k+1} = \widehat{Q_k}^H A_0 \widehat{Q_k}$ as usual. Suppose the subdiagonal elements of A_{k+1} are sufficiently small; we treat A_{k+1} as upper triangular. Given a j for which $\alpha_{ii}^{(k+1)} \neq a_{jj}^{(k+1)}$ for $i < j$, we fined a $\mathbf{y}_j$ of the upper triangular system $(A_{k+1} - \alpha_{jj}^{(k+1)} I)\mathbf{y}_j = 0$ by setting the $j+1, \ldots, n$ elements of $\mathbf{y}_j$ to 0, setting the jth element to 1, then backsolving. Then $\mathbf{x}_j = U\widehat{Q_k}\mathbf{y}_j$ is approximately the eigenvector of A corresponding to the approximate eigenvalue $\alpha_{jj}^{(k+1)}$.

Definition 6.79. An upper Hessenberg matrix $B \in \mathbb{C}^{n\times n}$ is called *unreduced* if $\beta_{i+1,i} \neq 0$ for $i = 1, \ldots, n-1$. Note that with deflation, the QR algorithm works only with unreduced Hessenberg matrices.

Theorem 6.80 (Implicit Q Theorem). Suppose $A \in \mathbb{C}^{n\times n}$ such that there exists a unitary $Q \in \mathbb{C}^{n\times n}$, and there exist unreduced upper Hessenberg $B \in \mathbb{C}^{n\times n}$ with $\beta_{i+1,i} > 0$ for $i = 1, \ldots, n-1$ such that $B = Q^H A Q$. Then B and Q are uniquely determined by $\mathbf{q}_1$, the first column of Q.

Theorem 6.81. Suppose $A \in \mathbb{C}^{n\times n}$ such that there exists a unitary $Q \in \mathbb{C}^{n\times n}$, and there exist unreduced upper Hessenberg $B \in \mathbb{C}^{n\times n}$ with $\beta_{i+1,i} > 0$ for $i = 1, \ldots, n-1$ such that $B = Q^H A Q$. Suppose Q' is any unitary matrix whose first column is the same as the first column of Q, and B' is any upper Hessenberg

matrix such that $B' = {Q'}^H A Q'$. Then there exists a diagonal unitary matrix D such that $B' = D^H B D$ and $Q' = QD$.

Lemma 6.82. Suppose $A \in \mathbb{C}^{n\times n}$ is an unreduced upper Hessenberg matrix, and we apply one step of the QR algorithm with shift κ: $A - \kappa I = QR$ (where Q is unitary and R is upper triangular), and let $B = Q^H A Q$. If $A - \kappa$ is invertible, then B is also an unrreduced upper Hessenberg matrix. If $A - \kappa$ is singular, then B is upper Hessenberg matrix with $\beta_{i+1,i} \neq 0$ for $i = 1, \ldots, n-2$, $\beta_{n,n-1} = 0$, and $\beta_{nn} = \kappa$.

Remark. The ideal behind the implicit shift strategy is the observation that we only need to find Q and B. Q can be found by QR factoring $A - \kappa I$. We could instead: (1) find a unitary P^H with the same first column as Q, (2) reduce PAP^H to upper Hessenberg form in the usual way: let $U_1, \ldots, U_{n-2}$ be the elementary reflectors used to reduce PAP^H to upper Hessenberg form:

$$U_{n-2} = U_1 P A P^H U_1 \cdots U_{n-1} = B',$$

and let $Q' = P^H U_1 \cdots U_{n-2}$; then $B' = {Q'}^H A Q'$, and since $U_1, \cdots, U_{n-2}$ leave $\mathbf{e}_1$ fixed, the first column of Q' is the firt column of P^H, which is the first column of Q. Then the uniqueness theorem above tells us that there exists a diagonal unitary D such that $Q' = QD$ and $B' = D^H B D$.

Remark. Let $R' = {Q'}^H (A - \kappa I)$. Then $R' = D^H Q^H (A - \kappa I) = D^H R$ is upper triangular, and $Q'R'$ is also a QR factorization of $A - \kappa I$. Notice also that the effect of this D disappears after the next QR step.

(Finding P): We first find the first column of Q. Since R is upper triangular,

$$p_{11}\mathbf{q}_1 = p_{11} Q \mathbf{e}_1 = QR\mathbf{e}_1 = (A - \kappa I)\mathbf{e}_1 = [\alpha_{11} - \kappa, \alpha_{21}, 0, \ldots, 0]^\top,$$

so $\mathbf{q}_1$ is a multiple of $\mathbf{a} = (A - \kappa I)\mathbf{e}_1$. Choose an elementary reflection P such that $P\mathbf{a} = \pm||a||_2 \mathbf{e}_1$; then $P^H \mathbf{e}_1 = \pm a/||a||_2$ is a multiple of $\mathbf{q}_1$.

(Reducing PAP^H to upper Hessenberg form): Since P is an element reflection in the (1, 2) plane, PAP^H has zero structure which is upper Hessenberg except for the 3, 1 element. We can "choose" this nonzero element to the 4, 2 element in $U_1 P A P^H U_1$. Suppose $U_{k-1} \cdots U_1 P A P^H U_1 \cdots U_{k-1}$ is upper Hessenberg except for the $k+2, k$ element. Choosing an element reflection U_k in the $(k+1, k+2)$ plane, we can choose this nonzero element to the $k+3, k+1$ element in $U_k \cdots U_1 P A P^H U_1 \cdots U_k$. After the $k = n-2$ step, the matrix is upper Hessenberg. Then the first column of $Q' = P^H U_1 \cdots U_{n-2}$ is a multiple of $\mathbf{q}_1$ and $B' = {Q'}^H A Q'$ is upper Hessenberg.

Theorem 6.83. Suppose $A \in \mathbb{C}^{n\times n}$ is unreduced upper Hessenberg and we apply two steps of the QR algorithm with shifts k_0 and k_1:

$$A - \kappa_0 I = Q_0 R_0, \quad A_1 = Q_0^H A Q_0, \quad A_1 - \kappa_1 I = Q_1 R_1, \quad B = Q_1^H A_1 Q_1.$$

Then $B = Q_1^H Q_0^H A Q_0 A_1$ and $(A - \kappa_0 I)(A - \kappa_1 I) = Q_0 Q_1 R_1 R_0$. In addition, we have the following conclusions.

(i) If neither of κ_0, κ_1 is an eigenvalue, or if $\kappa_0 = \kappa_1$ is an eigenvalue of algebraic multiplicity 1, then B is unreduced upper Hessenberg with $\beta_{i+1,i}$ for $i = 1, \ldots, n-2$, $\beta_{n,n-1} = 0$, and β_{nn} is the eigenvalue.

(ii) If $\kappa_0 = \kappa_1$ is an eigenvalue of algebraic multiplicity > 1, then B is upper Hessenberg with $\beta_{i+1,i} \neq 0$ for $i = 1, \ldots, n-3$, $\beta_{n-1,n-2} = \beta_{n,n-1} = 0$, and $\beta_{nn} = \kappa_0$ and $\beta_{n-1,n-1} = \kappa_1$.

Suppose, in addition, Q' is unitary with the same first column as $Q_0 Q_1$, and B' is upper Hessenberg with $B' = {Q'}^H A Q'$. Then, we have the following conclusions.

(i) If neither of κ_0, κ_1 is an eigenvalue, or if $\kappa_0 \neq \kappa_1$ and exactly one is an eigenvalue, or if $\kappa_0 = \kappa_1$ is an eigenvalue of algebraic multiplicity 1, then there exists a diagonal unitary D such that $Q' = QD$ and $B' = D^H B D$.

(ii) If $\kappa_0 = \kappa_1$ is an eigenvalue of algebraic multiplicity > 1, or if $\kappa_0 \neq \kappa_1$ are both eigenvalues, then $\beta'_{n-1,n-2} = 0$, there exists a unitary U which is diagonal except possibly for the $n, n-1$ and $n-1, n$ elements such that $Q' = QU$ and $B' = U^H BU$.

Remark. Based on the conclusions above, we want to find the first column of $Q_0 Q_1$:

$$p_{11}^{(1)} p_{11}^{(0)} (Q_0 Q_1 \mathbf{e}_1) = Q_0 Q_1 R_1 R_0 \mathbf{e}_1 = (A - \kappa_0 I)(A - \kappa_1 I)\mathbf{e}_1.$$

Since A is upper Hessenberg, this vector has at most nonzero entries in the 1, 2, 3 positions and depends only on $\kappa_0 + \kappa_1$, $\kappa_0 \kappa_1$, and the upper 3x2 part of A:

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \\ 0 & \alpha_{32} \end{bmatrix}.$$

Choose an elementary reflector P in this (1, 2, 3) three-spacing mapping this vector to a multiple of $\mathbf{e}_1$. Then the first column of P^H is a multiple of the first column of $Q_0 Q_1$.

6.8 Subspace Iteration

Theorem 6.84. Suppose $A \in \mathbb{C}^{n \times n}$. Let S be a subspace of $\mathbb{C}^n$ of dimension p, and define $AS = \{A\mathbf{x} : \mathbf{x} \in S\}$. Then $A(AS) = A^2 S$. In general, $A(A \cdots (AS)) = A^k S$.

Definition 6.85 (Power Method for Subspaces). Choose $\mathbf{x} \neq 0 \in \mathbb{C}^n$. Let $p = 1, S = \text{span}\{\mathbf{x}\}$. Suppose $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$ for the eigenvalues $\lambda_1, \ldots, \lambda_n$ of A, let $\mathbf{x}_1$ be the eigenvector corresponding to λ_1, and suppose the coefficient of $\mathbf{x}_1$ for $\mathbf{x}$ is nonzero. Consider the sequnce of subspaces $S, AS, A^2 S, \ldots$. As $k \to \infty$, the subspaces $A^k S$ "converge" to $\text{span}\{\mathbf{x}_1\}$.

Definition 6.86. Fix p (with $1 \leq p \leq n-1$). Let M_p be the set of all p-dimensional subspace of $\mathbb{C}^n$. For $S, T \in M_p$, and define

$$d(S,T) = \sup_{\mathbf{s} \in S, ||\mathbf{s}||=1} \inf_{\mathbf{t} \in T} ||\mathbf{s} - \mathbf{t}||_2$$

to be the distance between S and T.

Lemma 6.87. Let $S, T \in M_p$. Then

(i) $d(S,T) = \sup_{\mathbf{s} \neq 0 \in S} \frac{||\mathbf{s}-\mathbf{t}||_2}{||\mathbf{s}||_2}$,

(ii) $0 \leq d(S,T) \leq 1$,

(iii) d is a matrix on M_p.

Definition 6.88. Given $S \in M_p$, define $P_S \in \mathbb{C}^{n \times n}$ by:

$$P_S \mathbf{x} = \begin{cases} \mathbf{x} & \text{if } \mathbf{x} \in S, \\ 0 & \text{if } \mathbf{x} \in S^\perp = \{\mathbf{y} : \mathbf{y}^H \mathbf{s} = 0 \text{ for } \mathbf{s} \in S\} \end{cases}.$$

We call P_S an **orthogonal projection**.

Lemma 6.89. Let P_S be an orthogonal projection. Then

(i) $P_S^2 = P_S$;

(ii) $P_S^H = P_S$;

(iii) $I - P_S = P_{S^\perp}$;

(iv) for $\mathbf{x} \in \mathbb{C}^n$, $||P_S\mathbf{x}||_2^2 + ||(I - P_S)\mathbf{x}||_2^2 = ||\mathbf{x}||_2^2$.

Definition 6.90. Let $S, T \in M_p$. Define θ with $0 \leq \theta \leq \pi/2$ such that

$$\cos\theta = \min_{\mathbf{s}\neq 0 \in S} \max_{t \neq 0 \in T} \frac{|\mathbf{s}^H \mathbf{t}|}{||\mathbf{s}||_2 ||\mathbf{t}||_2}.$$

We call θ the **angle** between subspaces.

Theorem 6.91. If $S, T \in M_p$ and θ be the angle between them, then

$$d(S,T) = ||P_S - P_T||_2 = \sin\theta.$$

Corollary 6.92. If $S, T \in M_p$ and θ be the angle between them, then

$$d(S,T) = 1 \quad \Leftrightarrow \quad S \cap T^\perp \neq \{0\} \quad \Leftrightarrow \quad T \cap S^\perp \neq \{0\}.$$

Theorem 6.93. Suppose $A \in \mathbb{C}^{n\times n}$ is diagonalizable with eigenvalues $\lambda_1, \ldots, \lambda_n$ satisfying

$$|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|,$$

and eigenvectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$. Let $T = \text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ (called the dormant subspace), and $U = \text{span}\{\mathbf{x}_{p+1}, \ldots, \mathbf{x}_n\}$ (called the codominant subspace). Suppose $S \cap \mathbb{C}^n$ is any subspace of $\dim p$ such that $S \cap U = \{0\}$. Then there exists C such that

$$d(A^k S, T) \leq C \left| \frac{\lambda_{p+1}}{\lambda_p} \right|^k.$$

Remark. (i) If λ_{p+1} does not have complete set of eigenvectors, get $d(A^k S, T)$ bounded by a polynomial in k times $|\lambda_{p+1}/\lambda_p|^k$.

(ii) As in the power method, shifts can be used.

(iii) As in the powe rmethod, inverse subspace iteration is possible.

Remark. Let $W_k = [U_k, V_k] \in \mathbb{C}^{n\times n}$ be unitary such that the columns of U_k span $A^k S$. Write

$$W_k^H A W_k = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

It can be shown that $B_{21} \to 0$ at the same rate that $A^k S \to T$.

Lemma 6.94. Choose $S \in \dim S = p$ and $S \cap U = \{0\}$. Then since $|\lambda_p| > |\lambda_{p+1}|$, $\text{Null}(A) \subset U$, so $S \cap \text{Null}(A) = \{0\}$; also $A^k S \cap U = \{0\}$. Suppose $\mathbf{q}_1^{(0)}, \ldots, \mathbf{q}_p^{(0)}$ span S. Then $A^k\mathbf{q}_1^{(0)}, \ldots, A^k\mathbf{q}_p^{(0)}$ span $A^k S$.

Theorem 6.95. Suppose $A \cap \mathbb{C}^{n\times n}$ is invertible and S is a p-dim subspace. Then $A^k S$ and $(A^H)^{-k}(S^\perp)$ are orthogonal complements.

Theorem 6.96. For simplicity, consider the unshifted QR algorithm. Recall $A^{k+1} = \widetilde{Q}_k \widetilde{R}_k$. Let $\mathbf{q}_j^{(j)} = \mathbf{e}_j$ for $j = 1, \ldots, n$ and $p = n$ in subspace iteration, and let $S_j^{(k)} = \text{span}\{A^{k+1}\mathbf{e}_1, \ldots, A^{k+1}\mathbf{e}_j\}$; let $\widetilde{q}_j^{(k)}, \ldots, \widetilde{q}_n^{(k)}$ denote the columns of $\widetilde{Q}_k$. Then $\text{span}\{\widetilde{q}_j^{(k)}, \ldots, \widetilde{q}_n^{(k)}\} = S_j^{(k)}$.

7 Inner Product Spaces

7.1 Newton's Method

Definition 7.1. Suppose we have a non-linear system of n real equations with n unknowns $x_1, \ldots, x_n$:

$$\begin{cases} f_1(x_1, \ldots, x_n) & = 0, \\ f_2(x_1, \ldots, x_n) & = 0, \\ & \vdots \\ f_n(x_1, \ldots, x_n) & = 0. \end{cases}$$

If $\mathbf{x} = [x_1, \ldots, x_n]^\top$, $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})]^\top$, then the system becomes $\mathbf{f}(\mathbf{x}) = 0$.

At $\mathbf{x}_k$, we define an $n \times n$ matrix

$$J[\mathbf{x}_k] = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}_k) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}_k) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\mathbf{x}_k) & \cdots & \frac{\partial f_n}{\partial x_n}(\mathbf{x}_k) \end{bmatrix}$$

to be the *Jacobian matrix* of $\mathbf{f}$ at $\mathbf{x}_k$. If $J(\mathbf{x}_k)$ is invertible, we solve the linear system

$$J[\mathbf{x}_k] \cdot \Delta\mathbf{x}_k = -\mathbf{f}(\mathbf{x}_k) \text{ for } \Delta\mathbf{x}_k,$$

and then set $\mathbf{x}_{k+1} = \mathbf{x}_k + \Delta\mathbf{x}_k$. Theoretically, $\mathbf{x}_{k+1} = \mathbf{x}_k - J[\mathbf{x}_k]^{-1}\mathbf{f}(\mathbf{x}_k)$ which is the higher dimensional analogue of Newton's method.

Theorem 7.2 (Local Convergence Theorem). Suppose $\mathbf{f} : U \to \mathbb{R}^n$ is defined on an open set $U \subset \mathbb{R}^n$ such that $\mathbf{f} \in C^2$ and $\mathbf{f}(\mathbf{s}) = 0$ for some $\mathbf{s} \in U$, and that $J[\mathbf{s}]$ is invertible. Starting with an $\mathbf{x}_0 \in U$, we generate a sequence $\{\mathbf{x}_k\}$ by Newton's method $\mathbf{x}_{k+1} = \mathbf{x}_k - J[\mathbf{x}_k]^{-1}\mathbf{f}(\mathbf{x}_k)$ provided that each $\mathbf{x}_k \in U$ and each $J[\mathbf{x}_k]$ is invertible. Let $\mathbf{e}_k = \mathbf{s} - \mathbf{x}_k$ and let $B_p = \{\mathbf{x} \in \mathbb{R}^n : ||\mathbf{x} - \mathbf{s}||_2 \leq p\}$. Then there exists a constant $p > 0$ for which $B_p \subset U$ and for all $\mathbf{x} \in B_p$, $J[\mathbf{x}]$ is inverible, and if $\mathbf{x}_0 \in B_p$ then

(i) the sequence $\{\mathbf{x}_k\}$ is well-defined and $\mathbf{x}_k \in B_p$ for all $k \geq 0$,

(ii) $\mathbf{x}_k \to \mathbf{s}$, and

(iii) for some constant K, $||\mathbf{e}_{k+1}||_2 \leq K||\mathbf{e}_k||_2^2$.

7.2 Inner Product Spaces and Least Squares

Definition 7.3. A real *inner product space* is a real vector space V with an inner product $V \times V \to \mathbb{R}_{\geq 0}$, denoted by $\langle x, y\rangle$, satisfying:

(i) for all $x \in V$, $\langle x, x\rangle \geq 0$, $\langle x, x\rangle = 0$ if and only if $x = 0$;

(ii) for all $\alpha, \beta \in \mathbb{R}$, $x, y, z \in V$, $\langle \alpha x + \beta y, z\rangle = \alpha\langle x, z\rangle + \beta\langle y, z\rangle$;

(iii) for all $x, y \in V$, $\langle x, y\rangle = \langle y, x\rangle$.

Remark. The inner product is linear in the first variable, by (ii); symmetric, by (iii); and linear in the second variable, by (ii) and (iii), so the inner product is a *symmetric bilinear form* which is *positive definite*, by (i).

Theorem 7.4 (Cauchy Schwarz Inequality). Let V be a real inner product space. Then

$$|\langle x, y\rangle| \leq ||x|| \cdot ||y|| \text{ for all } x, y \in V.$$

Definition 7.5. Define the *norm* to be $||x|| = \sqrt{\langle x, x\rangle}$ for each $x \in V$.

Lemma 7.6. Suppose V is a real inner product space. Let $\alpha \in \mathbb{R}$ and $x, v \in V$.

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

(ii) $||\alpha x|| = |\alpha| \cdot ||x||$;

(iii) $||x + y|| \leq ||x|| + ||y||$.

Definition 7.7. If $\langle x, r\rangle = 0$, we say x and y are *orthogonal*; we write $x \perp y$. If S is a subspace of V, and $y \in V$ satisfies $y \perp x$ for all $x \in S$, then we say that y is orthogonal to S and write $y \perp S$. Similarly, we can define the set of vectors orthogonal to S by

$$S^{\perp} = \{y \in V : y \perp S\}.$$

Lemma 7.8. Let S be a subspace of V. Then

(i) $S^{\perp}$ is a subspace of V,

(ii) $S \cap S^{\perp} = \{0\}$.

Definition 7.9. If $\varphi_1, \varphi_2, \ldots, \varphi_n \in V$ satisfy

(i) $\langle \varphi_j, \varphi_k\rangle = 0$ for $j \neq k$, and

(ii) $\langle \varphi_j, \varphi_j\rangle \neq 0$ for $j = 1, \ldots, n$;

then we say that $\{\varphi_1, \ldots, \varphi_n\}$ is an *orthogonal system* in V. If in addition

(iii) $\langle \varphi_j, \varphi_j\rangle = 1$ for $j = 1, \ldots, n$;

then we say that $\{\varphi_1, \ldots, \varphi_n\}$ is an *orthonormal system* in V.

Lemma 7.10. Suppose V is a real inner product space.

(i) If $\{\varphi_1, \ldots, \varphi_n\}$ is an orthogonal system, then $\varphi_1, \ldots, \varphi_n$ are linearly independent.

(ii) If $\langle x, y\rangle = 0$, then $||x + y||^2 = ||x||^2 + ||y||^2$ (Pythagorean Theorem).

(iii) If $\{\varphi_1, \ldots, \varphi_n\}$ is an orthogonal system, $||\sum_{j=1}^n c_j\varphi_j||^2 = \sum_{j=1}^n |c_j|^2\langle \varphi_j, \varphi_j\rangle$.

(iv) If $\{\varphi_1, \ldots, \varphi_n\}$ is an orthonormal system, $||\sum_{j=1}^n c_j\varphi_j||^2 = \sum_{j=1}^n |c_j|^2$.

Theorem 7.11 (Gram-Schmidt Process). Let S be an n-dimensional subspace of an inner product space V. Then S has an orthonormal basis.

Remark. (Orthonormal version) Start with any basis $\varphi_1, \varphi_2, \ldots, \varphi_n$ of S. Let

$$\begin{aligned}
\zeta_1 &= \varphi_1, \\
\psi_1 &= \zeta_1/||\zeta_1||, \\
\zeta_2 &= \varphi_2 - \langle \varphi_2, \psi_1\rangle\psi_1, \\
\psi_2 &= \zeta_2/||\zeta_2||, \\
&\vdots \\
\zeta_j &= \varphi_j - \sum_{k=1}^{j-1}\langle \varphi_j, \psi_k\rangle\psi_k, \\
\psi_j &= \zeta_j/||\zeta_j||.
\end{aligned}$$

Then $\psi_1, \psi_2, \ldots, \psi_n$ is an orthonormal basis of S.

(Orthogonal version) Start with any basis $\varphi_1, \varphi_2, \ldots, \varphi_n$ of S. Let

$$\begin{aligned}
\eta_1 &= \varphi_1, \\
\eta_2 &= \varphi_2 - \frac{\langle \varphi_2, \eta_1 \rangle}{\langle \eta_1, \eta_1 \rangle} \eta_1, \\
&\vdots \\
\eta_j &= \varphi_2 - \sum_{k=1}^{n-1} \frac{\langle \varphi_j, \eta_k \rangle}{\langle \eta_k, \eta_k \rangle} \eta_k.
\end{aligned}$$

Then $\eta_1, \eta_2, \ldots, \eta_n$ is an orthogonal basis of S.

Theorem 7.12 (The Projection Theorem). Let V be an inner product space, and let S be a finite dimensional subspace. Then

(i) $V = S \oplus S^\perp$, i.e., given a $v \in V$, there are unique elements $x^* \in S$ and $e^* \in S^\perp$ for which $y = x + e$.

(ii) Given a $y \in V$, the x in (i) is the unique element of S which satisfies $\langle y - x^*, x \rangle = 0$ for all $x \in S$.

(iii) Given a $y \in V$, then the x^* in (i) is the unique element of S which minimizes $||y - x||^2$ over all $x \in S$.

Theorem 7.13 (The Normal Equations). Let V be an inner product space, and let S be a finite dimensional subspace, and let $\varphi_1, \ldots, \varphi_n \in S$ be a set of vectors which spans S. Represent each element $x \in S$ in terms of $\varphi_1, \ldots, \varphi_n$: $x = c_1\varphi_1 + \cdots + c_n\varphi_n$. Let $y \in V$. Then $x \in S$ minimizes $||y - x||^2 = ||y - \sum_{j=1}^n c_j\varphi_j||^2$ over all possible elements x of S if and only if the coefficients $c_1, \ldots, c_n$ of x satisfy the normal equations:

$$\begin{bmatrix}
\langle \varphi_1, \varphi_1 \rangle & \langle \varphi_1, \varphi_2 \rangle & \cdots & \langle \varphi_1, \varphi_n \rangle \\
\langle \varphi_2, \varphi_1 \rangle & \langle \varphi_2, \varphi_2 \rangle & \cdots & \langle \varphi_2, \varphi_n \rangle \\
\vdots & \vdots & \ddots & \vdots \\
\langle \varphi_n, \varphi_1 \rangle & \langle \varphi_n, \varphi_2 \rangle & \cdots & \langle \varphi_n, \varphi_n \rangle
\end{bmatrix}
\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}
=
\begin{bmatrix} \langle y, \varphi_1 \rangle \\ \langle y, \varphi_2 \rangle \\ \vdots \\ \langle y, \varphi_n \rangle \end{bmatrix}.$$

Lemma 7.14. The following are special cases of the normal equations.

(i) If $\varphi_1, \ldots, \varphi_n$ is an orthogonal basis of S, then

$$x^* = \sum_{j=1}^n \frac{\langle y, \varphi_j \rangle}{\langle \varphi_j, \varphi_j \rangle} \varphi_j.$$

(ii) If $\varphi_1, \ldots, \varphi_n$ is an orthonomal basis of S, then

$$x^* = \sum_{j=1}^n \langle y, \varphi_j \rangle \varphi_j.$$

(iii) If V is a finite dimensional inner product space and $\psi_1, \ldots, \psi_n$ is an orthonormal basis of V, then every $y \in V$, $y = \sum_{j=1}^n \langle y, \psi_j \rangle \psi_j$. Take $S = V$. Then $S^\perp = \{0\}$, $e^* = 0$ and $x^* = y$.

Theorem 7.15 (Bessel's Inequality). Let V be an inner product space, and suppose $\{\psi_1, \ldots, \psi_n\}$ is an orthonormal set in V. Then for every $y \in V$,

$$\sum_{j=1}^n |\langle y, \psi_j \rangle|^2 \leq ||y||^2.$$

Bessel's Inequalty holds for infinite inner product spaces. Let V be an inner product space, and suppose $\{\psi_1, \psi_2, \psi_3, \ldots\}$ is an orthonormal set in V. Then for every $y \in V$,

$$\sum_{j=1}^\infty |\langle y, \psi_j \rangle|^2 \leq ||y||^2.$$

Theorem 7.16. Let V be an inner product space, and suppose $\{\psi_1, \psi_2, \psi_3, \ldots\}$ is an orthonormal set in V. Then the following two conditions are equivalent:

(i) (Parseval's Equality) For every $y \in V$,

$$\sum_{j=1}^{\infty} |\langle y, \psi_j \rangle|^2 = ||y||^2.$$

(ii) For every $y \in V$, the series $\sum_{j=1}^{\infty} \langle y, \psi_j \rangle \psi_j$ converges in V to y, i.e.,

$$||y - \sum_{j-1}^{n} \langle y, \psi_j \rangle \psi_j||$$

goes to 0 as $n \to \infty$.

Definition 7.17. If $\{\psi_1, \psi_2, \psi_3, \ldots\}$ is an orthonormal set in an inner product space V which satisfies Parseval's equality, then $\{\psi_1, \psi_2, \psi_3, \ldots\}$ is called a **complete orthonormal system** in V.

7.3 Applications of Inner Product Spaces

Remark. (Application 1) Linear least squares in $\mathbb{R}^m$: Suppose $m \geq n$ and $\mathbf{a}^{(1)}, \mathbf{a}^{(2)}, \ldots, \mathbf{a}^{(n)} \in \mathbb{R}^m$. Let $S = \text{span}\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}\}$. Given $\mathbf{y} \in \mathbb{R}^m$, find $\mathbf{x}^* \in S$ closest to $\mathbf{y}$ and its coefficients $c_1, \ldots, c_n$ when represented as $\mathbf{x}^* = c_1 a^{(1)} + \cdots + c_n a^{(n)}$.

Let $A = [\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}]$ be a $m \times n$ matrix. Then for $\mathbf{x} = c_1 a^{(1)} + \cdots + c_n a^{(n)}$ in S, $\mathbf{x} = A\mathbf{c}$ so

$$||\mathbf{x} - \mathbf{y}||_2^2 = ||A\mathbf{c} - \mathbf{y}||_2^2.$$

We can restate the linear least squares problem as finding $\mathbf{c} \in \mathbb{R}^n$ that minimizes $||A\mathbf{c} - \mathbf{y}||_2^2$. The normal equations become $A^\top A\mathbf{c} = A^\top \mathbf{y}$.

Lemma 7.18. Suppose $A \in \mathbb{R}^{m \times n}$ with $m > n$, and that A has rank n. If $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ has rank n if and only if $A^\top A \in \mathbb{R}^{n \times n}$ is invertible.

Lemma 7.19. If $A \in \mathbb{R}^{m \times n}$ with $m \geq n$ has rank n, then the linear least squares problem $A^\top A\mathbf{c} = A^\top \mathbf{y}$ has a unique solution $\mathbf{c}$.

Remark. (QR factorization for LLS problems) Suppose $A \in \mathbb{R}^{m \times n}$ with $m > n$ and A is full rank. We can apply QR factorization to compute $A = QR$ where Q is an $m \times m$ orthogonal matrix and R is an $m \times n$ upper triangular matrix. Partition $Q = [\widetilde{Q}, \widetilde{\widetilde{Q}}]$ and $R = [\widetilde{R}, 0]^\top$ where $\widetilde{Q} \in \mathbb{R}^{m \times n}$, $\widetilde{\widetilde{Q}} \in \mathbb{R}^{m \times (m-n)}$, $\widetilde{R} \in \mathbb{R}^{n \times n}$. By block matrix multiplication $A = QR = \widetilde{Q}\widetilde{R}$. This form, $A = \widetilde{Q}\widetilde{R}$, where $\widetilde{Q} \in \mathbb{R}^{m \times n}$ has orthonormal columns, and $\widetilde{R}$ is upper triangular is often called the condensed QR-factorization. Moreover, since $\text{rank}(A) = n$ and $A = \widetilde{Q}\widetilde{R}$, $\text{rank}(\widetilde{R}) \geq n$, and thus $\widetilde{R}$ is invertible. The columns $\mathbf{q}^{(1)}, \ldots, \mathbf{q}^{(n)}$ of $\widetilde{Q}$ are an orthonormal basis of the range of A.

Given $\mathbf{y} \in \mathbb{R}^m$, for any $\mathbf{c} \in \mathbb{R}^n$, we have

$$||A\mathbf{c} - \mathbf{y}||_2^2 = ||\widetilde{R}\mathbf{c} - \widetilde{Q}^\top \mathbf{y}||_2^2 + ||\widetilde{\widetilde{Q}}^\top \mathbf{y}||_2^2.$$

Then $||A\mathbf{c} - \mathbf{y}||_2^2$ is minimized if and only if $\widetilde{R}\mathbf{c} = \widetilde{Q}^\top \mathbf{y}$ since $\widetilde{R}$ is invertible. We solve for $\mathbf{c}$ by back-subtitution. Since $Q^\top$ is orthogonal, the linear least-squares is equivalent to $Q^\top A\mathbf{c} = Q^\top \mathbf{y}$. We solve $\widetilde{R}\mathbf{c} = \widetilde{Q}^\top \mathbf{y}$ for $\mathbf{c}$. The closest element of S to $\mathbf{y}$ is $\mathbf{x}^* = A\mathbf{c}$, and $||\mathbf{e}^*||_2 = ||\widetilde{\widetilde{Q}}^\top \mathbf{y}||_2$.

Lemma 7.20. Let $\kappa_2(M) = ||\widetilde{M}||_2 \cdot ||\widetilde{R}^{-1}||_2$ be the condition number of $\widetilde{R}$. Then $\kappa_2(A^\top A) = \kappa_2(\widetilde{R})^2$.

Theorem 7.21. Suppose $A \in \mathbb{R}^{m\times n}$ with $m \geq n$, and A is full rank. The condensed form of QR factorization is $A = \widetilde{Q}\widetilde{R}$, where $\widetilde{Q} = [\mathbf{q}^{(1)} \cdots \mathbf{q}^{(n)}] \in \mathbb{R}^{m\times n}$ has orthonormal columns, and $\widetilde{R} \in \mathbb{R}^{n\times n}$ is upper triangular and invertible. For $j = 1, 2, \ldots, n$, $\{\mathbf{q}^{(1)} \cdots \mathbf{q}^{(j)}\}$ is an orthonormal basis of $S_j = \text{span}\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(j)}\}$.

Remark. Let $S = \text{span}\{\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)}\}$. Given $\mathbf{y} \in \mathbb{R}^m$, let $\mathbf{x}^*$ be the closest element of S to $\mathbf{y}$. Then $\mathbf{x}^* = c_1\mathbf{a}^{(1)} + \cdots + c_n\mathbf{a}^{(n)} = A\mathbf{c}$, where $\mathbf{c}$ is the solution of $A\mathbf{c} = \mathbf{y}$. If we represent $\mathbf{x}^*$ in terms of the orthonormal basis $\{\mathbf{q}^{(1)} \cdots \mathbf{q}^{(n)}\}$ of S: $\mathbf{x}^* = d_1\mathbf{q}^{(1)} + \cdots d_n\mathbf{q}^{(n)} = \widetilde{Q}\mathbf{d}$, then $d_j = {\mathbf{q}^{(j)}}^\top \mathbf{y}$, so $\mathbf{d} = \widetilde{Q}^\top\mathbf{y}$. So solving the LLS problem $A\mathbf{c} = \mathbf{y}$ can be broken down into

(1) Form the condensed QR factorization of A: $A = \widetilde{Q}\widetilde{R}$.
(2) Form $\mathbf{d} = \widetilde{Q}^\top\mathbf{y}$.
(3) Backsolve $\widetilde{R}\mathbf{c} = \mathbf{d}$ to find $\mathbf{c}$.

Remark. (Gram-Schmidt as a condensed QR factorization) Suppose $\mathbf{a}^{(1)}, \ldots, \mathbf{a}^{(n)} \in \mathbb{R}^m$ are linearly independent ($m \geq n$). Apply (orthonormal) Gram-Schmidt:

$$\boldsymbol{\zeta}^{(1)} = \mathbf{a}^{(1)}; \qquad \boldsymbol{\psi}^{(1)} = \boldsymbol{\zeta}^{(1)}/||\boldsymbol{\zeta}^{(1)}||_2$$
$$\boldsymbol{\zeta}^{(j)} = \mathbf{a}^{(j)} - \sum_{i=1}^{j-1}({\boldsymbol{\zeta}^{(i)}}^\top \mathbf{a}^{(j)})\boldsymbol{\psi}^i; \qquad \boldsymbol{\psi}^{(j)} = \boldsymbol{\zeta}^{(j)}/||\boldsymbol{\zeta}^{(j)}||_2$$

Using $\boldsymbol{\zeta}^{(j)} = t_{jj}\boldsymbol{\psi}^{(j)}$, we can rewrite this as $\mathbf{a}^{(1)} = t_{11}\boldsymbol{\psi}^{(1)}$, for $2 \leq j \leq n$, $\mathbf{a}^{(j)} = \sum_{i=1}^{j} t_{ij}\boldsymbol{\psi}^{(i)}$, Let

$$\Psi = [\boldsymbol{\psi}^{(1)} \cdots \boldsymbol{\psi}^{(n)}],$$
$$T = \begin{bmatrix} t_{11} & \cdots & t_{1n} \\ & \ddots & \vdots \\ & & t_{nn} \end{bmatrix}, \text{ and}$$
$$A = [\mathbf{a}^1 \cdots \mathbf{a}^n].$$

Then Ψ has orthonormal columns, T is $n \times n$ invertible, upper-triangular, and $A = \Psi T$, which is a condensed QR factorization of A.

Remark. (Modified Gram Schmidt) Computationally, the classic Gram Schmidt algorithm suffers from numerical instability. The classic Gram-Schmidt algorithm is as follows: For $j = 1, \ldots, n$,

$$\mathbf{b} = \mathbf{a}^{(j)}.$$

For $i = 1, \ldots, j-1$ (if $j > 1$),

$$\mathbf{b} = \mathbf{b} - ({\boldsymbol{\psi}^{(i)}}^\top \mathbf{a}^{(j)})\boldsymbol{\psi}^{(i)},$$
$$\boldsymbol{\psi}^{(j)} = \mathbf{b}/||\mathbf{b}||_2.$$

The modified Gram-Schmidt algorithm has a single change: For $j = 1, \ldots, n$,

$$\mathbf{b} = \mathbf{a}^{(j)}.$$

For $i = 1, \ldots, j-1$ (if $j > 1$),

$$\mathbf{b} = \mathbf{b} - ({\boldsymbol{\psi}^{(i)}}^\top \mathbf{b})\boldsymbol{\psi}^{(i)},$$
$$\boldsymbol{\psi}^{(j)} = \mathbf{b}/||\mathbf{b}||_2.$$

Theorem 7.22 (Singular Value Decomposition). Let $A \in \mathbb{R}^{m\times n}$ with $m \geq n$. Then there exists orthogonal matrices $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{m\times m}$ and a diagonal matrix $\Sigma \in \mathbb{R}^{m\times m}$ with $\sigma_1 \geq \cdots \geq \sigma \geq 0$ such that $A = V\Sigma U^\top$.

Corollary 7.23. Let $V_r = [\mathbf{v}^{(1)} \cdots \mathbf{v}^{(r)}] \in \mathbb{R}^{m\times r}$, $\Sigma_r \in \mathbb{R}^{r\times r}$ be diagonal, and $U_r = [\mathbf{u}^{(1)} \cdots \mathbf{u}^{(r)}] \in \mathbb{R}^{n\times r}$. Then $A = V_r\Sigma_r U_r^\top$.

Remark. (Weighted Discrete Inner Product) Fix distinct $x_0 < x_1 < \cdots < x_m \in \mathbb{R}$ and weights $w_0, w_1, \ldots, w_m > 0$. Let GF be the vector space of all grid functions $f : \{x_0, \ldots, x_m\} \to \mathbb{R}$. Represent each $f \in$ GF by its graph vector $\mathbf{F} = [f(x_0) \cdots f(x_m)]^\top \in \mathbb{R}^{m+1}$. (The mapping $f \to \mathbf{F}$ is a vector space isomorphism.) Define the inner product on GF to be

$$\langle f, g\rangle_w = \sum_{i=0}^{m} f(x_i)g(x_i)w_i.$$

Lemma 7.24.

(i) Let W be a $(m+1)\times(m+1)$ diagonal matrix. Then for $f, g \in$ GF,

$$\langle f, g\rangle_w = \mathbf{F}^\top W\mathbf{G}, \qquad ||f||_w = \sqrt{\mathbf{F}^\top\mathbf{W}\mathbf{F}};$$

(ii) A set of functions in GF is linearly independent if and only if their graph vector are linearly independent in $\mathbb{R}^{m+1}$;

(iii) $\dim(\text{GF}) = m + 1$;

Remark. We often above notation: suppose $a \leq x_0 < \cdots < x_m \leq b$. If $f \in C[a, b]$, we often speak as if $f \in$ GF, meaning the restriction of f to the domain $\{x_0, x_1, \ldots, x_m\}$.

Lemma 7.25. Let $R : C[a, b] \to$ GF be the restriction operator, mapping $f : [a, b] \to \mathbb{R}$ into $f|_{x_0,\ldots,x_m} : x_0, \ldots, x_m \to \mathbb{R}$, and let S be a subspace of $C[a, b]$ for which $R : S \to$ GF is one-to-one on S. Then $\langle f, g\rangle_w$ is an inner product on S.

Lemma 7.26. The set $\{g_0(x_i) = 1, g_1(x_i) = x_i, \ldots, g_n(x_i) = x_i^n\}$ are linearly independent in GF if and only if $n \leq m$.

Remark. (Least Squares Polynomial Fitting) Suppose $a \leq x_0 < \cdots < x_m \leq b$ and $n \leq m$. We can either view $S = \mathbb{P}_n \subset C[a, b]$ or $S = \text{span}\{g_0, \ldots, g_n\} \subset$ GF. Let $p(x) = c_0 + c_1x + \cdots + c_nx^n$ (or $p = c_0g_0 + \cdots c_ng_n$) Given $y \in$ GF (with $\mathbf{Y} = [y(x_0) \cdots y(x_m)]^\top$), we want to find $p \in S$ which minimizes $||p - y||_w^2$. Then

$$A = [\mathbf{G}_0 \cdots \mathbf{G}_n] = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & & \vdots \\ 1 & x_m & \cdots & x_m^n \end{pmatrix},$$

and $\mathbf{c}$ is the solution of the LLS problem $A\mathbf{c} = \mathbf{Y}$ and then p follows.

Remark. (Weighted Integral Inner Product) Fix $a, b \in \mathbb{R}$ with $a < b$, and fix $w(x) > 0$ that is continuous on (a, b) with $\int_a^b w(x)\,dx < \infty$. Let $V = C[a, b]$ the space of continuous function $f : [a, b] \to \mathbb{R}$ with inner product

$$\langle f, g\rangle_w = \int_a^b f(x)g(x)w(x)\,dx.$$

Theorem 7.27. $\langle f, g\rangle_w$ is an inner product on $C[a, b]$.

Remark. Let $g_0, \dots, g_n$ be linearly independent functions in $C[a,b]$, and let $S = \text{span}\{g_0, \dots, g_n\}$. Given $y \in C[a,b]$, we want to find the closest $f = c_0 g_0 + \cdots c_n g_n \in S$ to y in

$$||f - y||_w^2 = \int_a^b \left(\sum_{j=0}^{n} c_j g_j(x) - y(x) \right)^2 w(x)\, dx.$$

We can form the normal equations

$$\begin{bmatrix} \langle g_0, g_0 \rangle_w & \cdots & \langle g_0, g_n \rangle_w \\ \vdots & \ddots & \vdots \\ \langle g_n, g_0 \rangle_w & \cdots & \langle g_n, g_n \rangle_w \end{bmatrix} \begin{bmatrix} c_0 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} \langle g_0, y \rangle_w \\ \vdots \\ \langle g_n, y \rangle_w \end{bmatrix},$$

from which we can determine $f = c_0 g_0 + \cdots + c_n g_n$. To make the computation of f easier, we could find an orthogonal basis of S.

Remark. (Unnormalized Gram-Schmidt for polynomials) We can apply Gram-Schmidt on $g_0(x) = 1, \dots, g_n(x) = x^n$:

$$\begin{aligned} g_0(x) &= 1, \\ g_1(x) &= x - \frac{\langle x, g_0 \rangle_w}{\langle g_0, g_0 \rangle_w} g_0(x), \\ &\vdots \\ g_k(x) &= x^k - \sum_{j=0}^{k-1} \frac{\langle x^k, g_j \rangle_w}{\langle g_j, g_j \rangle_w} g_j(x) \text{ for } k \le n. \end{aligned}$$

For eack k, $g_k(x)$ is a monic polynomial of degree k. For $0 \le k \le n$, let $p_k(x) = g_k(x)/||g_k||_w$. Then $\{p_k\}_{k=0}^n$ is an orthogonal basis of $\mathbb{P}_n$. Given the inner product above, the monic orthogonal polynomials $g_0, \dots, g_n$ are unique.

Lemma 7.28. Let $r_0, r_1, \dots, r_n$ be orthogonal polynomials in $\langle \rangle_w$, where each r_k has exact degree k. Then for $e \le k \le n$, $xr_{k-1} \perp \mathbb{P}_{k-3}$, i.e., for all $p \in \mathbb{P}_{k-3}$, $\langle xr_{k-1}, p \rangle_w = 0$.

Corollary 7.29. In particular, if $g_0, \dots, g_n$ are the monic orthogonal polynomials in $\langle \rangle_w$, then for $1 \le k \le n$, xg_{k-1} is a monic polynomial of exact degree k, and $xg_{k-1} \perp \mathbb{P}_{k-3}$.

Lemma 7.30. Gram-Schmidt applied to $\{1, x, \dots, xg_{n-1}\}$ yields $\{g_0, \dots, g_n\}$.

Remark. In applying Gram-Schmidt to $\{1, x, \dots, xg_{n-1}\}$, we obtain g_{k-1} just before we need it to get xg_{k-1}. So we can rewrite Gram-Schmidt for $\{1, x, \dots, xg_{n-1}\}$, we get

$$\begin{aligned} g_0(x) &= 1, \\ g_1(x) &= xg_0 x - a_1 g_0(x), \\ &\vdots \\ g_k(x) &= xg_{k-1}(x) - a_k g_{k-1}(x) - b_k g_{k-2}(x) \text{ for } 2 \le k \le n. \end{aligned}$$

Remark. For $k \ge 2$,

$$b_k = \frac{\langle g_{k-1}, g_{k-1} \rangle_w}{\langle g_{k-2}, g_{k-2} \rangle_w}.$$

If we compute $l_j = \langle g_j, g_j \rangle_w$ after getting g_j, then to compute the next g_k, we set

$$a_k = \frac{\langle x g_{k-1}, g_{k-1} \rangle_w}{l_{k-1}} \quad \text{and} \quad b_k = \frac{l_{k-1}}{l_{k-2}};$$

so we just require two inner products.

Theorem 7.31. Given $f \in C[a,b]$ and $\epsilon > 0$, there exists N and $g_N \in \mathbb{P}_N$ such that

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

Theorem 7.32. $\{p_0, p_1, p_2, \ldots\}$ is a complete orthonormal system in $C[a,b]$ with $||\cdot||_w$.

Remark. Recall the Chebyshev polynomials of the first kind: Let $\cos^{-1} : [-1,1] \to [0,\pi]$ be the standard branch of the inverse cosine function, and for $k = 0, 1, 2, \ldots$, define T_k on $[-1,1]$ by $T_k(x) = \cos(k \cos^{-1}(x))$. On $[-1,1]$, T_k is a polynomial of exact degree k, with leading coefficient 2^{k-1} for $k \ge 1$. Moreover, $T_0(x) = 1$, $T_1(x) = x$, and for $k \ge 1$,

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

Theorem 7.33. The Chebyshev polynomials $\{T_0, T_1, \ldots\}$ are orthogonal polynomials in $C[-1,-1]$ with inner product

$$\langle f, g \rangle = \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}}\, dx$$

with $w(x) = 1/\sqrt{1-x^2}$ on $(-1,1)$.

Corollary 7.34. If $\{p_0, p_1, \ldots\}$ are defined by

$$p_0(x) = \frac{1}{\sqrt{\pi}} T_0(x), \qquad p_k(x) = \sqrt{\frac{2}{\pi}} T_k(x) \text{ for } k \ge 1,$$

then $\{p_0, p_1, \ldots\}$ are orthonormal polynomials.

Remark. (Linear Least Squares with Orthogonal Polynomials) Given the inner product $\langle \cdot, \cdot \rangle_w$, let $S = \mathbb{P}_n$, and suppose $r_0, r_1, \ldots, r_n$ are orthogonal polynomials with each r_k having exact degree k, so $\{r_0, \ldots, r_n\}$ is an orthogonal basis of $\mathbb{P}_n$. Given f, let p_n^* be the closest element of $S = \mathbb{P}_n$ to f in $||\cdot||_w$. Then inner product space theory gives

$$p_n^*(x) = \sum_{j=0}^{n} \frac{\langle f, r_j \rangle_w}{\langle r_j, r_j \rangle_w} r_j(x).$$

Theorem 7.35. Given a complete orthonormal system $\{p_0, p_1, p_2, \ldots\}$ in $C[a,b]$ with $\langle \cdot, \cdot \rangle$. Given $f \in C[a,b]$, let $p_n^*(x) = \sum_{j=0}^{n} \langle f, p_j \rangle_w p_j(x)$ be the closest element of $\mathbb{P}_n$ to f. Then $||f - p_n^*||_w \to 0$ as $n \to \infty$.

Definition 7.36. Given the conditions of Theorem 7.35, we say that the series $\sum_{j=0}^{\infty} \langle f, p_j \rangle_w p_j(x)$ converges in norm $||\cdot||_w$ to f, meaning that $||f - \sum_{j=0}^{n} \langle f, p_j \rangle p_j||_w \to 0$ as $n \to \infty$. The coefficients $\langle f, p_j \rangle$ are called *generalized Fourier coefficients* and the series is called the *generalized Fourier series* for f (with respect to $\{p_0, p_1, \ldots\}$.)

Theorem 7.37. Let f have the generalized Fourier series $\sum_{j=0}^{\infty} \langle f, p_j \rangle_w p_j(x)$. Then

$$||f - p_n^*||_w^2 = ||f||_w^2 - ||p_n^*||_w^2 = \sum_{k=0}^{\infty} \langle f, p_k \rangle_w^2 - \sum_{k=0}^{n} \langle f, p_k \rangle_w^2 = \sum_{k=n+1}^{\infty} \langle f, p_k \rangle_w^2.$$

Theorem 7.38. Given $f \in C[a,b]$ and $k \geq 0$, there exists a unique polynomial $\widetilde{p}_k(x) \in \mathbb{P}_k$ that minimizes $||f-p||_\infty = \max_{a\leq x\leq b} |f(x)-p(x)|$ over all $p \in \mathbb{P}_k$.

Remark. Define $E_k(f) = ||f - \widetilde{p}_k||_\infty$. Theorems that give upper bounds for $E_k(f)$ are called Jackson Theorems.

Theorem 7.39. If $f \in C^{k+1}[a,b]$, then

$$E_k(f) \leq \frac{1}{2^k(k+1)!}||f^{(k+1)}||_\infty \left(\frac{b-a}{2}\right)^{k+1}.$$

Theorem 7.40. If $f \in C^m[a,b]$ and $k \geq m$, then

$$E_k(f) \leq \left(\frac{\pi}{2}\right)^M \frac{1}{(k+1)k(k-1)\cdots(k-m+2)}||f^{(m)}||_\infty.$$

Theorem 7.41. Fix $a,b \in \mathbb{R}$ with $a < b$ and $w(x)$ on (a,b) with $\int_a^b w(x)\,dx < \infty$, and let $p_0, p_1, \ldots$ be orthonomal polynomials. Given an $f \in C[a,b]$ and $k \geq 1$, let $\widetilde{p}_{k-1}$ be the closest element of $\mathbb{P}_{k-1}$ to f in $||\cdot||_\infty$. Then $\langle \widetilde{p}_{k-1}, p_k\rangle_w = 0$, and

$$|\langle f, p_k\rangle_w| = |\langle f - \widetilde{p}_{k-1}, p_k\rangle_w| \leq \int_a^b |f(x) - \widetilde{p}_{k-1}(x)||p_k(x)|w(x)\,dx \leq E_{k-1}(f)\int_a^b |p_k(x)|w(x)\,dx.$$

By Cauchy-Schwarz,

$$|\langle f, p_k\rangle_w| \leq E_{k-1}(f)\int_a^b |p_k(x)|w(x)\,dx \leq E_{k-1}(f)\left(\int_a^b w(x)\,dx\right)^{1/2}.$$

Corollary 7.42. Let $T_0, T_1, \ldots$ be the Chebyshev polynomials, and $p_0 = \frac{1}{\sqrt{\pi}}T_0$, $p_k = \sqrt{\frac{2}{\pi}}T_k$ for $k \geq 1$. If $f \in C^2[-1,1]$, then

$$||f - p_n^*||_w^2 \leq \frac{\pi^3||f''||_\infty^2}{6(n-1)^3}.$$

Corollary 7.43. Let $T_0, T_1, \ldots$ be the Chebyshev polynomials, and $p_0 = \frac{1}{\sqrt{\pi}}T_0$, $p_k = \sqrt{\frac{2}{\pi}}T_k$ for $k \geq 1$. If $f \in C^2[-1,1]$, then $||f - p_n^*||_\infty \to 0$ as $n \to \infty$.

Corollary 7.44. Let $T_0, T_1, \ldots$ be the Chebyshev polynomials, and $p_0 = \frac{1}{\sqrt{\pi}}T_0$, $p_k = \sqrt{\frac{2}{\pi}}T_k$ for $k \geq 1$. If $f \in C^2[-1,1]$ and $Q_n(x) \in \mathbb{P}_n$ is the polynomial interpolant of f at the shifted Chebyshev nodes, then $||f - Q||_\infty \to 0$ as $n \to \infty$.

Theorem 7.45. In (a,b) (continuous case) or $x_0, \ldots, x_m$ (discrete case, with $k \leq m$), the orthogonal polynomial q_k has k distinct (and thus simple) zeroes in (a,b) or (x_0, x_m).

7.4 Gaussian Quadrature

Remark. (Gaussian Quadrature) Fix $a,b \in \mathbb{R}$ with $a < b$ and fix $w(x) > 0$ such that w is continuous on (a,b) with $\int_a^b w(x)\,dx < \infty$. Then $C[a,b]$ is an inner product space with inner product $\langle f,g\rangle_w = \int_a^b f(x)g(x)w(x)\,dx$. Suppose we want to construct a quadrature formula to approximate the weighted integral

$$I_w(f) = \int_a^b f(x)w(x)\,dx \text{ for } f \in C[a,b]$$

using $Q_n(f) = \sum_{j=0}^{n} A_j f(x_j)$, chosen so that $Q_n(f) = I_w(f)$. Recall that Q_n is defined to have precision (at least) m if for all $p \in \mathbb{P}_m$, $Q_n(p) = I_w(p)$.

If $x_0 < x_1 < \cdots < x_n \in [a, b]$ are given, we can solve for $A_0, \ldots, A_n$ to make Q_n have precision (at least) n by solving the linear system

$$\begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_n \end{bmatrix} = \begin{bmatrix} I_w(x^0) \\ I_w(x^1) \\ \vdots \\ I_w(x^n) \end{bmatrix}.$$

The matrix is Vandermonde so it is invertible.

Theorem 7.46. Given $x_0 < \cdots < x_n \in [a, b]$, there exists unique $A_0, \ldots, A_n$ such that Q_n has precision at least n.

Remark. Given $x_0 < \cdots < x_n \in [a, b]$, recall that the Lagrange polynomials are defined to be

$$l_k(x) = \prod_{0 \le j \le n} \left(\frac{x - x_j}{x_k - x_j} \right) \text{ for } 0 \le k \le n,$$

and satisfy

$$l_k(x_j) = \delta_{kj} = \begin{cases} 1 & \text{if } k = j \\ 0 & \text{if } k \neq j \end{cases}.$$

If $Q_n(f) = \sum_{j=0}^{n} A_j f(x_j)$ has precision (at least) n, then the solution of the linear system for $A_0, \ldots, A_n$ is explicitly given by $A_j = \int_a^b l_j(x) w(x)\, dx$.

Theorem 7.47. $Q_n(f) = \sum_{j=0}^{n} A_j f(x_j)$ has precision at least $2n + 1$ for $I_w(f) = \int_a^b f(x) w(x)\, dx$.

Corollary 7.48. $2n + 1$ is the maximal precision; Q_n can never have precision $2n + 2$.

Definition 7.49. The unique Q_n of precision $2n+1$ for I_w is called a *Gaussian quadrature*, and we say that Q_n is *Gaussian* for I_w.

Theorem 7.50. If Q_n is Gaussian, then $A_k > 0$ for $0 \le k \le n$.

Theorem 7.51. If Q_n is Gaussian, then $\sum_{j=0}^{n} A_j = \int_a^b w(x)\, dx$.

Corollary 7.52. Given $w(x) > 0$ such that w is continuous on (a, b) with $\int_a^b w(x)\, dx < \infty$, for each n, let Q_n be the Gaussian quadrature with nodes $x_0, \cdots, x_n$ and weights $A_0, \ldots, A_n$ for $I_w(f) = \int_a^b f(x) w(x)\, dx$. Then for all $f \in C[a, b]$, $\lim_{n \to \infty} Q_n(f) = I_w(f)$.

Theorem 7.53. If $f \in C[a, b]$, then

$$|I_w(f) - Q_n(f)| \le 2E_{2n+1}(f) \int_a^b w(x)\, dx.$$

Theorem 7.54. Fix $a, b \in \mathbb{R}$ with $a < b$ and fix $w(x) > 0$ such that w is continuous on (a, b) with $\int_a^b w(x)\, dx < \infty$. Fix n, and let $Q_n(f) = \sum_{j=0} A_j f(x_j)$ be Gaussian for $I_w(f) = \int_a^b f(x) w(x)\, dx$. Define $\langle f, g \rangle_w = \int_a^b f(x) g(x) w(x)\, dx$ as usual, and let $p_0, p_1, \ldots$ be the orthonormal polynomials, and define the

discrete inner product by

$$\langle f, g\rangle_d = \sum_{j=0}^{n} A_j f(x_j) g(x_j).$$

Then $p_0, p_1, \ldots, p_n$ are orthogonal in $\langle \cdot, \cdot \rangle_d$ too.

7.5 Periodic Functions

Definition 7.55. A function $f : \mathbb{R} \to \mathbb{R}$ is called 2π-*periodic* if for all $\theta \in \mathbb{R}$, $f(\theta + 2\pi) = f(\theta)$.

Remark. Let $C_{2\pi}$ denote the vector space of all continuous 2π periodic functions. If we consider the restrictions of functions in $C_{2\pi}$ to $[0, 2\pi]$, we can view $C_{2\pi}$ as $\{f : f \in C[0, 2\pi] \text{ and } f(0) = f(2\pi)\}$. Let $C^k_{2\pi} = \{f \in C_{2\pi} : f \in C^k(\mathbb{R})\}$. Notice that $f \in C_{2\pi}$ if and only if $f, f^{(1)}, \ldots, f^{(k)}$ are all in $C_{2\pi}$. If we restrict functions in $C^k_{2\pi}$ to $[0, 2\pi]$, we can view $C^k_{2\pi}$ a $\{f : f \in C^k[0, 2\pi] \text{ and } f^{(j)}(0) = f^{(j)}(2\pi) \text{ for } 0 \le j \le k\}$.

Definition 7.56. Let $\mathcal{T}_n$ be the vector space of all trigonometric polynomials of degree $\le n$, i.e., linear combinations of $\{1/2, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots, \cos n\theta, \sin n\theta\}$. We use $1/2$ instead of 1 for convenience.

Lemma 7.57. $\mathcal{T}_n$ is a subspace of $C_{2\pi}$ with $\dim(\mathcal{T}_n) = 2n+1$, and $\{1/2, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots, \cos n\theta, \sin n\theta\}$ is a basis.

Theorem 7.58. $\{1/2, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots, \cos n\theta, \sin n\theta\}$ is an *orthogonal basis* of $\mathcal{T}_n$.

$$\langle \cos k\theta, \cos j\theta \rangle = \begin{cases} 2\pi & \text{if } k = j = 0, \\ \pi & \text{if } k = j > 0, \\ 0 & \text{if } k \ne j \text{ and } k \ge 0, j \ge 0, \end{cases},$$

$$\langle \cos k\theta, \sin j\theta \rangle = 0 \text{ for all } k \ge 0, j \ge 1,$$

$$\langle \sin k\theta, \sin j\theta \rangle = \begin{cases} \pi & \text{if } k = j \ge 1, \\ 0 & \text{if } k \ne j \text{ and } k \ge 1, j \ge 1. \end{cases}$$

Given $f \in C_{2\pi}$, let $P^*_n(\theta)$ denote the closest element of $\mathcal{T}_n$ to f in the norm $||\cdot||$ induced by the inner product $\langle \cdot, \cdot \rangle$. Then, by general inner product space theory,

$$P^*_n(\theta) = \sum_{k=0}^{n} \frac{\langle f, \cos k\theta \rangle}{\langle \cos k\theta, \cos k\theta \rangle} \cos k\theta + \sum_{k=1}^{n} \frac{\langle f, \sin k\theta \rangle}{\langle \sin k\theta, \sin k\theta \rangle} \sin k\theta.$$

Thus

$$P^*_n(\theta) = \frac{A_0}{2} + \sum_{k=1}^{n} A_k \cos k\theta + \sum_{k=1}^{n} B_k \sin k\theta,$$

where

$$A_k = \frac{1}{\pi}\langle f, \cos k\theta \rangle = \frac{1}{\pi} \int_0^{2\pi} f(\theta) \cos(k\theta)\, d\theta \text{ for } k \ge 0$$

$$B_k = \frac{1}{\pi}\langle f, \sin k\theta \rangle = \frac{1}{\pi} \int_0^{2\pi} f(\theta) \sin(k\theta)\, d\theta \text{ for } k \ge 0.$$

Lemma 7.59. $\{1/2, \cos\theta, \sin\theta, \cos 2\theta, \sin 2\theta, \ldots\}$ is a *complete orthogonal system* in $C_{2\pi}$. The series

$$\frac{A_0}{2} + \sum_{k=1}^{n} (A_k \cos k\theta + B_k \sin k\theta)$$

is called the *Fourier series* of f. By general inner product space theory, it converges to f in the norm $||\cdot||$ induced by $\langle\cdot,\cdot\rangle$.

Theorem 7.60. If $f \in C^1_{2\pi}$, then the Fourier series of f converges uniformly to f.

Remark. (Discrete Fourier Transform) Fix an integer $m > 0$, let $h = 2\pi/m$, and let $\theta_j = jk$ for $0 \le j \le m-1$, so $\theta_0 = 0, \theta = h, \ldots, \theta_{m-1} = (m-1)h$. are equally spaced in $[0, 2\pi)$ and $mh = 2\pi$ would wrap around by periodicity to be the same as θ_0. Let G_m be the vector space of grid functions $f : \{\theta_0, \ldots, \theta_{m-1}\} \to \mathbb{R}$, and define an inner product $\langle\cdot,\cdot\rangle_d$ on G_m by $\langle f, g\rangle_d = \sum_{j=0}^{m-1} f(\theta_j)g(\theta_j)w_j$, $w_j = 2\pi/m$.

(Case 1) m is even, $m = 2n$. Then $1, \cos\theta, \sin\theta, \ldots, \cos((n-1)\theta), \sin((n-1)\theta), \cos(n\theta)$ is an orthogonal basis of G_m, each of $||\cdot||_d$ have length $\sqrt{\pi}$, except $||1||_d = \sqrt{2\pi}$.

(Case 2) m is odd, $m = 2n+1$. Then $1, \cos\theta, \sin\theta, \ldots, \cos(n\theta), \sin(n\theta)$ is an orthogonal basis of G_m, each of $||\cdot||_d$ of length $\sqrt{\pi}$, exept $||1||_d = \sqrt{2\pi}$.

Definition 7.61. The operator mapping on $f \in G_m$ into its discrete Fourier coefficients $\langle f, \cos k\theta\rangle_d$, $(0 \le k \le n)$ and $\langle f, \sin k\theta\rangle_d$ is called the *discrete Fourier Transform (DFT)*. The Fast Fourier Transform (FFT) is an algorithm that computes the DFT very quickly — in $\mathcal{O}(m\log_2(m))$ operations instead of $\mathcal{O}(m^2)$ operations.

7.6 Complex Inner Product Spaces

Definition 7.62. A complex inner product spaces is a complex vector space V together with an inner product: a function $V \times V$ into $\mathbb{C}$, denoted by $\langle u, v\rangle$, satisfying:

(i) for all $v \in V$, $\langle v, v\rangle \ge 0$; $\langle v, v\rangle$ if and only if $v = 0$;

(ii) for all $\alpha, \beta \in \mathbb{C}$, and for all $u, v, w \in V$, then $\langle \alpha u + \beta v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle$;

(iii) for all $u, v \in V$, $\langle v, u\rangle = \overline{\langle u, v\rangle}$.

Remark. Many definitions and facts about real inner product spaces carry over to complex inner product spaces:

- The norm: $||v|| = \sqrt{\langle v, v\rangle}$;
- Cauchy-Schwarz inequality: $|\langle u, v\rangle| \le ||u|| \cdot ||v||$;
- Pythagorean Theorem: If $\langle u, v\rangle = 0$, then $||u + v||^2 = ||u||^2 + ||v||^2$.
- Orthonormal system: If $\{\varphi_1, \ldots, \varphi_n\}$ is an orthonormal system, then $||\sum_{j=1}^n c_j\varphi_j||^2 = \sum_{j=1}^n |c_j|^2$.
- Bessel's Inequality: If $\{\varphi_1, \ldots, \varphi_n\}$ is an orthonormal system and $v \in V$, then $\sum_{j=1}^n |\langle v, \varphi_j\rangle|^2 \le ||v||^2$. If $\{\varphi_1, \varphi_2, \ldots\}$ is an orthonormal system and $v \in V$, then $\sum_{j=1}^\infty |\langle v, \varphi_j\rangle|^2 = ||v||^2$.

Theorem 7.63. If $\{\varphi_1, \varphi_2, \ldots\}$ is an orthonormal system in a complex inner product space V, then the following two conditions are equivalent:

(i) (Parseval's Equality) For all $v \in V$, $\sum_{j=1}^\infty |\langle v, \varphi_j\rangle|^2 = ||v||^2$.

(ii) For all $v \in V$, then $||v - \sum_{j=1}^\infty \langle v, \varphi_j\rangle\varphi_j|| \to 0$ as $n \to \infty$.

Definition 7.64. An orthonormal system $\{\varphi_1, \varphi_2, \ldots\}$ which satisfies either (i) or (ii) is called a *complete orthonormal system* in V.

Definition 7.65. A function $f : \mathbb{R} \to \mathbb{C}$ is called 2π-*periodic* if for all $x \in \mathbb{R}$, $f(x + 2\pi) = f(x)$. Let $C_{2\pi}$ denote the vector space of all continuous 2π-periodic complex-valued functions $f : \mathbb{R} \to \mathbb{C}$ with complex

scalars. Define an inner product on $C_{2\pi}$ by

$$\langle f, g\rangle = \frac{1}{2\pi}\int_0^{2\pi} f(g)\overline{g(x)}\,dx.$$

Definition 7.66. Let N be a nonnegative integer. Let $\mathcal{T}_N$ be the subspace of $C_{2\pi}$ consisting of all *trigonometric polynomials* of degree $\leq N$ of $\{1/2, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos Nx, \sin Nx\}$.

Lemma 7.67.

(i) $\dim(\mathcal{T}_N) = 2N + 1$;

(ii) $\{1/2, \cos x, \sin x, \cos 2x, \sin 2x, \ldots, \cos Nx, \sin Nx\}$ is an orthogonal basis of $\mathcal{T}_N$;

(iii) $\{e^{i\zeta x} : -N \leq \zeta \leq N\}$ is an orthonormal basis of $\mathcal{T}_N$:

$$\langle e^{i\zeta x, e^{i\eta x}}\rangle = \begin{cases} 1 & \text{if } \zeta = \eta, \\ 0 & \text{if } \zeta \neq \eta \end{cases} \qquad \text{for } \zeta, \eta \in \mathbb{Z}.$$

Definition 7.68. For $f \in C_{2\pi}$ and $\zeta \in \mathbb{Z}$, define the *Fourier coefficients* of f:

$$\hat{f}(\zeta) = \langle f, e^{i\zeta x}\rangle = \frac{1}{2\pi}\int_0^{2\pi} e^{-i\zeta x} f(x)\,dx.$$

The formal series $\sum_{\zeta=-\infty}^{\infty} \hat{f}(\zeta)e^{i\zeta x}$ is called the *Fourier series* of f. Let $S_n = \sum_{\zeta=-N}^{N} \hat{f}(\zeta)e^{i\zeta x}$ denote the Nth partial sum of the Fourier series of f, so $S_n \in \mathcal{T}_N$.

Theorem 7.69.

(i) $\{e^{i\zeta x} : -\infty \leq \zeta \leq \infty\}$ is complete orthonormal system in $C_{2\pi}$;

(ii) Parseval's Relation: $\sum_{\zeta=-\infty}^{\infty} |\hat{f}(\zeta)|^2 = ||f||^2$;

(iii) $||f - S_n||^2 \to 0$ as $N \to \infty$, so the Fourier series of f converges to f in the norm $||g|| = \left(\frac{1}{2\pi}\int_0^{2\pi} |g(x)|^2\,dx\right)^{1/2}$.

(iv) It can be shown that if $f \in C_{2\pi}$ and $f' \in C_{2\pi}$, then $S_n \to f$ uniformly, i.e. $\max_{x\in\mathbb{R}} |f(x) - S_N(x)| \to 0$ as $N \to \infty$.

Definition 7.70. Define the translation operator $T_h : C_{2\pi} \to C_{2\pi}$ by $T_h(f)(x) = f(x + h)$. For example, T_h applied to $e^{i\zeta x}$ is $ei\zeta(x + h)$.

Theorem 7.71. Let $g = T_h f$, i.e. $g(x) = f(x + h)$ Then $hatg(\zeta) = e^{i\zeta h}\hat{f}(\zeta)$.

Definition 7.72. Fix an integer $M > 0$. For each integer ν, define the grid point $x_\nu = 2\pi\nu/M$. Let the grid $G_M = \{x_\nu : \nu \text{ is an integer}\}$. Let $h = 2\pi/M$, so $x_\nu = \nu h$. A function $f : G_M \to \mathbb{C}$ is called 2π-*periodic* if $f(x + 2\pi) = f(x)$ for all $x \in G_M$ since $x_\nu + 2\pi = 2\pi\nu/M + 2\pi = 2\pi(\nu + M)/M = x_{\nu+M}$, this condition becomes $f(x_{\nu+M}) = f(x_\nu)$ for all integers ν.

Definition 7.73. Let G_m be the complex vector space of all 2π-periodic grid functions $f : G_M \to \mathbb{C}$. Since the values of f at the M grid points $\{x_\nu : 0 \leq \nu \leq M - 1\}$ are independent of each other and they determine f completely, e.g. $f(x)_{-2} = f(x_{M-2})$, $f(x_{-1}) = f(x_{M-1})$, $f(x_M) = f(x_0)$. We will view G_M as the vector space of all functions $f : \{\theta_0, \ldots, \theta_{M-1}\} \to \mathbb{C}$. So G_m is an M-dimensional vector space.

Definition 7.74. Define an inner product on G_M by

$$(f, g)_M = \frac{1}{M}\sum_{\nu=0}^{M-1} f(x_\nu)\overline{g(x_\nu)}.$$

Define the norm on G_M by

$$||f||_M = \sqrt{(f,f)_M} = \frac{h}{2\pi} \sum_{\nu=0}^{M-1} |f(x_\nu)|^2 \,() \quad .^{1/2}$$

For $f \in G_M$, and all integers ζ, define

$$\hat{f}_M(\zeta) = (f, e^{i\zeta x})_M = \frac{1}{M} \sum_{\nu=0}^{M-1} e^{-i\zeta x_\nu} f(x_\nu).$$

Theorem 7.75.

(i) If $\zeta \equiv \nu \pmod{M}$, then $e^{i\zeta x}$ and $e^{i\nu x}$ restrict to the same grid function in G_M.

(ii) For any integer ζ, η:

$$(e^{i\zeta x}, e^{i\eta x}) = \begin{cases} 1 & \text{if } \zeta \equiv \eta \pmod{M} \\ 0 & \text{if } \zeta \not\equiv \eta \pmod{M}. \end{cases}$$

(iii) Suppose M is odd, say $M = 2N+1$. Then $\{e^{i\zeta x} : -N \le \zeta \le N\}$ is an orthonormal basis of G_M.

(iv) Suppose M is even, say $M = 2N$. Then e^{iNx}, e^{-iNx}, and $\cos Nx$ all restrict to the same grid function, $\{e^{i\zeta x} : -N \le \zeta \le N\}$ is an orthonormal basis of G_M, and $\{e^{i\zeta x} : -N \le \zeta \le N\} \cup \{\cos Nx\}$ is an orthonormal basis of G_M.

(v) If $\zeta \equiv \eta \pmod{M}$, then $e^{i\zeta x}$ and $e^{i\eta x}$ restrict to the same grid functions, so

$$\hat{f}_M(\zeta) = (f, e^{i\zeta x})_M = (f, e^{i\eta x})_M = \hat{f}_M(\eta).$$

In particular, $\hat{f}_M(\zeta + M) = \hat{f}_M(\zeta)$. So $\hat{f}_M(\zeta)$ is an M-periodic function on ζ.

(vi) (Inversion formula) If M is odd, say $M = 2N+1$. Then for any $f \in G_M$,

$$f(x_\nu) = \sum_{\zeta=-N}^{N} \hat{f}_M(\zeta) e^{i\zeta x_\nu} \qquad \nu = 0, 1, \ldots, M-1.$$

If M is even, say $M = 2N$. Then for any $f \in G_M$,

$$f(x_\nu) = \sum_{\zeta=-N+1}^{N} \hat{f}_M(\zeta) e^{i\zeta x_\nu} \qquad \nu = 0, 1, \ldots, M-1.$$

In general,

$$f(x_\nu) = \sum_{\zeta=-N+1}^{N} \hat{f}_M(\zeta) e^{i\zeta x_\nu} + \hat{f}_M(N) \cos(N x_\nu), \qquad \nu = 0, 1, \ldots, M-1.$$

(vii) Recall $h = 2\pi/M$. If $g = T_h f$, then $\hat{g}_M(\zeta) = e^{i\zeta h} \hat{f}_M(\zeta)$. If $g = T_{-h} f$, then $\hat{g}_M(\zeta) = e^{-i\zeta h} \hat{f}_M(\zeta)$.

Definition 7.76 (Discrete Fourier Transform). Fix an integer M. Let $f \in G_M$ be a grid function. Define an M-vector

$$\mathbf{f}_M = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \cdots \\ f(x_{M-1}) \end{bmatrix}.$$

Using $\{e^{i\zeta x} : 0 \le \zeta \le M-1\}$ as an orthonormal basis of G_M, define the M-vector

$$\hat{\mathbf{f}}_M = \begin{bmatrix} \hat{f}(x_0) \\ \hat{f}(x_1) \\ \cdots \\ \hat{f}(x_{M-1}) \end{bmatrix}.$$

Let $\omega = e^{-(2\pi i)/M}$. Then

$$\hat{f}_M(\zeta) = \frac{1}{M}\sum_{\nu=0}^{M-1} e^{-i\zeta x_\nu} f(x_\nu) = \frac{1}{M}\sum_{\nu=0}^{M-1} \omega^{\zeta\nu} f(x_\nu).$$

So $\hat{\mathbf{f}}_M = W_M \mathbf{f}_M / M$ where W_M is a $M \times M$ matrix with

$$W_M = \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{M-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(M-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^{M-1} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix}.$$

The mapping $F_M : \mathbb{C}^M \to \mathbb{C}^M$ given by $F_M : \mathbf{f}_M \to \hat{\mathbf{f}}_M$ is called the *discrete Fourier transform.*

Note that $\overline{\omega} = \cos(2\pi/) + i\sin(2\pi/M) = e^{2\pi i/M}$. Since

$$f(x_\nu) = \sum_{\zeta=0}^{M-1} \hat{(f_M)}(\zeta) e^{i\zeta x_\nu} = \sum_{\zeta=0}^{M-1} \overline{\omega}^{\zeta\nu} \hat{f}_M(\zeta)$$

for $\nu = 0, \ldots, M-1$ In addition, $W_M = W_M^\top$,

$$\mathbf{f}_M = \overline{W}_M \hat{\mathbf{f}}_M = \overline{W}_M^\top \hat{\mathbf{f}}_M.$$

The mapping $F_M^{-1} : \mathbb{C}^M \to \mathbb{C}^M$ given by $F_M^{-1} : \hat{\mathbf{f}}_M \to \hat{\mathbf{f}}_M$ is called the *inverse discrete Fourier transform.* In particular, $F_M^{-1} \circ F_M = I$. Then $\left(\overline{W}_M^\top\right)\left(\frac{1}{M} W_M\right) = I$, so $\left(\frac{1}{\sqrt{M}} W_M\right)^\top \left(\frac{1}{\sqrt{M}} W_M\right) = I$. Then $\frac{1}{\sqrt{M}} W_M$ is a unitary matrix.

8 Special Topics

8.1 Singular Value Decomposition

Theorem 8.1. Let $A \in \mathbb{C}^{m\times n}$ have rank r. Then there exist unitary matrices $U \in \mathbb{C}^{n\times n}$, $V \in \mathbb{C}^{m\times m}$ such that

$$V^H A U = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Sigma \in \mathbb{R}^{r\times r}$ is a diagonal matrix with elements $\sigma_1, \ldots, \sigma_r$ such that $\sigma \ge \cdots \ge \sigma_r > 0$.

Definition 8.2. Writing A as

$$V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^H$$

is called the *singular value decomposition* (SVD) of A, and the values $\sigma_1, \ldots, \sigma_n$ are called the singular values of A. The columns of U are called *right singular vectors* of A. The rows of V^H are called the *left singular vectors* of A.

Lemma 8.3. Let $A \in \mathbb{C}^{m\times n}$ have rank r. The singular values of A are unique.

Lemma 8.4. Let $A \in \mathbb{C}^{m\times n}$ have rank r with a singular value decomposition

$$A = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^H.$$

Then

$$\begin{cases} Au_i = \sigma_i v_i & \text{for } i = 1, \dots, r; \\ Au_i = 0 & \text{for } i = r+1, \dots, n. \end{cases}$$

Theorem 8.5. If $A \in \mathbb{C}^{n\times n}$ is Hermitian with eigenvalues $\lambda_1, \dots, \lambda_n$, then the singular values of A are $|\lambda_1|, \dots, |\lambda_n|$.

Theorem 8.6. Let $A \in \mathbb{C}^{m\times n}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_n$. Then $||A||_2 = \sigma_1$ and $||A||_F^2 = \sigma_1^2 + \cdots + \sigma_n^2$ where $||\cdot||_F$ is the Frobenius norm.

Corollary 8.7. $||A||_2 = \sqrt{\rho(A^H A)}$ where ρ is the spectral radius of A.

Theorem 8.8. Let $A \in \mathbb{C}^{m\times n}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_n$. Then for each k,

$$\sigma_k = \min_{\dim S = n-k+1} \left(\max_{x\in S, x\neq 0} \frac{||Ax||_2}{||x||_2} \right) \qquad \text{(minimax)},$$
$$\sigma_k = \max_{\dim S = k} \left(\min_{x\in S, x\neq 0} \frac{||Ax||_2}{||x||_2} \right) \qquad \text{(maximin)}$$

where S is an arbitrary subspace of $\mathbb{C}^n$.

Theorem 8.9. Let $A, B \in \mathbb{C}^{m\times n}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_n$ and $\tau_1 \geq \cdots \geq \tau_n$. Then

$$|\sigma_k - \tau_k| \leq ||A - B||_2 \text{ for } k = 1, \dots, n.$$

Theorem 8.10. Let $A \in \mathbb{C}^{m\times n}$ have the singular value decomposition

$$A = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^H.$$

Fix $b \in \mathbb{C}^m$ and consider the linear least squares problem: if

$$y = U \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^H b \in \mathbb{C}^n$$

Then y minimizes $||b - Ax||_2^2$ over $x \in \mathbb{C}^n$. Moreover, y is the unique minimum element: if $||b - Ax||_2^2 = ||b - Ay||_2^2$ and $x \neq y$, then $||x||_2 > ||y||_2$.

Definition 8.11. The *pseudo-inverse* of A is

$$A^+ = U \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{bmatrix} V^H.$$

Lemma 8.12.

(i) $y = A^+ b$ is the solution of LLS problem of smallest norm;

(ii) A^+ does not depend on U, V because of (i);

(iii) If $\text{null}(A) = \{0\}$, it can be shown that $A^+ = (A^H A)^{-1} A^H$;

(iv) If $A \in \mathbb{C}^{n\times n}$ is invertible, then $A^{-1} U \Sigma^{-1} V^H$, so $||A^{-1}||_2 = 1/\sigma$;

(v) So $\kappa_2(A) = ||A||_2 \cdot ||A^{-1}||_2 = \sigma_1/\sigma_2$ if $A \in \mathbb{C}^{n\times n}$ is invertible.

Theorem 8.13. Let $A, B \in \mathbb{C}^{m\times n}$ have singular values $\sigma_1 \geq \cdots \geq \sigma_n$ and $\tau_1 \geq \cdots \geq \tau_n$. Then $\sum_{k=1}^n (\sigma_k - \tau_k)^2 \leq ||A - B||_F^2$.

Theorem 8.14. Suppose $A \in \mathbb{C}^{m\times n}$ has rank r, and s is an integer with $1 \leq s < r$. Then $\min_{\text{rank}(B)=s} ||A - B||_2 = \sigma_{s+1}$ and $\min_{\text{rank}(B)=s} ||A - B||_F = \sqrt{\sigma_{s+1}^2 + \cdots + \sigma_r^2}$, where $\sigma_1 \geq \cdots \geq \sigma_n$ are the singular values of A. Moreover, if the SVD of A is

$$A = V \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} U^H$$

the minimum for both norms for $B = A'$ where

$$A' = V \begin{bmatrix} \Sigma' & 0 \\ 0 & 0 \end{bmatrix} U^H,$$

where $\Sigma' = \text{diag}(\sigma_1, \ldots, \sigma_s) \in \mathbb{C}^{s\times s}$.

Theorem 8.15. Given any $A \in \mathbb{C}^{m\times n}$ and $\epsilon > 0$, there exists $B \in \mathbb{C}^{m\times n}$ of full rank such that $||A - B||_2 \leq \epsilon$.

8.2 Lanczos Method

Definition 8.16. Let $A \in \mathbb{C}^{n\times n}$ and $x \neq 0 \in \mathbb{C}^n$. The *grade* of x with respect to A is the smallest positive integer m such that $\{x, Ax, \ldots, A^m x\}$ is linearly dependent.

Lemma 8.17. Let $A \in \mathbb{C}^{n\times n}$ and $x \neq 0 \in \mathbb{C}^n$, and let m be the grade.

(i) $m \leq n$;

(ii) there exist constants $\gamma_0, \gamma_1, \ldots, \gamma_{m-1}$ such that $A^m x + \gamma_{m-1} A^{m-1} x + \cdots + \gamma_0 x = 0$;

(iii) $\gamma_0, \gamma_1, \ldots, \gamma_{m-1}$ are unique.

Definition 8.18. The *minimum polynomial* of x to A is

$$\pi_{A,x}(\lambda) = \lambda^m + \gamma_{m-1}\lambda^{m-1} + \cdots + \lambda_0.$$

Corollary 8.19. $\pi_{A,x}(\lambda)$ divides $\pi_A(\lambda)$.

Definition 8.20. For $s = 1, 2, \ldots$ define $M_s(x) = \text{span}\{x, Ax, \ldots, A^{s-1}x\}$. The sequence $x, Ax, A^2x, \ldots$ is called a Kyrlov sequence. The sequence $M_1, M_2, \ldots$ is called Kyrlov sequence of subspaces.

Lemma 8.21.

(i) $M_1(x) \subset M_2(x) \subset \cdots \subset M_m(x) = M_{m+1}(x) = \cdots$

(ii) $\dim M_s(x) = \min(s, m)$.

Theorem 8.22. M_m is an invariant subspace of A. The matrix of A with respect to the basis $\{x, Ax, \ldots, A^{m-1}x\}$

of M_m is

$$\begin{bmatrix} 0 & 0 & \cdots & 0 & -\gamma_0 \\ 1 & 0 & \cdots & 0 & -\gamma_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -\gamma_{m-1}. \end{bmatrix}.$$

Corollary 8.23. If Y is an invertible matrix such that

$$\operatorname{span}\{y_1, \ldots, y_m\} = \operatorname{span}\{x, Ax, \ldots, A^{m-1}x\},$$

then $Y^{-1}AY$ is block upper triangular with $m \times m$ and $(n-m) \times (n-m)$ diagonal blocks.

Theorem 8.24. If A is derogatory, then every vector $x \neq 0 \in \mathbb{C}^n$ has grade $< n$.

Theorem 8.25. Suppose A has distinct eigenvalues. Then x is a grade n if and only if x is note a linear combination of $< n$ eigenvectors.

Corollary 8.26. If A has distinct eigenvalues, the set of vectors x of grade less than n is the union of n subspaces, each of $\dim n - 1$.

Remark. (Lanczos Method) We will consider only the variant appropriate for real symmetric matrices $A \in \mathbb{R}^{n \times n}$. This is a proess for reducing A to the diagonal form using Krylov sequences. Suppose $\mathbf{x} \neq 0 \in \mathbb{R}^n$ is of grade m. Let $\mathbf{v}_1 = x/||x||_2$. Suppose that $\{v_1, \ldots, v_k\}$ have been constructed where $1 \leq k \leq m$. Let

$$\begin{aligned} a_k &= v_k^\top A v_k, \\ w_{k+1} &= Av_k - \alpha_k v_k - \beta_{k-1} v_{k-1} \quad (\text{where } \beta_c = 0), \\ \beta_k &= ||w_{k+1}||_2, \\ v_{k+1} &= w_{k+1}/\beta_k \quad (\text{if } k < m). \end{aligned}$$

Lemma 8.27. For $k = 1, \ldots, m$, $\{v_1, \ldots, v_k\}$ is an orthogonal basis of M_k.

Lemma 8.28. $w_{m+1} = 0$.

Theorem 8.29. If $1 \leq s < m$, then

$$AV_s = V_s T_s + \beta_s v_{s+1} e_s^\top.$$

If $s = m$, then

$$AV_m = V_m T_m.$$

Remark. In practice, the Lanczos method was originally proposed as a way of computing a tridiagonal matrix similar to A, but numerically , the v_j's lack orthogonality due to round-off error. However, the method can be used to get approximate eigenvalues. This may be effective when n is large and $s << n$.

Theorem 8.30. Suppose $z = [\zeta_1, \ldots, \zeta_s]^\top$ with $||z||_2 = 1$ is an eigenvalues of T_s corresponding to eigenvalue μ. Let $y = V_s z$. Then $||y||_2 = 1$ and $||Ay - \mu y|| = \beta_s |\zeta_s|$.

8.3 Conjugate Gradient Method

Remark. Suppose $Q \in \mathbb{R}^{n \times n}$ is symmetric and positive definite, and $b \in \mathbb{R}^n$. Consider minimizing

$$\phi(x) = \frac{1}{2} x^\top Q x - b^\top x$$

over $x \in \mathbb{R}^n$. Then $\Delta\phi(x) = Qx - b$, so the unique minimum of ϕ is at the solution x^* of the linear equation $Qx = b$.

Definition 8.31. Two vectors p_1, p_2 are called Q-conjugate if $p_1^\top Q p_2 = 0$.

Theorem 8.32. If $p_1, \ldots, p_k$ are Q-conjugate (pairwise) and nonzero, they are linearly independent.

Remark. Suppose $p_1, \ldots, p_n$ are nonzero and Q-conjugate. Represent $x^* = \alpha_1 p_1 + \cdots + \alpha_n p_n$. Then $p_j^\top b = p_j^\top Q x^* = \alpha_j p_j^\top Q p_j$, so $\alpha_j = (p_j^\top b)/(p_j^\top A p_j)$. Note that α_j can be evaluated from b without knowing x^*; x^* can be obtained from

$$x^* = \sum_{j=1}^{n} \frac{p_j^\top b}{p_j^\top Q p_j} p_j.$$

Theorem 8.33 (Conjugate Direction Theorem). Suppose $p_1, \ldots, p_n$ are nonzero and Q-conjugate. Let $x_1 \in \mathbb{R}^n$. Generate a sequence $x_1, \ldots, x_{n+1}$ by

$$x_{k+1} = x_k + \gamma_k p_k \text{ where } \gamma_k = -\frac{g_k^\top p_k}{p_k^\top Q p_k}$$

and $g_k = (\Delta\phi)(x_k) = Qx_k - b$. Then $x_{n+1} = x^*$.

Theorem 8.34. Let $x_1, \ldots, x_{n+1}$ be the sequence in Theorem 8.33, and let $\mathcal{B}_k = \operatorname{span}\{p_1, \ldots, p_n\}$.

(i) x_{k+1} minimizes $\phi(x)$ on the line $x = x_k + \gamma p_k$ for $\gamma \in \mathbb{R}$; and

(ii) x_{k+1} minimizes $\phi(x)$ on the affine subspace $x_1 + \mathcal{B}_k$.

Corollary 8.35. For $1 \le j < k \le n+1$, $g_k^\top p_j = 0$.

Remark. Suppose $p_1, \ldots, p_l$ are nonzero, Q-conjugate where $l < n$. Given x_1, generate $x_1, \ldots, x_{l+1}$, $\gamma_1, \ldots, \gamma_l$, and $g_1, \ldots, g_{l+1}$ as in the Conjugate Direction Theorem. The proof of the expanding subspace theorem implies $g_k \perp \mathcal{B}_{k-1}$ for $k = 1, \ldots, l+1$.

Theorem 8.36 (Conjugate Gradient Algorithm). Let $x_1 \in \mathbb{R}^n$. Choose $p_1 = -g_1 = b - Qx_1$ (if $p_1 = 0$, step: $x_1 = x^*$). Let $\gamma_1 = -(g_1^\top p_1)/(p_1^\top Q p_1)$, $x_2 = x_1 + \gamma_1 p_1$, $g_2 = \Delta\phi(x_2) = Qx_2 - b$, $\zeta_1 = (g_2^\top Q p_1)/(p_1^\top Q p_1)$, and $p_2 = -g_2 + \zeta_1 p_1$ (if $p_2 = 0$, stop: $g_2 = 0$, so $x_2 = x^*$). For $k \ge 2$, suppose p_k is nonzero. Let $\gamma_k = -(g_k^\top p_k)/(p_k^\top Q p_k)$, $x_{k+1} = x_k + \gamma_k p_k$, $g_{k+1} = \Delta\phi(x_{k+1}) = Qx_{k+1} - b$, $\zeta_k = (g_{k+1}^\top Q p_k)/(p_k^\top Q p_k)$, and $p_{k+1} = -g_{k+1} + \zeta_k p_k$ (if $p_{k+1} = 0$, stop: $g_{k+1} = 0$, so $x_{k+1} = x^*$).

Theorem 8.37. It is clear that $g_{k+1} = 0$ implies $p_{k+1} = 0$, and the algorithm terminates. The following theorem shows that $p_{k+1} = 0$ implies $g_{k+1} = 0$, and the conjugate gradient algorithm is a conjugate direction method.

Theorem 8.38 (Conjugate Gradient Theorem). Suppose $k \ge 1$ and $p_k \ne 0$. Then

(i) $\operatorname{span}\{g_1, \ldots, g_k\} = \operatorname{span}\{p_1, \ldots, p_k\} = \operatorname{span}\{g_1, Qg_1, \ldots, Q^{k-1} g_1\}$;

(ii) $p_k^\top Q p_j = 0$ for $1 \le j \le k-1$;

(iii) $\gamma_k = (g_k^\top g_k)/(p_k^\top Q pk)$ and $\zeta_k = (g_{k+1}^\top g_{k+1})/(g_k^\top g_k)$.

The last set of (i) is clearly $\operatorname{span}\{p_1, Qp_1, \ldots, Q^{k-1} p_1\}$;

Corollary 8.39. If $p_1, \ldots, p_l$ are nonzero, then $g_1, \ldots, g_l$ are nonzero, and $g_k^\top g_j = 0$ for $1 \le j < k \le l+1$; so $\{g_1, \ldots, g_k\}$ is an orthogonal basis of $\operatorname{span}\{g_1, Qg_1, \ldots, Q^{k-1} g_1\}$ for $k = 1, \ldots, l$.

Corollary 8.40. Suppose that Lanczos algorithm is applied to Q with starting vector g_1, and let $m = \text{grade}(g_1)$. Then $p_1, \ldots, p_m$ in the conjugate gradient are all nonzero, and $p_{m+1} = g_{m+1} = 0$ for $j = 1, \ldots, m$; g_j is a multiple of v_j in Lanczos.