

A SAMPLING OF NUMERICAL TECHNIQUES

FREDDY BOULTON

ABSTRACT. This paper is meant to be an introduction to the study of numerical analysis. We explore a sampling of numerical techniques used to address several elementary, yet fundamental queries, such as finding roots to continuous real-valued functions and solving linear systems of the form $A\mathbf{x} = \mathbf{b}$. We approach these problems from a numerical perspective, describing the algorithms we use.

CONTENTS

1. Introduction	1
2. Root-Finding for Continuous Real-Valued Functions	2
2.1. The Contraction Mapping Theorem and Iterations	2
2.2. Finding Roots: Iterations through the Bisection Method	3
3. Matrix Notation	4
4. Finding LU (Solving $A\mathbf{x} = \mathbf{b}$ when x exists)	4
4.1. An Inefficient Approach	5
4.2. A More Efficient Approach	6
4.3. When LU Factorization Fails	9
5. Finding QR (Solving $A\mathbf{x} = \mathbf{b}$ when \mathbf{x} exists)	9
5.1. Gram-Schmidt Algorithm	10
6. Norms and Conditions Numbers	13
6.1. Norms	13
6.2. Condition Numbers	16
7. Least Squares: where solutions to $A\mathbf{x} = \mathbf{b}$ might not exist	17
7.1. Geometric Interpretation	18
7.2. Least Squares through QR Factorization	18
8. Acknowledgements	18
References	18

1. INTRODUCTION

Numerical analysis is concerned with the construction of approximate solutions to problems in scientific applications. The discipline began with babylonian scholars approximating $\sqrt{2}$ while constructing right triangles with legs of unit length. The scholars further used these approximations to construct a tangible measurement guide with countless applications to architecture and carpentry. In this study, the scholars saw first-hand the interplay between theory and practice. This expository paper seeks to analyze a sampling of techniques used by numerical analysts to solve various classical, mathematical questions. In studying these problems, we touch upon fundamental topics in other areas of mathematics such as Linear Algebra, Functional Analysis, and Real Analysis. We have adopted the spirit of the first numerical analysts by exploiting the symbiotic relationship between theory and practice in order to attain a deeper understanding of the material covered. Although numerical analysis is often studied through the lens of computer science, we are primarily concerned with the mathematical interpretation of the numerical algorithms. We still address the driving questions of numerical analysis, such as “how good is the approximate solution?” and “is the algorithm computationally feasible?”

This paper is organized as follows: In Section 2, we consider techniques for finding and approximating

the roots of real-valued continuous functions. In Section 3, we offer a brief overview of matrix notation. Sections 4 and 5 deal with methods to solving systems of linear equations, when the systems have explicit solutions. We conclude with a discussion of the theory of matrix norms and condition numbers in Section 6, as a means of introducing the method of Least Squares in Section 7, a method for analyzing solutions to systems of linear equations which may not have solutions.

2. ROOT-FINDING FOR CONTINUOUS REAL-VALUED FUNCTIONS

We begin by identifying roots of real valued functions. In the case of linear or quadratic polynomials, it is easy to find an explicit formula for the roots. However, significant obstacles arise once the degree of the polynomial is greater than or equal to five, as no such “quadratic formula equivalent ” exists (as proved by Abel in 1824 [2]). Consequently, we cannot expect a formula for arbitrary functions and as such we seek some type of algorithm to find the roots. We will introduce an iterative process which yields function values sufficiently close to the desired root. We first establish the Contraction Mapping Theorem, which is the necessary analytical tool to ensure that our iterative processes converge.

2.1. The Contraction Mapping Theorem and Iterations. We begin our discussion with the notion of a fixed point and Brouwer’s Fixed Point Theorem.

Definition 2.1 (Fixed Point). Let $f : [a, b] \rightarrow [a, b]$ be a continuous function defined on $[a, b]$. We say that $\zeta \in [a, b]$ is a fixed point of the function f if $f(\zeta) = \zeta$.

Given this definition, it is natural to wonder under which hypotheses does a function f have a fixed point. We address this question in Brouwer’s Fixed Point Theorem.

Theorem 2.2 (Brouwer’s Fixed Point Theorem). Let $f : [a, b] \rightarrow [a, b]$ for some $[a, b] \subset \mathbb{R}$. Then there exists a fixed point $\zeta \in [a, b]$ such that $f(\zeta) = \zeta$.

Proof. Without loss of generality, we may assume that $f(a) \neq a$ and $f(b) \neq b$; otherwise, a or b would be fixed points. We now define the function $g : [a, b] \rightarrow \mathbb{R}$ such that $g(x) = x - f(x)$. It follows that $g(a) = a - f(a) < 0$ since $f(a) \in [a, b]$ and similarly, $g(b) = b - f(b) > 0$. Since the sum of two continuous functions is a continuous function, then by the Intermediate Value Theorem, it follows that there exists a $\zeta \in (a, b)$ such that $g(\zeta) = 0$. Thus, $\zeta - f(\zeta) = 0$, implying that $f(\zeta) = \zeta$. Thus f has a fixed point ζ . \square

Definition 2.3 (Simple Iteration). Let g be a real valued function continuous on $[a, b]$ such that $g(x) \in [a, b]$ for all $x \in [a, b]$. We call the recursive sequence defined by

$$x_{k+1} = g(x_k) \text{ for } k \in \mathbb{N}$$

a simple iteration of g .

Fixed points are a crucial component of our goal to find the real-roots of functions because if f is real-valued and continuous, and the iteration above converges, then it must converge to a fixed point. Suppose that $\lim_{x \rightarrow \infty} (g_k) = \zeta$. Then we have that

$$\zeta = \lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} g(x_k) = g(\lim_{k \rightarrow \infty} x_k) = g(\zeta)$$

because g is continuous and limits pass through continuous functions. We are finding roots of a function f by finding fixed points of the recursive iterative function g .

With this in mind, we must find sufficient conditions for when the simple iteration converges. As it turns out, contractions provide an adequate condition:

Definition 2.4 (Contraction). Let f be a real valued function, continuous on a closed interval $[a, b] \subset \mathbb{R}$. We call f a contraction on $[a, b]$ if there exists an $L \in \mathbb{R}$, with $0 < L < 1$, such that

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

for all $x, y \in [a, b]$.

We now demonstrate the role of contractions and iterations in our pursuit of finding roots:

Theorem 2.5 (Contraction Mapping Theorem). *Let $f : [a, b] \rightarrow [a, b]$ with $[a, b] \subset \mathbb{R}$. Then f has a unique fixed point ζ in $[a, b]$. Moreover, the sequence (x_k) as defined in Definition 2.3 converges to ζ as $k \rightarrow \infty$ for any starting value $x_0 \in [a, b]$.*

Proof. By Brouwer's fixed point theorem, a fixed point ζ exists. To show uniqueness of ζ , suppose that there exists another fixed point $\zeta' \in [a, b]$ such that $\zeta' \neq \zeta$. Since f is a contraction, then

$$|f(\zeta) - f(\zeta')| = |\zeta - \zeta'| \leq L|\zeta - \zeta'|.$$

Since $|\zeta - \zeta'| > 0$ by assumption, then $\frac{|\zeta - \zeta'|}{|\zeta - \zeta'|} \leq L$. However, this implies that $1 \leq L$, which contradicts the definition of a contraction. Thus, $\zeta = \zeta'$. We will now show that the iteration defined in Definition 2.3 converges to ζ for any starting point $x_0 \in [a, b]$. We claim that $|x_k - \zeta| \leq L^k|x_0 - \zeta|$ for all $k \in \mathbb{N}$.

First, consider the case where $k = 1$, notice that $|x_1 - \zeta| = |f(x_0) - f(\zeta)| \leq L|x_0 - \zeta|$ since f is a contraction. Now suppose that for any $k \in \mathbb{N}$, the statement holds. We will now prove the statement for the case of x_{k+1} . By the inductive hypothesis, $|x_k - \zeta| \leq L^k|x_0 - \zeta|$. Thus,

$$|x_{k+1} - \zeta| = |f(x_k) - f(\zeta)| \leq L|x_k - \zeta| \leq L^{k+1}|x_0 - \zeta|.$$

This shows that the statement holds for all $k \in \mathbb{N}$.

Since $0 < L < 1$, then it follows that $\lim_{k \rightarrow \infty} L^k|x_0 - \zeta| = 0$. Since $0 \leq |x_k - \zeta| \leq L^k|x_0 - \zeta|$, we have that $\lim_{k \rightarrow \infty} |x_k - \zeta| = 0$. Thus, $\lim_{k \rightarrow \infty} x_k = \zeta$. \square

In order to illustrate an application of the Contraction Mapping Theorem, we consider the function $f : [1, 2] \rightarrow [1, 2]$ defined by

$$f(x) = e^x - 2x - 1.$$

We wish to find a root of f in this closed interval, which involves finding a fixed point of the function $x - f(x)$. Setting $f(x) = 0$, we see that $x = \ln(2x + 1)$. This motivates the choice of the function g as follows:

$$g(x) = \ln(2x + 1)$$

for the simple iteration $x_{k+1} = g(x_k)$ for $k \in \mathbb{N}$. Since g is continuous and differentiable on $[1, 2]$ for any $x, y \in [1, 2]$, the Mean Value Theorem yields that for all $x, y \in [1, 2]$, there exists a there exists some $p \in (x, y)$ such that

$$g(x) - g(y) = g'(p)(x - y),$$

which implies that

$$|g(x) - g(y)| \leq |g'(p)||x - y|.$$

By the chain rule, we have that $g'(x) = \frac{2}{2x+1}$ and thus $|g'| < 1$, which implies that g is a contraction.

By Theorem 2.5, g will converge to its fixed point $\zeta \in [1, 2]$ for any starting point $x_0 \in [1, 2]$. After a couple of iterations of the function, we see that $\zeta \approx 1.26$.

2.2. Finding Roots: Iterations through the Bisection Method. Suppose that f is a real valued and continuous function on a closed subset $[a, b]$ such that $f(a)$ and $f(b)$ are opposite signs. Thus, since f is continuous then by the Intermediate Value Theorem, there exists a $\zeta \in (a, b)$ such that $f(\zeta) = 0$. If we can identify a small enough interval around ζ then this will count as a good enough approximation of the root. This is precisely the goal of the Bisection Method. We shrink the interval in a systematic way based on "halving":

Let $c_k = \frac{1}{2}(a_k + b_k)$ be the average of the k^{th} interval of the iteration, where a_k, b_k have opposite signs, thus guaranteeing that $\zeta \in (a_k, b_k)$. Define (a_{k+1}, b_{k+1}) by:

$$(a_{k+1}, b_{k+1}) = \begin{cases} (a_k, c_k), & \text{if } f(c_k)f(b_k) > 0 \\ (c_k, b_k), & \text{if } f(c_k)f(b_k) < 0 \end{cases}$$

If $f(c_k)f(b_k) > 0$, then $f(c_k)$ must be the same sign as $f(b_k)$, thus the root must lie in (a_k, c_k) (as long as $c_k \neq 0$) by the intermediate value theorem since a_k and b_k have opposite signs by assumption. The

same argument symmetrically applies to the other half of the definition. Notice that the bisection method completely relies on the Intermediate Value Theorem, which requires that f be continuous. Moreover, it is clear that the bisection method gives a good approximation as $k \rightarrow \infty$ as each interval is successively halved.

The bisection method is not as sophisticated or elegant as the contraction method but it only requires that the function be continuous and for us to provide an interval, i.e. a guess, as to where the root might be. Thus, we avoid verifying that the function f is a contraction, which can be challenging, time-consuming, and even impossible depending on the function.

3. MATRIX NOTATION

In this section, we briefly introduce the basics of matrix notation, and explain an efficient strategy to solve systems of the form $A\mathbf{x} = \mathbf{b}$. We begin our discussion of matrix notation by defining a Matrix space:

Definition 3.1 (Matrix Space). The set of real-valued matrices with m rows and n columns is a vector space over the field of real numbers \mathbb{R} and is denoted by $\mathbb{R}^{m \times n}$.

Remark 3.2. For a definition of a vector space, consult [4].

We begin with the definition of matrix-vector multiplication on the left-hand side:

Definition 3.3. Let $x \in \mathbb{R}^n$ and let $A \in \mathbb{R}^{m \times n}$. The matrix vector product $\mathbf{b} = A\mathbf{x}$ is defined as:

$$b_i = \sum_{j=1}^n a_{ij}x_j, i = 1, 2, \dots, m.$$

where b_i denotes the i^{th} entry of \mathbf{b} , a_{ij} denotes the entry in the i^{th} row and j^{th} column of A and x_j denotes the j^{th} entry of \mathbf{x} .

We now walk through the notation for matrix multiplication:

Definition 3.4 (Matrix Multiplication). Let $A \in \mathbb{R}^{l \times m}$ and let $B \in \mathbb{R}^{m \times n}$. The entries of the resulting matrix $C \in \mathbb{R}^{l \times n}$ are defined as follows:

$$c_{ij} = \sum_{k=1}^m a_{ik}b_{kj}.$$

where b_{ij} is the i, j entry of B , a_{ik} is the entry in the i^{th} row and k^{th} entry of A , and similarly for b_{jk} .

We will now explain a general strategy to solve linear systems. Suppose we have a matrix $A \in \mathbb{R}^{3 \times 3}$ and a vector $\mathbf{b} \in \mathbb{R}^3$ and that we are trying to find a vector $\mathbf{x} \in \mathbb{R}^3$ such that $A\mathbf{x} = \mathbf{b}$. Typically, this would entail a system of the form

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & t \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix},$$

which is not a trivial exercise. However, if the entries of A satisfied $a_{ij} = 0$ for all $i < j$, then the system would be of the form

$$\begin{bmatrix} a & b & c \\ 0 & e & f \\ 0 & 0 & t \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 5 \\ 10 \\ 15 \end{bmatrix},$$

which can be solved easily by working from the 3^{rd} row up: $z = \frac{15}{t}$, $y = \frac{10 - \frac{15}{t}f}{e}$ and so forth. This motivates trying to find a way to reduce any matrix $A \in \mathbb{R}^{n \times n}$ to an upper-triangular form so that the problem of $A\mathbf{x} = \mathbf{b}$ can be solved efficiently.

4. FINDING LU (SOLVING $A\mathbf{x} = \mathbf{b}$ WHEN x EXISTS)

We now examine one of the fundamental problems of Linear Algebra: given a matrix $A \in \mathbb{R}^{n \times n}$ and a vector \mathbf{b} in \mathbb{R}^n , we want to find a vector $\mathbf{x} \in \mathbb{R}^n$ such that $A\mathbf{x} = \mathbf{b}$. There are many ways to approach the problem from the numerical perspective. We will show several approaches in order to compare and contrast their strengths and weaknesses.

4.1. An Inefficient Approach. One way of solving $A\mathbf{x} = \mathbf{b}$ is to find the inverse of A and compute $A^{-1}\mathbf{b}$. We will begin our analysis of this approach by introducing inverse matrices and the concept of the determinant.

Definition 4.1 (Inverse of a Matrix). Let $A \in \mathbb{R}^{n \times n}$. A matrix W is the inverse of a matrix A if

$$AW = WA = I_n$$

where

$$I_n = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{bmatrix},$$

and is referred to as the $n \times n$ *identity matrix*.

Remark 4.2. It is easy to show that inverse matrices are unique by using the associativity of matrix multiplication (consult reference [2]). Thus we denote the inverse of the matrix A by A^{-1} .

In order to explicitly calculate the inverse of a square matrix A we introduce the concept of the determinant.

Definition 4.3 (Determinant). Let $A \in \mathbb{R}^{n \times n}$, and let S_n be the set of all permutations of the elements of the set $\{1, 2, 3, \dots, n\}$. A permutation of the set $\{1, 2, 3, \dots, n\}$ is a bijection $\sigma : \{1, 2, 3, \dots, n\} \rightarrow \{1, 2, 3, \dots, n\}$. We define the sign of a permutation σ , denoted $\text{sign}(\sigma)$ to be $(-1)^{\text{inv}(\sigma)}$, where $\text{inv}(\sigma)$ denotes the number of inversions of σ , i.e. the number of exchanges of two adjacent elements of the set $\{1, 2, 3, \dots, n\}$. We define the determinant of A , denoted by $\det(A)$, as

$$\det(A) = \sum_{\sigma \in S} \text{sign}(\sigma) \prod_{i=1}^n a_{i, \sigma(i)},$$

where $a_{i, \sigma(i)}$ denotes the entry of the matrix A in the i^{th} row and the column i is mapped to in the permutation.

Example 1 (Determinant of a 2×2 matrix). We will compute the determinant of a matrix $A \in \mathbb{R}^{2 \times 2}$. Let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

Proof. There are 2 total permutations of the set $\{1, 2\}$:

- (1) $\sigma = \{1, 2\}$
- (2) $\sigma' = \{2, 1\}$

It is clear that σ has zero inversions since no adjacent elements were exchanged; its sign is therefore 1. Thus, $\text{sign}(\sigma) \prod_{i=1}^n a_{i, \sigma(i)} = a_{11}(a_{21}) = ad$. Similarly, σ' has one inversion, and thus its sign is -1 . It follows that $\text{sign}(\sigma') \prod_{i=1}^n a_{i, \sigma(i)} = -(a_{12})a_{21} = -bc$. It follows that $\det(A) = ad - bc$. \square

Remark 4.4. We can already see the immense computational cost of computing the determinant of a matrix. If we wanted to compute the determinant of a 50×50 matrix, we would have to compute the sum over $50!$ terms (the total number of permutations in a set of 50 elements) (see reference [3]).

We now define the cofactor of a matrix, which will be used to compute the inverse of a matrix.

Definition 4.5 (Cofactor). Let $A \in \mathbb{R}^{n \times n}$. The *cofactor of entry* a_{ij} , denoted as $\text{cof}(a_{ij})$, is defined as the determinant of the $(n-1) \times (n-1)$ matrix obtained by deleting the i^{th} row and the j^{th} column.

Example 2. Consider the 3×3 matrix

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix};$$

$\text{cof}(a_{11})$ is the determinant of the resultant matrix given by deleting the first row and column of the matrix.

Thus, we must compute the determinant of $\begin{bmatrix} e & f \\ h & i \end{bmatrix}$, which we know is equal to $ei - fh$.

We now introduce Cramer's Rule, which utilizes cofactors to compute the inverse of a square matrix A .

Theorem 4.6 (Cramer's Rule). *For a matrix $A \in \mathbb{R}^{n \times n}$ such that $\det(A) \neq 0$:*

$$A^{-1} = \frac{1}{\det(A)} \begin{bmatrix} A_{11} & A_{21} & \dots & A_{n1} \\ A_{12} & A_{22} & \dots & A_{n2} \\ \dots & \dots & \dots & \dots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}$$

where A_{ij} denotes the $\text{cof}(a_{ij})$

Remark 4.7. The proof of Cramer's Rule depends on an equivalent formulation of the determinant, namely $\det(A) = \sum_{k=1}^n a_{ik} A_{ki}$, and the fact that

$$\sum_{k=1}^n a_{jk} A_{ik} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

However, the proof of these properties will not be provided here, as it deters from the main focus of this paper. For more information see [1].

Using Cramer's Rule, we can directly compute A^{-1} and solve the system $A\mathbf{x} = \mathbf{b}$. However, from the formulation of Cramer's Rule, it is evident that the inverse of a square matrix only exists if $\det(A) \neq 0$. This poses an obstacle to the analysis: given a large square matrix, say 30×30 , we would have to compute a sum of $30!$ terms in order to even begin to calculate A^{-1} . To make matters worse, even if we were somehow guaranteed that $\det(A) \neq 0$, we'd have to calculate 90 determinants of 29×29 matrices in order to compute A^{-1} , which is clearly undesirable. This suggests that from the numerical perspective, we seek a more efficient algorithm.

4.2. A More Efficient Approach. From the computational point of view, a better way to solve systems of linear equations is to describe the matrix A in terms of matrices which are easier to invert. In particular, we factor our matrix A into the product of a unit lower triangular matrix and an upper triangular matrix. This process is called *LU* factorization and it is done through row operations and Gaussian Elimination. We begin our analysis by introducing Elementary Matrices, which perform desired row operations for Gaussian Elimination.

Definition 4.8 (Elementary Matrix). Multiplication by Elementary Matrices perform linear operations on the rows of A . If we want to add a scalar multiple, $\mu_{(rs)}$, of row s to row r , we perform the following operation:

$$A \rightarrow A(I_n + \mu_{(rs)}E^{(rs)}),$$

where

$$E_{(ij)}^{(rs)} = \begin{cases} 1, & \text{if } i = r, j = s \\ 0, & \text{otherwise} \end{cases},$$

and $\mu_{(rs)} \in \mathbb{R}$.

We next provide an example.

Example 3. *Suppose that we have the matrix*

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 8 \end{bmatrix},$$

and that we want to add 3 times the first row to the second one. It is clear that our new matrix A' should read $A' = \begin{bmatrix} 1 & 2 \\ 6 & 14 \end{bmatrix}$. However, using our definition of the elimination matrix E , we see that

$$(I_n + \mu_{(12)}E^{(21)}) = \begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix}.$$

Therefore by computing $(I_n + \mu_{(12)}E^{(21)})A$ we have that

$$\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 6 & 14 \end{bmatrix}.$$

Notice that the elementary Matrix $(I_n + \mu_{(12)}E^{(21)})$ has entries equal to 0 above its diagonal. We formalize this concept in the following definition.

Definition 4.9 (Lower and Upper Triangular Matrix). A matrix $L \in \mathbb{R}^{n \times n}$, where $n \geq 2$ is called lower triangular if $l_{ij} = 0$ for all $i < j$. In other words, all entries above the main diagonal are 0. L is **unit lower triangular** if it is lower triangular and $l_{ii} = 1$ for all $i \in \{1, 2, \dots, n\}$; i.e. all entries along the main diagonal are 1. A matrix $U \in \mathbb{R}^{n \times n}$ where $n \geq 2$ is called upper triangular if $u_{ij} = 0$ for all $j < i$.

Thus, if $r < s$, then matrix $(I_n + \mu_{(rs)}E^{(rs)})$ is lower triangular by definition 4.9. We will now show that the product of two lower triangular matrices is lower triangular. In order to do this, we will use the following lemma:

Lemma 4.10. For the matrix $E^{(rs)}$, as defined above, where $r \neq s$, then

$$E^{(rs)}E^{(rs)} = 0,$$

where 0 is the $n \times n$ with all 0 entries.

Proof. The proof is omitted. For more information consult [1]. □

Theorem 4.11. The inverse of the $n \times n$ elementary matrix $(I_n + \mu_{(rs)}E^{(rs)})$ is given by $(I_n - \mu_{(rs)}E^{(rs)})$.

Proof. By the distributive property of matrix multiplication, we have

$$(I_n + \mu_{(rs)}E^{(rs)})(I_n - \mu_{(rs)}E^{(rs)}) = I_n - \mu_{(rs)}E^{(rs)} + \mu_{(rs)}E^{(rs)} + \mu_{(rs)}^2E^{(rs)}E^{(rs)}.$$

By the above lemma, $E^{(rs)}E^{(rs)} = 0$. Thus,

$$(I_n + \mu_{(rs)}E^{(rs)})(I_n - \mu_{(rs)}E^{(rs)}) = I + 0 + 0 = I.$$

Since multiplicative inverses are unique, it follows that $(I + \mu_{(rs)}E^{(rs)})^{-1} = (I - \mu_{(rs)}E^{(rs)})$. □

The goal of LU Factorization will be to convert our original matrix A to an upper triangular matrix, as defined in definition 4.9. We will perform this factorization by multiplying A by a series of elementary matrices similar to the ones above. Let $(I + \mu_{(rs)}E^{(rs)}) = L^{(rs)}$, let $L^{-(rs)} = (I + \mu_{(rs)}E^{(rs)})^{-1}$ and let U be an upper triangular matrix. Thus, we have:

$$L^{(n,s)}L^{(n-1,s-1)} \dots L^{(1,2)}A = U$$

By multiplying by the inverses of each unit lower triangular matrix, we have that

$$A = L^{-(1,2)} \dots L^{-(n,s)}U$$

Since the product of all the L matrices is still a unit lower triangular matrix, we have that

$$A = LU,$$

as desired.

Example 4. We perform the LU factorization of the matrix

$$B = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & 5 & -4 \end{bmatrix}.$$

We begin by making the first column of B a vector of the form $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$. Thus, the first step is to multiply by

$$L^1 = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ This yields,}$$

$$L^1 B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ -1 & 5 & -4 \end{bmatrix}.$$

Now we add the first row to the third by multiplying by $L^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$, and obtain that

$$L^2 L^1 B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 6 & -3 \end{bmatrix}.$$

We now move to the second column. We will add -3 times the second row to the third by multiplying by

$$L^3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -3 & 1 \end{bmatrix}. \text{ Thus,}$$

$$L^3 L^2 L^1 B = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix}$$

Isolating B yields,

$$\begin{aligned} B &= L^1 L^2 L^3 \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix}. \end{aligned}$$

$$\text{Thus, } L = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix} \text{ and } U = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix}.$$

4.2.1. *Using LU Factorization to solve Equations.* Once we obtain the LU factorization for our matrix A , solving the system $A\mathbf{x} = \mathbf{b}$ is equivalent to solving

$$LU\mathbf{x} = \mathbf{b}.$$

It may be tempting to multiply by L^{-1} and obtain the system

$$U\mathbf{x} = \mathbf{c},$$

where $\mathbf{c} = L^{-1}\mathbf{b}$ which can be solved in reverse order. However, instead of computing L and L^{-1} explicitly, we invert each elementary matrix at each step. This is how a computer would solve the system of Linear Equations. We demonstrate this approach, and what we mean by “invert at each step” with the following example:

Example 5 (Solving a System of Linear Equations). *Suppose we have the system of linear equations:*

$$\begin{aligned} x + y + z &= 6 \\ 2x + 4y + 2z &= 16 \\ -x + 5y - 4z &= -3 \end{aligned}$$

We can express the system as a 3×3 matrix multiplied by a column vector as follows:

$$\begin{bmatrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ -1 & 5 & -4 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 16 \\ -3 \end{bmatrix}$$

Recalling that this is the matrix B of Example 4, we can multiply by the elementary matrices L^1, L^2, L^3 (defined in Example 4), on both sides to obtain:

$$L^3 L^2 L^1 B = L^3 L^2 L^1 \begin{bmatrix} 6 \\ 16 \\ -3 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 6 \\ 4 \\ -9 \end{bmatrix}$$

Thus, solving in reverse order, we have that $z = 3$, $y = 2$, and $x = 1$.

Note that in Example 4, the matrix $L^3 L^2 L^1$ is the inverse of the matrix $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 3 & 1 \end{bmatrix}$, which is the lower triangular matrix in the LU factorization of B . Thus, $L^3 L^2 L^1 = L^{-1}$ in the example above, which allows us to avoid explicitly inverting L .

4.3. When LU Factorization Fails. The process of elimination above implicitly assumes that each entry along the diagonal of A is nonzero. Notice if $b_{22} = 0$ in the matrix above, then we would not be able to multiply the second row by a scalar in order to eliminate the term in the b_{23} position above. In order to solve this problem, we introduce the concept of the permutation matrix:

Definition 4.12. A permutation matrix $P \in \mathbb{R}^{n \times n}$ is a matrix in whose entries are either 1 or 0, such that every row and column contains precisely one nonzero element.

Permutation matrices exchange rows. Thus, if our matrix B above were changed to $\begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & 2 \\ -1 & 5 & -4 \end{bmatrix}$, then

we could multiply by the Permutation Matrix $P = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ in order to exchange the third and second

rows and obtain the matrix $\begin{bmatrix} 1 & 1 & 1 \\ -1 & 5 & -4 \\ 2 & 0 & 2 \end{bmatrix}$, which can be factored into its LU form.

In general, if we have a system $A\mathbf{x} = \mathbf{b}$ we can multiply by a permutation matrix P and obtain that $PA\mathbf{x} = P\mathbf{b}$, where PA can be decomposed into LU . Thus, we have that $U\mathbf{x} = L^{-1}P\mathbf{b}$, which can be solved backwards as before.

5. FINDING QR (SOLVING $A\mathbf{x} = \mathbf{b}$ WHEN \mathbf{x} EXISTS)

Despite the seeming elementary nature of LU factorization, the LU approach is not always ideal, as we are transforming our matrix entry-by-entry. Another way to solve $A\mathbf{x} = \mathbf{b}$, which bypasses the fault of LU , is to compute $A = QR$, where Q is an orthogonal matrix and R is an upper triangular matrix. In order to compute the QR factorization, we first need to consider how we can “convert” A into an orthogonal matrix. The algorithm for this conversion is called Gram-Schmidt Orthogonalization and it is a standard procedure in Linear Algebra used to create a collection of orthogonal vectors.

5.1. Gram-Schmidt Algorithm. Our discussion of the Gram-Schmidt Algorithm begins with an introduction to the notions of orthogonal vectors and orthogonal matrices.

Definition 5.1 (Orthogonal and Orthonormal Vectors). Two vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ are orthogonal if $\langle u, v \rangle = 0$, where $\langle u, v \rangle = \sum_{i=1}^n u_i v_i$. We say that $\langle u, v \rangle$ as the inner product of u and v . Similarly, we say that a vector q is orthonormal if $\langle q, q \rangle = 1$. For a list of properties of the inner product, consult [1].

Definition 5.2 (Orthogonal Matrices). A matrix $Q \in \mathbb{R}^{n \times n}$ is orthogonal if its columns are orthogonal to each other, and the length of each column-vector is one. An alternative formulation, for Q , if q_i and q_j are columns of Q such that $i \neq j$, then $\langle q_j, q_i \rangle = 0$, whereas $\langle q_i, q_i \rangle = 1$.

The advantage of orthogonal matrices is that the inverses of orthogonal matrices are trivial to compute. In order to invert these matrices we must introduce the following concept:

Definition 5.3 (Matrix Transpose). Let $A \in \mathbb{R}^{n \times n}$. We define $A^T \in \mathbb{R}^{n \times n}$, called the transpose of A , to be the reflection of A along the main diagonal. In other words, if a_{ij}^T is an entry in A^T , then $a_{ij}^T = a_{ji}$.

Equipped with this definition, we now show the simple structure of Q^{-1} when Q is orthonormal.

Theorem 5.4 (Inverses of Orthogonal Matrices). Let $Q \in \mathbb{R}^{n \times n}$ be an orthogonal matrix. Let Q^T be the transpose of matrix Q . Then,

$$Q^T = Q^{-1}$$

Proof. We will first compute $Q^T Q$. Let $(qq^T)_{ii}$ denote any entry of the resulting product matrix along the main diagonal. By the definition of matrix multiplication, $(qq^T)_{ii} = \sum_{k=1}^n q_{ik} q_{ki}^T$; where q_{ik} denotes the entries along the i^{th} row of Q and q_{ki}^T denotes the entries along the i^{th} column of Q^T . However, by the definition of the transpose, the i^{th} row of Q is equal to the i^{th} column of Q^T . Thus,

$$\sum_{k=1}^n q_{ik} q_{ki}^T = \langle q_i^T, q_i^T \rangle = 1$$

by the definition of an orthogonal matrix.

Now let qq_{ij}^T , where $i \neq j$, denote any entry off the main diagonal. By the definition of matrix multiplication, $(qq)_{ij}^T = \sum_{k=1}^n q_{ik} q_{kj}^T = \langle q_i^T, q_j^T \rangle = 0$ since q_{ik} is equal to the i^{th} column of Q^T . Thus, it follows that $Q^T Q = I_n$. A symmetrical argument will show that $Q Q^T = I_n$, implying that the statement holds. \square

This shows that a factorization of A into a QR matrix product is extremely useful, since inverting the Q matrix requires no additional work. We next introduce the concept of linear independence and then describe the procedure for factoring A into QR .

Definition 5.5 (Linearly Independent Vectors). A collection of vectors v_1, \dots, v_n is said to be linearly independent if for any $\alpha_1, \dots, \alpha_n$ such that $\alpha_1 v_1 + \dots + \alpha_n v_n = 0$, then $\alpha_1 = \alpha_2 = \dots = \alpha_n = 0$.

Procedure (Gram-Schmidt Algorithm). Let v_1, v_2, \dots, v_n be a set of linearly independent vectors in \mathbb{R}^n . The following vectors:

$$\begin{aligned} q_1 &= \frac{v_1}{\|v_1\|} \\ q_2 &= \frac{v_2^*}{\|v_2^*\|}; \text{ where } v_2^* = v_2 - \langle v_2, q_1 \rangle q_1 \\ &\dots\dots\dots \\ q_n &= \frac{v_n^*}{\|v_n^*\|}; \text{ where } v_n^* = v_n - \langle v_n, q_{n-1} \rangle q_{n-1} - \dots - \langle v_n, q_1 \rangle q_1 \end{aligned}$$

produce an orthonormal set of vectors q_1, \dots, q_n which span the same space as v_1, \dots, v_n .

We will now prove that the set of vectors $\{q_1, \dots, q_n\}$ is orthonormal.

Proof. The proof is inductive. For the case of $n = 1$, it is clear that q_1 is orthogonal to all previous vectors and it is trivial to check that q_1 has length 1. To see this, note that

$$\langle q_1, q_1 \rangle = \frac{1}{\|v_1\|_2^2} \|v_1\|_2^2 = 1.$$

Now assume that vectors q_1, \dots, q_{n-1} are orthogonal. We will now show that q_n is orthogonal to q_1, \dots, q_{n-1} . Let $i \in \{1, \dots, n-1\}$. Note that by the bilinearity of the inner product, we have:

$$\langle q_i, q_n \rangle = \frac{1}{\|v_n^*\|} \langle q_i, v_n \rangle - \langle v_n, q_{n-1} \rangle \langle q_i, q_{n-1} \rangle - \dots - \langle v_n, q_1 \rangle \langle q_i, q_1 \rangle.$$

Since $\langle q_i, q_{n-1} \rangle = 0$ by the inductive step, then we have

$$\frac{1}{\|v_n^*\|} \langle q_i, q_n \rangle = \frac{1}{\|v_n^*\|} (\langle q_i, q_n \rangle - \langle q_i, q_n \rangle \langle q_i, q_i \rangle) = 0.$$

Thus, vector q_n is orthogonal to vectors q_1, \dots, q_{n-1} . The approach as in the base case can be used to show that it has length 1. Thus, vectors q_1, \dots, q_n are orthogonal. \square

In order to motivate the procedure of Gram-Schmidt Algorithm, we provide a geometric interpretation for each step. By taking the dot product of v_2 with the previous vector, we are essentially removing the projection of v_2 in the direction of q_1 . Thus, we are leaving only the component of vector v_2 which is orthogonal to q_1 . In general, at step i , we produce a vector that is orthogonal to vectors v_1 to v_{i-1} by taking away all components which are parallel to v_1, \dots, v_i .

If $A = [v_1 \ v_2 \ \dots \ v_n]$, we can treat the columns of the matrix as separate vectors and convert the columns to an orthonormal basis, however, notice that we are left with the following:

$$\begin{aligned} v_1 &= r_{11}q_1 \\ v_2 &= r_{12}q_1 + r_{22}q_2 \\ &\vdots \\ v_n &= r_{1n}q_1 + r_{2n}q_2 + \dots + r_{nn}q_n \end{aligned}$$

In light of this observation, we see that the columns of the original matrix A are linear combinations of the columns of the matrix Q . Therefore, we have that

$$A = QR = [q_1 \ q_2 \ \dots \ q_n] \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ 0 & r_{22} & \dots & r_{2n} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & r_{nn} \end{bmatrix}.$$

Example 6. Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 0 & 5 \\ 0 & 3 & 6 \end{bmatrix}.$$

According to the algorithm: $q_1 = \frac{1}{\sqrt{1}}v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$,

By definition,

$$\begin{aligned} v_2^* &= v_2 - (v_2 \cdot q_1)q_1 \\ &= \begin{bmatrix} 2 \\ 0 \\ 3 \end{bmatrix} - 2 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 0 \\ 3 \end{bmatrix}. \end{aligned}$$

Since $\|v_2^*\|_2 = \sqrt{0^2 + 0^2 + 3^2} = 3$, then $q_2 = \frac{v_2^*}{\|v_2^*\|} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

For q_3 , we define the vector v_3^* as:

$$\begin{aligned} v_3^* &= v_3 - (v_3 \cdot q_2)(q_2) - (v_3 \cdot q_1)(q_1) \\ &= \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix} - 6 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} - 4 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix}. \end{aligned}$$

Since $\|v_3^*\|_2 = 5$, then $q_3 = \frac{1}{5} \begin{bmatrix} 0 \\ 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. Thus, for our orthonormalized matrix we have that

$$Q = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

Moreover, notice that:

$$\begin{aligned} v_1 &= 1q_1 \\ v_2 &= 2q_1 + 3q_2 \\ v_3 &= 4q_1 + 6q_2 + 5q_3. \end{aligned}$$

It follows that our upper triangular matrix R is of the form:

$$R = \begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 6 \\ 0 & 0 & 5 \end{bmatrix}.$$

Thus, we have factored A as

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 6 \\ 0 & 0 & 5 \end{bmatrix}.$$

Once we compute $A = QR$, then we can set:

$$QR\mathbf{x} = \mathbf{b}.$$

Using that Q is orthogonal, we have that $R\mathbf{x} = \mathbf{Q}^T\mathbf{b}$. The system could then be solved backwards. As an example, consider the system of linear equations:

$$\begin{aligned} x + 2y + 4z &= 31 \\ 5z &= 25 \\ 3y + 6z &= 42, \end{aligned}$$

which is equivalent to solving:

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 0 & 5 \\ 0 & 3 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 31 \\ 25 \\ 42 \end{bmatrix}.$$

Using the QR factorization of A computed above, we have that

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 6 \\ 0 & 0 & 5 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 31 \\ 25 \\ 42 \end{bmatrix}.$$

Thus,

$$\begin{bmatrix} 1 & 2 & 4 \\ 0 & 3 & 6 \\ 0 & 0 & 5 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 31 \\ 25 \\ 42 \end{bmatrix} = \begin{bmatrix} 31 \\ 42 \\ 25 \end{bmatrix}.$$

By solving backwards, we have $z = 5, y = 4, x = 3$, or $\mathbf{x} = \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix}$.

6. NORMS AND CONDITIONS NUMBERS

6.1. Norms. Norms provide a rigorous way to quantify size and distance in vector spaces. Because of this, they provide a measure of how “close” an approximate solution is to the actual solution, such as in root-finding. Additionally, they are a suitable measure of the effects of rounding errors on solutions of systems of linear equations and fundamental to the study of numerical analysis. Moreover, they are the basis for the study of functional analysis.

Definition 6.1 (Norm). Let \mathcal{V} be a vector space over the field \mathbb{R} of real numbers. We call a nonnegative function $\|\cdot\| : \mathcal{V} \rightarrow \mathbb{R}$ a norm on \mathcal{V} if it satisfies the following:

- (1) $\|\mathbf{v}\| = 0$ if and only if $\mathbf{v} = \mathbf{0}$, for all $\mathbf{v} \in \mathcal{V}$.
- (2) $\|\lambda\mathbf{v}\| = |\lambda|\|\mathbf{v}\|$, for all $\lambda \in \mathbb{R}$ and $\mathbf{v} \in \mathcal{V}$.
- (3) $\|\mathbf{v} + \mathbf{u}\| \leq \|\mathbf{v}\| + \|\mathbf{u}\|$ for all $\mathbf{v}, \mathbf{u} \in \mathcal{V}$.

We now define a series of norms for vectors $\mathbf{v} \in \mathbb{R}^n$.

Definition 6.2 (1-norm). The vector 1-norm of a vector $\mathbf{v} \in \mathbb{R}^n$, denoted by $\|\mathbf{v}\|$ is defined as

$$\|\mathbf{v}\| = \sum_{i=1}^n |\mathbf{v}_i|.$$

Definition 6.3 (2-norm). The vector 2-norm of a vector $\mathbf{v} \in \mathbb{R}^n$, denoted by $\|\mathbf{v}\|_2$, is defined as:

$$\|\mathbf{v}\|_2 = \left[\sum_{i=1}^n |\mathbf{v}_i|^2 \right]^{\frac{1}{2}}.$$

Remark 6.4 (Relationship to Dot-Product). The vector 2-norm is closely related to the dot product of a vector. More specifically:

$$\|\mathbf{v}\|_2^2 = \sum_{i=1}^n |\mathbf{v}_i|^2 = \langle \mathbf{v}_i, \mathbf{v}_i \rangle.$$

Thus, the dot-product of a vector with it itself is its length squared. This property of the 2-norm will be useful in later parts of the paper.

Definition 6.5 (∞ -norm). The ∞ -norm (infinity norm) of a vector $\mathbf{v} \in \mathbb{R}^n$, denoted by $\|\mathbf{v}\|_\infty$ is defined as

$$\|\mathbf{v}\|_\infty = \max_{1 \leq i \leq n} |\mathbf{v}_i|.$$

We can generalize the concept of the 1-norm and 2-norm, for all $p \in [1, \infty)$:

Definition 6.6 (p -norm). Let $p \in \mathbb{R}$ such that $p \geq 1$. We define the p -norm of the vector $\mathbf{v} \in \mathbb{R}^n$, denoted as $\|\mathbf{v}\|_p$ by

$$\|\mathbf{v}\|_p = \left[\sum_{i=1}^n |\mathbf{v}_i|^p \right]^{\frac{1}{p}}.$$

Calling these objects “norms” may seem a bit premature, since we have yet to prove they satisfy the norm properties. For $p = 1$ or $p = \infty$, the proofs follow directly from the properties of absolute value and maximums. Moreover, proving axioms 1 and 2 is trivial for any $p \in (1, \infty)$. The main difficulty lies in proving property 3. Our discussion begins with Young’s inequality, which is a statement regarding convex functions.

Definition 6.7 (Convexity). We call a twice differentiable function f convex on $[a, b]$ if $\frac{d^2}{dx^2} f > 0$ for all $x \in [a, b]$. Equivalently, this implies that for all $\theta \in (0, 1)$ and for all $x, y \in [a, b]$, $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$ for all $x, y \in [a, b]$. As an example, note that the function $f(x) = e^x$ is convex on \mathbb{R} because $\frac{d^2}{dx^2} e^x = e^x > 0$ for all $x \in \mathbb{R}$.

Theorem 6.8 (Young's Inequality). Let $a, b \in \mathbb{R}$. Let $p, q \in \mathbb{R}$ such that $p, q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$. Then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

Proof. Let $\theta = \frac{1}{p}$. Thus, $1 - \theta = \frac{1}{q}$. Since $f(x) = e^x$ is a convex function on \mathbb{R} ,

$$ab = e^{\ln a + \ln b} = e^{\frac{1}{p} \ln a^p + \frac{1}{q} \ln b^q} \leq \frac{1}{p} e^{\ln a^p} + \frac{1}{q} e^{\ln b^q} = \frac{a^p}{p} + \frac{b^q}{q}.$$

□

Young's Inequality is the crucial ingredient needed to prove Holder's Inequality, which is a generalization of the Cauchy-Schwarz Inequality. The Cauchy-Schwarz Inequality states that for any vectors \mathbf{v}, \mathbf{u} , $|\sum_{i=1}^n v_i u_i| \leq \|\mathbf{v}\|_2 \|\mathbf{u}\|_2$. Notice that Cauchy-Schwarz is the special case where $p = q = \frac{1}{2}$ in the statement below:

Theorem 6.9 (Holder's Inequality). Let $p, q \in \mathbb{R}$ with $\frac{1}{p} + \frac{1}{q} = 1$. For any \mathbf{v} and $\mathbf{u} \in \mathbb{R}^n$, we have

$$\left| \sum_{i=1}^n v_i u_i \right| \leq \|\mathbf{v}\|_p \|\mathbf{u}\|_q$$

Proof. Without loss of generality, suppose that $\mathbf{v}, \mathbf{u} \neq \mathbf{0}$; otherwise, the inequality trivially holds. Now define the vectors $\hat{\mathbf{v}}$ where

$$\hat{v}_i = \frac{v_i}{\|\mathbf{v}\|_p}$$

for all $i \in [n]$ and similarly for $\hat{\mathbf{u}}$. By the triangle inequality for real numbers we have that

$$\left| \sum_{i=1}^n \hat{v}_i \hat{u}_i \right| \leq \sum_{i=1}^n |\hat{v}_i \hat{u}_i|.$$

By applying Young's Inequality to the right hand side, it follows that

$$\sum_{i=1}^n |\hat{v}_i \hat{u}_i| \leq \frac{1}{p} \sum_{i=1}^n |\hat{v}_i|^p + \frac{1}{q} \sum_{i=1}^n |\hat{u}_i|^q = \frac{1}{p \|\hat{\mathbf{v}}\|_p} \sum_{i=1}^n |v_i|^p + \frac{1}{p \|\hat{\mathbf{v}}\|_p} \sum_{i=1}^n |u_i|^p = \frac{1}{p} + \frac{1}{q} = 1.$$

Thus, we have that

$$\left| \sum_{i=1}^n \hat{v}_i \hat{u}_i \right| = \frac{1}{\|\mathbf{v}\|_p \|\mathbf{u}\|_p} \left| \sum_{i=1}^n v_i u_i \right| \leq 1,$$

implying that $|\sum_{i=1}^n v_i u_i| \leq \|\mathbf{v}\|_p \|\mathbf{u}\|_q$. □

Using Holder's Inequality, we next state and prove Minkowski's Inequality, which is equivalent to proving property 3.

Theorem 6.10 (Minkowski's Inequality). Let $1 \leq p \leq \infty$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$. Then,

$$\|\mathbf{v} + \mathbf{u}\|_p \leq \|\mathbf{v}\|_p + \|\mathbf{u}\|_p.$$

Proof. Consider $\|\mathbf{v} + \mathbf{u}\|_p^p$. By the triangle inequality for real numbers; we have that

$$\|\mathbf{v} + \mathbf{u}\|_p^p = \sum_{i=1}^n |v_i + u_i|^p \leq \sum_{i=1}^n |v_i + u_i|^{p-1} (|v_i| + |u_i|).$$

By distributivity,

$$\sum_{i=1}^n |v_i + u_i|^{p-1} (|v_i| + |u_i|) = \sum_{i=1}^n |v_i + u_i|^{p-1} |v_i| + \sum_{i=1}^n |v_i + u_i|^{p-1} |u_i|.$$

By Holder's Inequality

$$\sum_{i=1}^n |v_i + u_i|^{p-1} |v_i| + \sum_{i=1}^n |v_i + u_i|^{p-1} |u_i| \leq \left(\sum_{i=1}^n |v_i + u_i|^p \right)^{\frac{p-1}{p}} \left(\sum_{i=1}^n (|v_i| + |u_i|)^p \right)^{\frac{1}{p}}.$$

By the triangle inequality,

$$\begin{aligned} \left(\sum_{i=1}^n |v_i + u_i|^p \right)^{\frac{p-1}{p}} \left(\sum_{i=1}^n (|v_i| + |u_i|)^p \right)^{\frac{1}{p}} &\leq \left(\sum_{i=1}^n |v_i + u_i|^p \right)^{\frac{p-1}{p}} \left(\left(\sum_{i=1}^n |v_i|^p \right)^{\frac{1}{p}} + \left(\sum_{i=1}^n |u_i|^p \right)^{\frac{1}{p}} \right) \\ &= \|\mathbf{u} + \mathbf{v}\|_{\mathbf{p}}^{p-1} (\|\mathbf{v}\|_{\mathbf{p}} + \|\mathbf{u}\|_{\mathbf{p}}). \end{aligned}$$

Since $\|\mathbf{v} + \mathbf{u}\|_{\mathbf{p}}^p \leq \|\mathbf{v} + \mathbf{u}\|_{\mathbf{p}}^{p-1} (\|\mathbf{v}\|_{\mathbf{p}} + \|\mathbf{u}\|_{\mathbf{p}})$, Minkowski's Inequality follows by dividing both sides of the inequality by $\|\mathbf{v} + \mathbf{u}\|_{\mathbf{p}}^{p-1}$. \square

Thus, Minkowski's Inequality reveals that property 3 of a norm is satisfied for any p between 1 and ∞ . We next compare the sizes of different norms:

Proposition 6.11. (1) $\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_2$

Proof. Consider $\|\mathbf{x}\|_{\infty}^2$. By definition, that $\|\mathbf{x}\|_{\infty}^2 = \max_{1 \leq i \leq n} |x_i|^2$. Let $\max_{1 \leq i \leq n} |x_i|^2 = v_k$. Thus, since for all $i \in \{1, 2, \dots, n\}$, $|v_i|^2 \geq 0$,

$$\|\mathbf{x}\|_{\infty}^2 = |\mathbf{x}_k|^2 \leq |\mathbf{x}_k|^2 + \sum_{j \neq k} |\mathbf{x}_j|^2 = \sum_{i=1}^n |\mathbf{x}_i|^2 = \|\mathbf{x}\|_2^2.$$

Thus, it follows that $\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_2$. \square

(2) $\|\mathbf{x}\|_2 \leq \sqrt{\mathbf{n}} \|\mathbf{x}\|_{\infty}$

Proof. If we let $\|\mathbf{x}\|_{\infty} = |\mathbf{x}_k|$ where $k \in \{1, 2, \dots, n\}$, then by definition it follows

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^n |\mathbf{x}_i|^2 \leq \sum_{i=1}^n |\mathbf{x}_k|^2 = \mathbf{n} |\mathbf{x}_k|^2 = \mathbf{n} \|\mathbf{x}\|_{\infty}^2.$$

Thus, $\|\mathbf{x}\|_2 \leq \sqrt{\mathbf{n}} \|\mathbf{x}\|_{\infty}$. \square

Given these observations, a interesting mathematical question to ask is whether all norms are comparable. We make this notion precise by defining norm equivalence.

Definition 6.12 (Norm Equivalence). We say that two norms $\|\cdot\|_a, \|\cdot\|_b$, are equivalent if there exist a $c, C \in \mathbb{R}$ such that for all $\mathbf{x} \in \mathbb{R}^n$

$$C\|x\|_b \leq \|x\|_a \leq c\|x\|_b.$$

From the previous exercise we have the following relationship between the 2-norm and the ∞ -norm:

$$\|\mathbf{x}\|_{\infty} \leq \|\mathbf{x}\|_2 \leq \sqrt{\mathbf{n}} \|\mathbf{x}\|_{\infty},$$

which implies that the 2-norm and ∞ -norm are equivalent for any finite dimensional space \mathbb{R}^n . It turns out we can generalize this for any norm.

Theorem 6.13 (Norms in Finite-Dimensional Vector Spaces are equivalent). *Let V be a finite dimensional vector space. If $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on V , then there exists $c, C \in \mathbb{R}$ such that for all, $v \in V$ $c\|v\|_b \leq \|v\|_a \leq C\|v\|_b$*

Proof. Without loss of generality, we may prove the theorem for vectors $v \in V$ such that $\|v\| = 1$. For an arbitrary $v \in V$, let $v = \|v\|e_1$ where e_1 is the canonical basis vector and $\|e_1\| = 1$. If there exists c, C such that $c\|e_1\|_b \leq \|e_1\|_a \leq C\|e_1\|_b$ then this implies that

$$c\|v\| \|e_1\|_b \leq \|v\| \|e_1\|_a \leq C\|v\| \|e_1\|_b$$

which proves the claim for the arbitrary vector v . Thus, in essence we are showing that norms are equivalent on the compact unit ball in the 2-norm or Euclidean Distance.

Now we will show that the function $\|v\|_i$ is continuous with respect to $\|\cdot\|_2$. First, recall that we showed above that the infinity norm is less than or equal to the 2-norm of a vector. Thus, $\max_{i \in [n]} |x_i - y_i| < \|x - y\|_2$. Now, let $\epsilon > 0$ and let $\delta = \frac{\epsilon}{\sum_{i=1}^n \|e_i\|_2}$. If $\|x - y\|_2 < \delta$, then $\max_{i \in \{1, 2, \dots, n\}} |x_i - y_i| < \delta$. This implies that, $\max_{i \in [n]} |x_i - y_i| \sum_{i=1}^n \|e_i\|_2 < \epsilon$. Therefore:

$$\begin{aligned} \left| \|x\| - \|y\| \right| &\leq \|x - y\| \\ &= \left\| \sum_{i=1}^n (x_i - y_i) e_i \right\| \\ &\leq \sum_{i=1}^n |x_i - y_i| \|e_i\| \\ &\leq \max_{i \in [n]} |x_i - y_i| \sum_{i=1}^n \|e_i\| \\ &\leq \epsilon. \end{aligned}$$

It follows that $\left| \|x\| - \|y\| \right| < \epsilon$ if $\|x - y\|_2 < \delta$.

Since $\|v\|_a$ and $\|v\|_b$ are continuous functions with respect to the 2-norm, then $f(v) = \frac{\|v\|_a}{\|v\|_b}$ is continuous as well provided that $\|\cdot\|_b$ is nonzero, which holds since $\|v\| \neq 0$. Since $f(v)$ is continuous on the compact unit ball, there exist c, C such that

$$c \leq \frac{\|v\|_a}{\|v\|_b} \leq C.$$

Thus, by multiplying by $\|v\|_b$ we see that the norms are equivalent. \square

Since all norms are equivalent on any finite dimensional space, all forms of error analysis are equivalent. This implies that we are free to choose which norm we deem suitable. Moreover, we can extend the concept of norms to finite-dimensional matrix spaces.

Definition 6.14 (Induced Matrix Norm). The induced Matrix Norm for a matrix is defined as

$$\|A\| = \max_{\mathbf{x} \in \mathbb{C} \setminus \{\mathbf{0}\}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

It is important to notice that we are in essence defining the norm of a matrix as the ratio of the norms of two vectors. By multiplying matrix A on the left, we are forming a linear combination of its columns, i.e. a vector with n entries. We'll continue our discussion of matrix norms with condition numbers, a precise method of measuring error with norms.

6.2. Condition Numbers. Condition numbers quantify the “sensitivity” of the output of a function to changes of its input. We will then, apply this notion to analyze the error in our numerical algorithms. We first consider the general concept of condition numbers of functions, which as we will see, closely resemble the notion of the derivative of a function.

Definition 6.15 (Absolute Condition Number). If f is a function between finite-dimensional spaces V and W with norms $\|\cdot\|_V, \|\cdot\|_W$, then we define the absolute condition number of f , denoted by $\text{cond } f$, as

$$\text{cond}(f) = \sup_{x, y \in V, x \neq y} \frac{\|f(y) - f(x)\|_W}{\|y - x\|_V}.$$

If $\text{cond}(f) = +\infty$ or $1 \ll \text{cond}(f) < +\infty$, then we say that f is ill-conditioned.

We can also create a local version of the above by considering a neighborhood of an arbitrary $x \in V$.

Definition 6.16 (Relative Local Condition Number). Let $x \in V$ and let $\|\delta(x)\|_\alpha > 0$, where $\alpha \in [1, \infty)$ such that $\delta(x) \in V$. We define the relative local condition number of f as

$$\text{cond}_x(f) = \sup_{x + \delta(x) \in V} \frac{\|f(x + \delta(x)) - f(x)\|_W / \|f(x)\|_W}{\|\delta(x)\|_V / \|x\|_V}.$$

Example 7. We consider the function $f : (0, +\infty) \rightarrow \mathbb{R}$ where $f(x) = \sqrt{x}$ for all $x \in (0, +\infty)$. Since all norms are equivalent, we will work with $|\cdot|$ for convenience. Since f is differentiable,

$$\begin{aligned} \text{cond}_x(f) &= \sup_{x+\delta(x) \in V} \frac{|f(x+\delta(x)) - f(x)|/|f(x)|}{|\delta(x)|/|x|} \\ &= \sup_{x+\delta(x) \in V} \frac{|f(x+\delta(x)) - f(x)|}{|\delta(x)|} \frac{|x|}{\sqrt{x}} \\ &= f'(x)\sqrt{x} \\ &= \frac{1}{2\sqrt{x}}\sqrt{x} \\ &= \frac{1}{2}. \end{aligned}$$

Thus, by analyzing the absolute local condition number on the space \mathbb{R} we can see that the notion of a condition number is somewhat connected to the first derivative of the function in question. Thinking of our past days in Calculus, this result verifies our intuition as first derivatives express how changes in function input affect output.

We will now use the the relative local condition number to derive the notion of the condition number for a nonsingular matrix. Let $A \in \mathbb{R}^{n \times n}$ suppose that \mathbb{R}^n is equipped with a vector norm. Consider the function $A^{-1} : \mathbf{b} \rightarrow A^{-1}\mathbf{b}$. The relative local condition number of A^{-1} is defined as

$$\begin{aligned} \text{cond}_{\mathbf{b}}(A^{-1}) &= \sup_{\delta\mathbf{b}} \frac{\|A^{-1}(\mathbf{b} + \delta\mathbf{b}) - A^{-1}\mathbf{b}\|/\|A^{-1}\mathbf{b}\|}{\|\delta\mathbf{b}\|/\|\mathbf{b}\|} \\ &\leq \|A^{-1}\| \frac{\|\mathbf{b}\|}{\|A^{-1}\mathbf{b}\|} \\ &= \|A^{-1}\| \frac{\|A(A^{-1}\mathbf{b})\|}{\|A^{-1}\mathbf{b}\|} \\ &\leq \|A^{-1}\| \frac{\|A\| \|(A^{-1}\mathbf{b})\|}{\|A^{-1}\mathbf{b}\|} \\ &= \|A^{-1}\| \|A\|. \end{aligned}$$

An identical argument will show that $\text{cond}_{\mathbf{x}}(A) \leq \|A\| \|A^{-1}\|$, when we consider the function $A : \mathbf{x} \in \mathbb{R}^n \rightarrow A\mathbf{x} \in \mathbb{R}^n$. The fact that $\text{cond}_{\mathbf{b}}(A^{-1}) \leq \|A^{-1}\| \|A\|$ and $\text{cond}_{\mathbf{x}}(A) \leq \|A\| \|A^{-1}\|$ motivates the following definition:

Definition 6.17 (Condition number of a nonsingular Matrix). We define the condition number of a nonsingular matrix A as $\kappa(A) = \|A\| \|A^{-1}\|$

7. LEAST SQUARES: WHERE SOLUTIONS TO $A\mathbf{x} = \mathbf{b}$ MIGHT NOT EXIST

In sections 4 and 5 we have worked with invertible square matrices $A \in \mathbb{R}^{n \times n}$. However, it may be the case where we have to solve a rectangular system of linear equations of the form $A\mathbf{x} = \mathbf{b}$ where $A \in \mathbb{R}^{m \times n}$, $m \geq n$ and, $\text{rank}(A) = n$. Typically, such a system will not have a solution. A common example of this problem is polynomial data fitting. Given x_1, \dots, x_m data points and y_1, \dots, y_m observations, we are interested in finding an $n - 1$ (where $n < m$) degree polynomial p such that $p(x_i) = y_i$ for all i . We may still approximate a solution to such a system by finding a vector $\mathbf{x} \in \mathbb{R}^m$ which minimizes $A\mathbf{x} - \mathbf{b}$. Given that all norms are equivalent, we will minimize this quantity in the 2-norm. Thus, the least squares problems for a rectangular system with no solution is as follows:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \|A\mathbf{x} - \mathbf{b}\|_2.$$

The desired polynomial in the above situation is called the Least Squares Polynomial, due to its connection with the Least Squares problem. We will now provide a brief geometric interpretation of the Least Squares problem followed by an overview of the QR algorithm used to solve tackle such problems.

7.1. Geometric Interpretation. Note that by trying to $\min_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{Ax} - \mathbf{b}\|_2$, we are essentially finding the point $\mathbf{Ax} \in \text{range}(A)$ that is closest to \mathbf{b} . Basic geometric intuition tells us that we must find the perpendicular distance of \mathbf{b} to \mathbf{Ax} . Thus, to find \mathbf{x} , we project \mathbf{b} into the space \mathbf{Ax} using an orthogonal projection P . Therefore, the Least Squares problem reduces to finding the point \mathbf{x} where $\mathbf{Ax} = P\mathbf{b}$.

7.2. Least Squares through QR Factorization. We can approach the Least Squares problem through QR factorization. The Gram-Schmidt Algorithm will produce a matrix $Q \in \mathbb{R}^{m \times n}$ with orthonormal columns and an upper triangular matrix $R \in \mathbb{R}^{n \times n}$. It is important to note that Gram-Schmidt will always produce a QR factorization, however some of the rows of the R matrix may be empty. We can project \mathbf{b} into range A , by $P = QQ^T$. By using the $A = QR$ factorization of A ; we can insert both expressions into the expression $\mathbf{Ax} = \mathbf{b}$ and obtain the expression $QR = QQ^T\mathbf{b}$, or equivalently, $R\mathbf{x} = Q^T\mathbf{b}$. Thus, once the vector $Q^T\mathbf{b}$ is computed, all that is left is to set the vector equal to $R\mathbf{x}$ and solve the system backwards for \mathbf{x} . Expressed Algorithmically:

- (1) Compute $A = QR$.
- (2) Compute $Q^T\mathbf{b}$.
- (3) Solve for \mathbf{x} backwards, $R\mathbf{x} = Q^T\mathbf{b}$.

To conclude, one can compute the stability of a least squares solution by computing the condition number of the matrix A , as defined in Section 6.

8. ACKNOWLEDGEMENTS

Working on this paper has definitely reinforced my love for mathematics while simultaneously widening my knowledge and making me a more mathematically mature person. I would like to thank the Department of Mathematics at the University of Chicago for providing me with this opportunity and to Peter May for providing the template for this paper. I hope that by now, the reader is as excited about Numerical Analysis as I am.

REFERENCES

- [1] Strang, Gilbert. *Linear Algebra and Its Applications*; pgs. 136-137. New York: Academic Press, INC, 1976. Print.
- [2] Suli, Endre, and David F. Mayers. *An Introduction to Numerical Analysis*. N.p.: Cambridge University Press, 2003. Print.
- [3] Trefethen, Lloyd N., and David Bau, III. *Numerical Linear Algebra*. N.p.: SIAM, 1997. Print.
- [4] Tulsiani, Madhur *REU 2013: Apprentice Class, Lecture 1 Notes*. N.p., n.d. Web. 31 Aug. 2013. <http://ttic.uchicago.edu/~madhurt/courses/reu2013/lecture1.pdf>.