Numerical Analysis - Math 464 and 465 Notes

Brett Saiki

March 2022

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

Contents

1 Floating Point and Roundoff Error

1.1 Number Representation

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

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

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

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

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

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

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

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

Remark.

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

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

1.2 Normalized Scientific Notation in Base β

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

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

with $d_1 \neq 0$.

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

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

with $d_1 \neq 0$.

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

1.3 Floating Point Arithmetic

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

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

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

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

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

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

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

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

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

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

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

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

1.4 Absolute and Relative Error

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

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

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

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

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

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

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

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

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

1.5 Arithmetic Operations with Floating-Point Numbers

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

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

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

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

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

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

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

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

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

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

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

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

This kind of analysis is called backward error analysis.

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

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

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

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

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

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

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

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

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

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

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

and similarly,

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

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

1.6 Converting Between Bases

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

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

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

where each operation is in base β arithmetic.

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

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

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

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

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

where each operation is in base β arithmetic.

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

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

2 Solutions of Linear Systems

2.1 Solutions of Linear Systems using Elimination

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

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

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

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

This method of solving is called *back subtitution.*

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

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

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

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

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

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 Pivoting

Definition 2.5. 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, $\dots$, n-th equations.

Theorem 2.6. 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 2.7. 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 2.8. An invertible matrix A has an LU-factorization if and only if each of the upper left hand submatrices

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

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

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

(i) Do not choose near-zero pivots.
(ii) Cannot just use absolute comparison of $a_{ik}^{(k-1)}$.
(iii) The best pivot maximizes the ratio of the size of pivot entry to the size of the row.

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

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

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

Definition 2.9. The technique of *simple pivoting* involves choosing the pivot row with the smallest $I \geq k$ for which $A_{Ik}^{(k-1)} \neq 0$, and interchanging the k-th row and the I-th row.

Definition 2.10. 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 2.11. The technique of *scaled partial pivoting* involves computing scale factors for each row:

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

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

Definition 2.12. In *total pivoting*, the columns are also interchanged. At the k-th step, choose $I \geq k$ and $J \geq k$ for which $|a_{IJ}^{(k-1)}|$ is the maximum of $|a_{ij}|$ for $i = k, \ldots, n$ and $j = k, \ldots, n$. Interchange the k-th row and the I-th row and interchange the k-th column and the J-th column.

Lemma 2.13. 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: $\sum_{k=1}^{n-1} \left[(n-k+1)^2 - 1\right] \approx n^3/3$.

2.3 Interchanging

Theorem 2.14. 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 2.15. 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 2.16. 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.

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

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

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

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

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

Definition 2.26. Define the Frobenius norm of A to be

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

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

2.5 Residual Error

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

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

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

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

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

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

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

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

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

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

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

then the following is usually true,

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

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

2.6 General Iterative Methods

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

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

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

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

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

(i) $I - M$ is invertible,

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

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

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

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

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

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

Lemma 2.39. For the methods descibed above,

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

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

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

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

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

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

This is equivalent to the equation:

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

for $1 \leq i \leq n$ 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