Editor's Note. Some theorems and lemmas are not included as they were not covered in lecture or homework. Developed from *Analysis of Numerical Methods* by Isaacson and Keller as well as notes based off of the lectures of Prof. Olga Holtz of UC Berkeley. All content is copyright of its respective author(s).

1 Norms, Arithmetic, and Well-Posed Computations

1.1 Norms of Vectors and Matrices

Theorem 1. For any square matrix A of order n there exists a non-singular matrix P, of order n, such that

$$B = P^{-1}AP$$

is upper triangular and has the eigenvalues of A, say $\lambda_j := \lambda_j(A), j = 1, 2, \ldots, n$, on the principal diagonal.

Remark. The above theorem can also be stated as: Any matrix A of order n is similar to an upper-triangular matrix B.

Definition 1. A map $\|\cdot\| : \mathcal{V} \to \mathbb{R}$ (also denoted as N) is a vector norm if it satisfies all of the following properties:

(i) $\|\mathbf{x}\| \geq 0$ for all $\mathbf{x} \in \mathcal{V}$ and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = \mathbf{0}$.

(ii) $\|\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$ for all scalars α and all $\mathbf{x} \in \mathcal{V}$.

(iii) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ (The triangle inequality).

Definition 2. The p-norm of a vector $\mathbf{x}$ is defined as

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p} \qquad p \geq 1$$

where x_i are the entries of $\mathbf{x}$.

Remark. Note that the case of $p = 2$ is known as the *Euclidean norm.*

Remark. Note that the case of $p \to \infty$, the norm is the same as

$$\|\mathbf{x}\|_\infty = \max_i |x_i|$$

Lemma 1. Every vector norm $N(\mathbf{x})$ is a continuous function of the components of $\mathbf{x}$.

Theorem 2. For each pair of vector norms $N(\mathbf{x})$ and $N'(\mathbf{x})$, there exist positive constants m and M such that for all $\mathbf{x} \in \mathbb{C}^n$

$$mN'(\mathbf{x}) \leq N(\mathbf{x}) \leq MN'(\mathbf{x})$$

Definition 3. A matrix norm is a map $\|\cdot\| : \mathcal{M} \to \mathbb{R}$ that satisfies the conditions (i)–(iii) from the definition of vector norm and also satisfies

(iv) $\|AB\| \leq \|A\| \cdot \|B\|$

for any (square) matrices of order n.

Definition 4. The induced norm (also known the a natural norm) for a given vector norm is defined thusly: if $\|\cdot\|$ is a vector norm, the induced norm for a matrix A is

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

or, equivalently,

$$\|A\| := \max_{\|\mathbf{u}\|=1} \|A\mathbf{u}\|.$$

Remark. Note that this implies

(A) $\|A\|_\infty$ is the maximum absolute row sum and

(B) $\|A\|_1$ is the maximum absolute column sum.

Definition 5. The spectral radius of any square matrix A is

$$\rho(A) := \max_s |\lambda_s(A)|$$

where λ_s is the sth eigenvalue of A.

Remark. Lemma 1 and Theorem 2 also apply identically to matrix norms.

Lemma 2. For any induced norm $\|\cdot\|$ and square matrix A,

$$\rho(A) \leq \|A\|$$

Theorem 3. For each nth order matrix A and arbitrary $\epsilon > 0$, an induced norm $\|\cdot\|$ can be found such that

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon$$

Definition 6. A matrix A for which

$$\lim_{m \to \infty} A^m = 0$$

is said to be convergent.

Theorem 4. The following three statements are equivalent:

(a) A is convergent

(b) $\lim_{m \to \infty} \|A^m\| = 0$ for some matrix norm

(c) $\rho(A) < 1$

Corollary. The matrix A is convergent if for some matrix norm

$$\|A\| < 1$$

Theorem 5. (a) The geometric series

$$I + A + A^2 + A^3 + \cdots,$$

converges iff A is convergent.

(b) If A is convergent, then $I - A$ is non-singular and

$$(I - A)^{-1} = I + A + A^2 + A^3 + \cdots$$

Corollary. If for some natural norm $\|A\| < 1$, then $I - A$ is non singular and

$$\frac{1}{1 + \|A\|} \leq \left\|(I - A)^{-1}\right\| \leq \frac{1}{1 - \|A\|}$$

1.2 Floating-Point Arithmetic and Rounding Errors

Definition 1. If the number $a \neq 0$ has the exact decimal representation

$$a = \pm 10^q (.d_1 d_2 \cdots)$$

then the t-digit floating decimal representation of a is

$$\text{fl}(a) \coloneqq \pm 10^q (.\delta_1 \delta_2 \cdots \delta_t)$$

Remark. In the previous definition, $(.\delta_1 \delta_2 \cdots \delta_t)$ and q are known as the *mantissa* and *exponent* of $\text{fl}(a)$, respectively. The exponent q is generally bounded by $-N$ and M where N and M are large positive integers determined by the machine.

Lemma 1. The error in t-digit floating decimal representation of a number $a \neq 0$ is bounded:

$$|a - \text{fl}(a)| \leq K|a|10^{-t}$$

where

$$K = \begin{cases} 5 & \text{(rounding)} \\ 10 & \text{(chopping)} \end{cases}$$

Algorithm (Floating-point addition)**.** If adding two numbers, each at t-digit decimal precision, the smaller of the two numbers is shifted so that the exponent of the two numbers are the same (with the extra precision rounded or truncated). The mantissa of the numbers are then summed and either rounded or chopped.

Definition 2. Guard digits are extra digits used to increase the accuracy of subtractions: if subtracting one number of t-digit decimal precisision from another and using s guard digits, the computer would shift and subtact as though it were using $(t+s)$-digit decimal precision, and then round to t-digit decimal precision.

1.3 Well-Posed Computations

Definition 1. A function f is Lipschitz continuous if there exists a constant M such that for any x and y

$$|f(x) - f(y)| \leq M|x - y|$$

Remark. If f is differentiable on the interval $[x, y]$, then we can take

$$M = \sup_{\xi \in [x,y]} \left|f'(\xi)\right|$$

Definition 2. A problem (e.g. $A\mathbf{x} = \mathbf{f}$) is well-posed if it satisfies all of the following:

(a) It has a solution (in the example, x) for every input (in the example, f).

(b) The solution is unique.

(c) The solution depends Lipschitz continuously on the input.

2 Numerical Solution of Linear Systems and Matrix Inversion

2.1 Gaussian Elimination

Definition 1. If A is non-singular, then a solution $\mathbf{x}$ to the equation $A\mathbf{x} = \mathbf{f}$ exists and is unique for any $\mathbf{f}$. If A is singular, then the solution $\mathbf{x}$ may not exist or be unique.

Algorithm (Gaussian Elimination/LU Decomposition)**.** Consider the problem $A\mathbf{x} = \mathbf{f}$ where A is an $n \times n$ matrix and $\mathbf{x}$ and $\mathbf{f}$ both have n entries. Consider the subproblem $A^{(k-1)}\mathbf{x} = \mathbf{f}^{(k-1)}$ where $A^{(k-1)}$ is both upper triangular and has non-zero diagonal elements for its first $k-1$ rows. We define $A^{(1)} \coloneqq A$ and $\mathbf{f}^{(1)} \coloneqq \mathbf{f}$. If we denote the entries of $A^{(k)}$ and $\mathbf{f}^{(k)}$ as $a_{ij}^{(k)}$ and $f_i^{(k)}$ respectively, then we may generate them with the following method:

$$a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{for } i \leq k-1 \\ 0 & \text{for } i \leq k, j \leq k-1 \\ a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)} & \text{for } i \geq k, j \geq k \end{cases}$$

$$f_i^{(k)} = \begin{cases} f_i^{(k-1)} & \text{for } i \leq k-1 \\ f_i^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} f_{k-1}^{(k-1)} & \text{for } i \geq k \end{cases}$$

Performing this method recursively, with $k = 1, \dots, n$, yields a new system that can be solved with simple backsubstitution (described further on).

Remark. A more intuitive way to look at the previous algorithm is in the row perspective: to generate $A^{(k)}$ from $A^{(k-1)}$, for each row $i \geq k$, simply subtract the $(k-1)$th row, scaled such that the $(k-1)$th column is eliminated for those rows. The same corresponding operations are performed to generate $\mathbf{f}^{(k)}$.

Theorem 1. Let the matrix A be such that the Gaussian elimination procedure defined in the Algorithm for Gaussian elimination yields non-zero diagonal elements $a_{kk}^{(k)}, k = 1, \dots, n$. Then A is non-singular and in fact,

$$\det A = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)}.$$

The final matrix $A^{(n)} =: U$ is upper triangular and A has the factorization

$$LU = A$$

where $L \coloneqq (m_{ik})$ is lower triangular with the elements

$$m_{ik} = \begin{cases} 0 & \text{for } i < k \\ 1 & \text{for } i = k \\ a_{ik}^{k} / a_{kk}^{(k)} & \text{for } i > k. \end{cases}$$

and the final vector $\mathbf{f}^{(n)} =: \mathbf{g}$ is

$$\mathbf{g} = L^{-1}\mathbf{f}$$

Algorithm (Backsubstitution)**.** If we define $(u_{ij}) \coloneqq U$ where U is an upper triangular matrix with its diagonal elements as non-zero, then we can solve the equation $U\mathbf{x} = \mathbf{g}$ with the equations

$$x_n = \frac{1}{u_{nn}} g_n$$
$$x_i = \frac{1}{u_{ii}} \left(g_i - \sum_{j=i+1}^{n} u_{ij} x_j \right), \ i = n-1, \dots, 1.$$

Theorem 2. Let the matrix A have rank r. Then we can find a sequence of distinct row and column indices $(i_1, j_1), \dots, (i_r, j_r)$ such that the corresponding pivot elements in $A^{(1)}, \dots, A^{(r)}$ are non-zero and $a_{ij}^{(r)} = 0$ if $i \neq i_1, \dots, i_r$. Let us define the permutation matrices, whose columns are unit vectors:

$$P \coloneqq \left(\mathbf{e}^{(i_1)}, \mathbf{e}^{(i_2)}, \dots, \mathbf{e}^{(i_r)}, \dots, \mathbf{e}^{(i_n)} \right)$$
$$Q \coloneqq \left(\mathbf{e}^{(j_1)}, \mathbf{e}^{(j_2)}, \dots, \mathbf{e}^{(j_r)}, \dots, \mathbf{e}^{(j_n)} \right)$$

where i_k, j_k, for $1 \leq k \leq r$, are the indices of the pivots and the sets $\{i_k\}$ and $\{j_k\}$ are permutations of $1, 2, \dots n$. Then the system

$$B\mathbf{y} = \mathbf{g}$$

where

$$B \coloneqq P^T A Q, \qquad \mathbf{y} \coloneqq Q^T \mathbf{x} \qquad \mathbf{g} \coloneqq P^T \mathbf{f}$$

can be solved Gaussian elimination with the natural order of pivots $(1,1), (2,2), \dots, (r,r)$.

Definition 2. Partial pivoting is the swapping of rows in order to place the largest elements (by absolute value) along the diagonal. Complete pivoting or maximal pivoting is the swapping of both rows and columns for the same purpose.

2.1.1 Operational Counts

Definition 3. The arithmetic complexity of an algorithm is the total number of additions, subtractions, multiplications and divisions that occur while performing it. The multiplicative complexity of algorithm is the total number of multiplications and divisions that occur while performing it.

Remark. To solve the m systems

$$A\mathbf{x} = \mathbf{f}(j), \qquad j = 1, \dots, m$$

with arbitrary $\mathbf{f}(j)$ it is *always* more efficient to solve them by Gaussian elimination than to calculate A^{-1}.

Formula.

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

Lemma 1. Solving the m systems

$$A\mathbf{x} = \mathbf{f}(j), \qquad j = 1, \dots, m$$

has a *multiplicative* complexity of

$$\frac{n^3}{3} + mn^2 - \frac{n}{3} \quad \text{ops.}$$

Lemma 2. There is a required multiplicative complexity of

$$n^3 \quad \text{ops.}$$

to computing A^{-1}.

2.1.2 A Priori Error Estimates; Condition Number

Definition 4. The condition number μ is defined as

$$\mu = \mu(A) \coloneqq \left\|A^{-1}\right\| \cdot \|A\|$$

Theorem 3. Consider solving the perturbed equation

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{f} + \delta\mathbf{f}$$

where $A\mathbf{x} = \mathbf{f}$, A is non-singular, and δA is so small that

$$\|\delta A\| < 1/\left\|A^{-1}\right\|.$$

If $\mathbf{x}$ and $\delta\mathbf{x}$ satisfy the perturbed and unperturbed equations, then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left[\frac{\mu}{1 - \mu\|\delta A\|/\|A\|}\right]\left(\frac{\|\delta\mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\delta A\|}{\|A\|}\right)$$

Remark. The previous theorem basically states that as long as the bracketed term is not too large, small relative changes in $\mathbf{f}$ and A only yield small relative changes in $\mathbf{x}$.

Lemma 3. If A is non-singular and t sufficiently large, then the Gaussian elimination method, with maximal pivots and floating-point arithmetic (with t-digit mantissas), yields multipliers s_{ij} with $|s_{ij}| \leq 1$ and pivots $b_{jj}^{(j)} \neq 0$, where $b_{ij}^{(k)}$ and s_{ij} are defined in the book. (Sorry! Wouldn't fit in the template.)

2.1.3 A Posteriori Error Estimates

Theorem 4. Let C be a computed or approximate inverse of the matrix A. If we define the error in the inverse as

$$F \coloneqq C - A^{-1}$$

and the residual matrix as

$$R \coloneqq AC - I,$$

then, given $\|R\| < 1$, all of the following statements are true:

(a) A and C are non-singular

(b) $\left\|A^{-1}\right\| \leq \|C\|/(1 - \|R\|)$

(c) $\|F\| \leq \|C\| \cdot \|R\|/(1 - \|R\|)$

Corollary. Under the same conditions as the previous theorem,

(d) $\left\|C^{-1}\right\| \leq \|A\|/(1 - \|R\|)$

(e) $\left\|A - C^{-1}\right\| \leq \|A\| \cdot \|R\|(1 - \|R\|)$

3 Iterative Solutions of Non-Linear Equations

Definition 5. A fixed point of a function g is a value α such that

$$\alpha = g(\alpha)$$

Definition 6. A mapping $g : D \to D$ is contracting if

$$\|g(x) - g(y)\| \leq M\|x - y\|$$

for all $x, y \in D$ where $M < 1$ provided D is closed and bounded (e.g. an interval on $\mathbb{R}$) or D is the entire domain and range of g.

3.1 Functional Iteration for a Single Equation

Theorem 1. If g satisfies the Lipschitz condition

$$|g(x) - g(y)| \leq \lambda|x - y|$$

for all x, y in the interval $I \coloneqq [x_0 - \rho, x_0 + \rho]$ such that $0 \leq \lambda < 1$ and initial guess x_0 be such that

$$|x_0 - g(x_0)| \leq (1 - \lambda)\rho, \qquad (*)$$

then all of the following are true:

(i) all iterates x_ν defined by $x_{\nu+1} \coloneqq g(x_\nu)$, lie within the interval I

(ii) (Existence) the iterates converge to some point

$$\lim_{\nu\to\infty} x_\nu = \alpha$$

which is a root of $x - g(x) = 0$.

(iii) (Uniqueness) α is the only root in I.

Corollary. If $|g'(x)| \leq \lambda < 1$ for $x \in I$ and $(*)$ is still satisfied, then all of the conclusions of the previous theorem still hold.

3.1.1 Error Propagation

Theorem 2. If $x = g(x)$ has a root at $x = \alpha$ and in the interval $I \coloneqq (\alpha - \rho, \alpha + \rho)$ $g(x)$ satisfies

$$|g(x) - g(\alpha)| \leq \lambda|x - \alpha| \qquad (**)$$

with $\lambda < 1$, then for any x_0 in the interval I, all

(i) all iterates x_ν defined by $x_{\nu+1} \coloneqq g(x_\nu)$, lie within the interval I

(ii) the iterates x_ν converge to α

(iii) α is the only root in I.

Corollary. If $|g'(x)| \leq \lambda < 1$ for $x \in I$ and $(**)$ is still satisfied, then all of the conclusions of the previous theorem still hold.

Theorem 3. Let $x = g(x)$ satisfy the conditions of the previous theorem and let $X_{\nu+1} \coloneqq g(X_\nu) + \delta_\nu$ (where δ_ν is an error term small enough that $\delta_\nu < \delta$ for all ν for some fixed δ). Let X_0 be any point in the interval

$$|\alpha - x| \leq \rho_0$$

where

$$0 < \rho_0 \leq \rho - \frac{\delta}{1-\lambda}.$$

Then the iterates X_ν lie in the interval

$$|\alpha - X_\nu| \leq \rho$$

and

$$|\alpha - X_k| \leq \frac{\delta}{1-\lambda} + \lambda^k\left(\rho_0 - \frac{\delta}{1-\lambda}\right)$$

where $\lambda^k \to 0$ as $\to \infty$.

3.2 Second and Higher Order Iteration Methods

Definition 1. Consider the Taylor series expansion of a function g at a fixed point α:

$$g(x) = \alpha + g'(\alpha)(x-\alpha) + \frac{g''(\alpha)}{2}(x-\alpha)^2 + \cdots + \frac{g^{(k)}(\alpha)}{k!}(x-\alpha)^k + \cdots$$

If k is the lowest positive integer such that $g^{(k)}(\alpha) \neq 0$, then the function has an order of convergence of k. When $k = 1$ or $k = 2$, the convergence is "linear" and "quadratic," respectively.

Algorithm (Chord Method)**.** Suppose we're trying to find a root of $f(x)$ which we will denote α. The iterative scheme

$$x_{\nu+1} = x_\nu - mf(x_\nu)$$

will, assuming m is chosen such that $0 < mf'(\alpha) < 2$, converge to the root linearly.

Algorithm (Newton's Method)**.** Suppose we're trying to find a root of $f(x)$ which we will denote α. The iterative scheme

$$x_{\nu+1} = x_\nu - \frac{f(x_\nu)}{f'(x_\nu)}$$

will converge to the root quadratically, assuming $f'(\alpha) \neq 0$ or that the root $x = \alpha$ has a multiplicity of one.