UNIVERSITY OF CAPE COAST

REGULARIZATION OF ILL-CONDITIONED LINEAR SYSTEMS

JOSEPH ACQUAH

2009

UNIVERSITY OF CAPE COAST

REGULARIZATION OF ILL-CONDITIONED LINEAR SYSTEMS

BY

JOSEPH ACQUAH

THESIS SUBMITTED TO
THE DEPARTMENT OF MATHEMATICS & STATISTICS OF THE SCHOOL
OF PHYSICAL SCIENCES, UNIVERSITY OF CAPE COAST
IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR AWARD OF
MASTER OF PHILOSOPHY DEGREE IN MATHEMATICS

DECEMBER 2009

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this University or elsewhere.

Candidate's Signature ..... Date .....

Name .....

Supervisors' Declaration

We hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Principal Supervisor's Signature ..... Date .....

Name: .....

Co-Supervisor's Signature ..... Date .....

Name: .....

ABSTRACT

The numerical solution of the linear system $\mathbf{Ax} = \mathbf{b}$, arises in many branches of applied mathematics, sciences, engineering and statistics. The most common source of these problems is in the numerical solutions of ordinary and partial differential equations, as well as integral equations. The process of discretization by means of finite differences often leads to the solution of linear systems, whose solution is an approximation to the solution of the original differential equation. If the coefficient matrix is ill-conditioned or rank-deficient, then the computed solution is often a meaningless approximation to the unknown solution. Regularization methods are often used to obtain reasonable approximations to ill-conditioned systems.

However, the methods for choosing an optimal regularization parameter is not always clearly defined. In this dissertation, we have studied various methods for solving ill-conditioned linear systems, using the Hilbert system as a prototype. These systems are highly ill-conditioned. We have examined various regularization methods for obtaining meaningfully approximations to such systems. Tikhonov Regularization method proved to be the method of choice for regularizing rank deficient and discrete ill-posed problems compared to the Truncated Singular Values decomposition and the Jacobi and Gauss-Seidel Preconditioner's for boundary values problems. The truncated singular value decomposition truncate the harmful effect of the small singular values on the solution by replacing them with exact zero. The truncation improves the solution to an extent and the solution deteriorates again. The maximum error in the solution occurs at the cut-off level $\lambda = 2.2520 \times 10^{-10}$. The optimal solution was obtained at $\lambda = 2.2520 \times 10^{-10}$. The Jacobi and the Gauss-Seidel preconditioner's for sparse systems also gave an optimal solution.

Our results using Tikhonov method, shows that in *order* 0 regularization, the accuracy of the solution increases with increasing values of the regularization parameter, with optimal regularization parameter of 10^0, giving an accuracy of 15 digits. *Order* 2 regularization gave an accuracy of about 13 digits, with optimal parameter of 10^{-9}, while *order* 1 regularization gave an accuracy of only 3 digits, corresponding to an optimal parameter of 10^{-13}. In all the three cases, the optimal regularization parameters were determined by inspection, using the *minimum error*. The L-Curve method failed to indicate the optimal regularization parameter. We applied regularization to the solution of an ill-conditioned discretized Fredholm integral equation of the first kind.

First, we transformed the integral equation into a linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 17×17 positive definite matrix. None of the standard methods for solving linear systems gave the desired solution. The accuracy of the solution increases with increasing values of the regularization parameter. The regularized solutions of order one and two, gave an accuracy of about 3 digits accuracy, with parameter values $\lambda = 10^{-1}$ and $\lambda = 10^{-3}$ respectively, while order zero shows no accuracy in the regularized solutions. The L-curve was applied to determine the optimal regularization parameter. The optimal regularization parameter corresponding to the optimal solution was determined at the corner part of the L-curve. The L-curve for order two regularization gave us the optimal solution with a regularization parameter value $\lambda = 10^{-3}$. For order zero and one, the regularization parameter concentrated at the sharp corner of the L-curve did not approximate to the exact solution. Order zero and one in this case has no optimal regularization parameter on the L-curve.

ACKNOWLEDGEMENTS

I wish to express my profound gratitude to my Principal Supervisor, Professor Francis Benyah of University of Western Cape, Cape Town, South Africa and my Co-Supervisor, Dr .E.K. Essel of the Department of Mathematics and Statistics, University of Cape Coast, for introducing me to this area of research. I have benefited immensely from their experience, constant encouragement, patience, pieces of advice and understanding which together have enhanced my research work.

My sincere thanks also goes to Dr(Mrs) N.G. Mensah and Professor B.K. Gordor, of the Department of Mathematics and Statistics for their motivation and guidance.

I would also like to thank Professor F.K. Allotey, President of the Institute for Mathematical Science, Accra, Ghana for enriching me academically through a lot of conferences and assisting me financially throughout this study.

Last but not the least, my special thanks goes to all the staff(both the teaching and non-teaching) and my fellow Master students at the Department of Mathematics and Statistics, U.C.C for sharing constructive ideas and supporting me in diverse ways during my graduate studies at the Department.

Finally, I wish to thank my wife, Mrs Ernestina Acquah, my brother Dr .H. De-Graft Acquah, and my son Micheal De-Graft Acquah for their encouragement and prayers.

DEDICATION

To

My Family

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Background of the Study

The concept of ill-posed problems goes back to Hadamard, at the beginning of the 19^{th} century, Hadamard(1923). Hadamard essentially defined a problem as ill-posed if the solution is not unique, or if it is not a continuous function of a given data. That is, if an arbitrary small perturbation of a data can cause an arbitrarily large perturbation of the solution. Hadamard believed that ill-posed problems were "artificial", in the sense that they could not describe physical systems. However, this analogy was wrong according to Hansen(1992). Ill-posed problems arise in many areas of science and engineering in the form of inverse problems. Inverse problems arise naturally in determining an unknown input that gives rise to a measured output signal. Rank-deficient or discrete ill-posed problems are characterized by having a coefficient matrix that is very ill-conditioned.

Given an m $\times$ n matrix $\mathbf{A}$, for m $\geq$ n and a vector $\mathbf{b} \in \mathbb{R}^n$, **the linear systems** problem is to find a vector $\hat{\mathbf{x}} \in \mathbb{R}$ satisfying the linear equation

$$\mathbf{Ax} = \mathbf{b}. \tag{1.1}$$

The linear system in Equation 1.1 arises in many branches of applied mathematics, sciences, engineering and statistics. If the matrix is ill-conditioned or rank-deficient, then the computed solution is often a meaningless approximation to the exact solution. The need arises to find answers to the following questions : What kind of ill-conditioning system do we have at

hand and how do we deal with it? Is the problem rank deficient problem or ill-posed ? Is it possible to include additional information to stabilize the solution? Which additional information is available and is it suitable for stabilization purposes? How much stabilization should be added?

Purpose of the Study

Ill-conditioned linear systems arise naturally in many applications, especially in the numerical solution of ordinary and partial differential equations. Such systems are typically ill-conditioned. Regularization methods are therefore needed in order to obtain meaningful solutions to them.

The aim of this thesis is to study various methods for solving ill-conditioned linear systems, including Hilbert systems, as well as those arising from the discretization of differential and integral equations. The study will also cover different regularization methods for stabilizing computations and methods for selecting an optimal regularization parameter.

The purpose of the study is summarized as follows:

1. To study various ill-conditioned linear systems, using the Hilbert system as a prototype.

2. To investigate various methods for stabilizing the computation of a Hilbert system.

3. To study various regularization methods.

4. To apply the methods to stabilize the solutions of boundary-value problems, and integral equations.

Mathematical Background

Matrix and Vector Norms

A vector space V is said to be a **normed linear space** if for each vector $\mathbf{v} \in \mathrm{V}$, there exist a real number $\|\mathbf{v}\|$ called the **norm** of $\mathbf{v}$, satisfying

1. $\|\mathbf{v}\| \geq 0 \quad$ and $\quad \|\mathbf{v}\| = 0 \quad$ if and only if $\mathbf{v} = 0$
2. $\|\lambda\mathbf{v}\| = |\lambda| \, \|\mathbf{v}\| \quad$ for any scalar $\lambda \in \mathbb{R}$
3. $\|\mathbf{v} + \mathbf{w}\| \leq \|\mathbf{v}\| + \|\mathbf{w}\| \quad$ for all $\mathbf{v}, \mathbf{w} \in \mathrm{V} \quad$ (**triangle inequality**)

Examples of Vector Norms

Let $\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n}]^\mathrm{T}$ be a vector in $\mathbb{R}^n$. Then, each of the following defines a norm on $\mathbb{R}^n$. In fact, they are the most commonly used norms on $\mathbb{R}^n$.

$$\|\mathbf{x}\|_1 = \sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i}| = |\mathrm{x}_1| + |\mathrm{x}_2| + \cdots + |\mathrm{x}_\mathrm{n}| \quad (\mathbf{l_1 \, norm}) \tag{1.2}$$

$$\|\mathbf{x}\|_2 = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i}|^2 \right)^{1/2} = \sqrt{\mathrm{x}_1^2 + \mathrm{x}_2^2 + \cdots + \mathrm{x}_\mathrm{n}^2} \quad (\mathbf{l_2 \, norm}) \tag{1.3}$$

$$\|\mathbf{x}\|_\infty = \max_{\mathrm{i}} |\mathrm{x_i}| \quad (\mathbf{l_\infty \, norm}) \tag{1.4}$$

$$\|\mathbf{x}\|_\mathrm{p} = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} |\mathrm{x_i}|^\mathrm{p} \right)^{1/\mathrm{p}} \quad \mathrm{p} \geq 1, \quad (\mathbf{l_p \, norm}). \tag{1.5}$$

Example 1.1. Let $\mathbf{x} = [2, -3, 0, 1, -4] \in \mathbb{R}^5$. *Then*

(a) $\|\mathbf{x}\|_1 = |2| + |-3| + |0| + |1| + |-4| = 10.$

(b) $\|\mathbf{x}\|_2 = \sqrt{2^2 + (-3)^2 + (0)^2 + (1)^2 + (-4)^2} = \sqrt{30}$

(c) $\|\mathbf{x}\|_\infty = 4.$

It can be verified that Equations 1.2—1.4 satisfies the definition of a norm. However, the proof of the Triangle Inequality for $\|\cdot\|_2$ requires the following well-known inequality.

Theorem 1.1 (Cauchy-Schwartz Inequality)**.** For each $\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n}]^\mathrm{T}$ and

$\mathbf{y} = [\mathrm{y}_1, \mathrm{y}_2, \cdots, \mathrm{y}_\mathrm{n}]^\mathrm{T} \in \mathbb{R}^\mathrm{n}$,

$$\sum_{\mathrm{i}}^{\mathrm{n}} |\mathrm{x}_\mathrm{i}\mathrm{y}_\mathrm{i}| \ \leq \ \left[\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_\mathrm{i}^2\right)\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{y}_\mathrm{i}^2\right)\right]^{1/2} = \|\mathbf{x}\|_2 \, \|\mathbf{y}\|_2 \tag{1.6}$$

Definition 1.1 (Equivalent Norms)**.** Two vector norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are said to be **equivalent** if the ratio of a vector's length in one norm to a vector's length in another norm is bounded from above and below by constants, say c_1 and c_2. Thus, are independent of the vectors.

For example, it can be shown that

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_1 \ \leq \mathrm{n} \, \|\mathbf{x}\|_\infty \tag{1.7}$$

The inequality above shows that $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are equivalent, since

$$1 \quad \leq \frac{\|\mathbf{x}\|_1}{\|\mathbf{x}\|_\infty} \quad \leq \mathrm{n}$$

In general, if $\mathrm{p} \geq \mathrm{q}$ then

$$\|\mathbf{x}\|_\mathrm{p} \ \leq \|\mathbf{x}\|_\mathrm{q} \ \leq \mathrm{n}^{(\mathrm{p}-\mathrm{q})/(\mathrm{pq})} \, \|\mathbf{x}\|_\mathrm{p} \tag{1.8}$$

Since, $1 \quad \leq \quad \|\mathbf{x}\|_\mathrm{q}/\|\mathbf{x}\|_\mathrm{p} \quad \leq \quad \mathrm{n}^{(\mathrm{p}-\mathrm{q})/(\mathrm{pq})}$.

Definition 1.2. A sequence of vectors, $\{\mathbf{x}^\mathrm{k}\}_1^\infty$ is said to converge to a vector $\mathbf{x}$ with respect to the norm $\|\cdot\|$, if and only if

$$\lim_{\mathrm{k}\to\infty} \|\mathbf{x}^\mathrm{k} - \mathbf{x}\| = 0. \tag{1.9}$$

For example, a sequence of vectors $\{\mathbf{x}^k\}_1^\infty$ in $\mathbb{R}^n$ converges to the vector $\mathbf{x}$ with respect to the norm $\|\cdot\|_\infty$, if and only if

$$\lim_{\mathrm{k}\to\infty} \mathrm{x_i^k} = \mathrm{x_i} \text{ for } 1 \ \leq \ \mathrm{i} \ \leq \ \mathrm{n}.$$

In Definition 1.2, the norm was not specified. This is because, in $\mathbb{R}^n$, all norm are equivalent. Therefore convergence in one norm automatically implies convergence in another norm.

Example 1.2. The sequence $\{\mathbf{x}^k\}$ where

$$\mathbf{x}^{\mathrm{k}} = \left[\mathrm{e}^{-\mathrm{k}},\ \mathrm{k}\sin\frac{1}{\mathrm{k}},\ 2+\mathrm{k}^{-2}\right]^{\mathrm{T}}$$

converges to $\mathbf{x} = (1,\ 1,\ 2)^T$.

Matrix Norms

In the previous section, norms were used to measure the length of vectors. In this section we study some of the standard matrix norms defined on the vector space $\mathbf{M}_{nn}$, of n × n matrices. Matrix norms play a crucial role in numerical linear algebra. For instance, the norm of an m × n matrix $\mathbf{A}$ is useful in determining the accuracy of the computed solution of the linear system $A\mathbf{x} = \mathbf{b}$.

Since $\mathbf{M}_{nn}$ is isomorphic to $\mathbb{R}^{n^2}$, the vector norms we used in the previous section can also be used, in a limited way, to measure the *size* of matrices. In addition to the standard vector space operations on $\mathbf{M}_{nn}$, the operation of *multiplication* is also defined. There is therefore the need to relate the norm of the product $\mathbf{AB}$ to the norms of $\mathbf{A}$ and $\mathbf{B}$.

Definition 1.3. A **matrix norm** on $\mathbf{M}_{nn}$ is a real-valued function $\|\cdot\|$ from $\mathbf{M}_{nn}$ into $\mathbb{R}$ with the following properties:

$$\|\mathbf{A}\| \geq 0 \text{ for all } \mathbf{A} \in \mathbf{M}_{n,n}, \quad \text{and } \|\mathbf{A}\| = 0, \text{iff } \mathbf{A} = 0. \tag{1.10}$$

$$\|\alpha\mathbf{A}\| = |\alpha| \, \|\mathbf{A}\| \quad \text{for all } \alpha \in \mathbb{R} \text{ and } \mathbf{A} \in \mathbf{M}_{nn}. \tag{1.11}$$

$$\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\| \quad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbf{M}_{nn}. \tag{1.12}$$

$$\|\mathbf{AB}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|. \tag{1.13}$$

There are many ways in which we can construct matrix norms to satisfy the properties given above. However, since matrices operate on vectors, we define a matrix norm $\|\cdot\|$ so as to be **compatible** with a vector norm by imposing the requirement that

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\|, \tag{1.14}$$

for all $\mathbf{A} \in \mathrm{M_{nn}}$ and for all $\mathbf{x} \in \mathbb{R}^n$.

We now derive a matrix norm form a given vector norm. For any $\mathbf{x} \in \mathbb{R}^n$, the matrix $\mathbf{A}$ transforms the vector $\mathbf{x}$ into another vector $\mathbf{Ax}$. One way of measuring the "size" of $\mathbf{A}$ is by comparing $\|\mathbf{x}\|$ with $\|\mathbf{Ax}\|$ using any convenient vector norm. The ratio $\frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}$ is a measure of the stretching capability of $\mathbf{A}$. We also expect the matrix norm $\|\mathbf{A}\|$, to be compatible with the vector norm being used, see Equation 1.14. Therefore, we must have

$$\|\mathbf{Ax}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\| \qquad \text{or} \qquad \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|.$$

The maximum stretch over all possible $\mathbf{x}$ can then be taken as the definition of $\|\mathbf{A}\|$. That is,

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\| \neq 0} \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|}. \tag{1.15}$$

This matrix norm is often referred to as a **compatible matrix norm** or a **natural norm**. Thus, if let $\mathbf{y} = \mathbf{x}/\|\mathbf{x}\|$ for any nonzero vector $\mathbf{x} \in \mathbb{R}^n$, then $\|\mathbf{y}\| = 1$ and $\|\mathbf{Ay}\| = \|\mathbf{Ax}\|/\|\mathbf{x}\|$. The definition in Equation 1.15 is equivalent to $\|\mathbf{Ay}\|$, and can be prove by the theorem below.

Theorem 1.2. The norm defined by

$$\|\mathbf{A}\| = \max_{\|\mathbf{y}\|=1} \|\mathbf{A}\mathbf{y}\| \tag{1.16}$$

defines a matrix norm.

Proof. We verify properties 1.10 — 1.13.

(i) Since $\|\mathbf{y}\| = 1, \quad \mathbf{y} \neq \mathbf{0}$, and if $\mathbf{A} \neq \mathbf{0}$, then $\|\mathbf{A}\| = \max_{\|\mathbf{y}\|=1} \|\mathbf{A}\mathbf{y}\| \ \geq \ 0$.
If $\mathbf{A} = \mathbf{0}$, then $\mathbf{A}\mathbf{y} = \mathbf{0}$ for all $\mathbf{y} = 1$, and so $\|\mathbf{A}\mathbf{y}\| = 0$ for all $\|\mathbf{y}\| = 1$.
Hence $\|\mathbf{A}\| = \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y}\| = 0$.
Conversely, $\|\mathbf{A}\| = 0 \Rightarrow \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y}\| = 0, \Rightarrow \|\mathbf{A}\mathbf{y}\| = 0 \Rightarrow \mathbf{A}\mathbf{y} = \mathbf{0}$
for all $\mathbf{y}$, $\Rightarrow \mathbf{A} = \mathbf{0}$.

(ii) $\|\alpha\mathbf{A}\| = \max_{\mathbf{y}=1} \|\alpha\mathbf{A}\mathbf{y}\| = \max_{\mathbf{y}=1} |\alpha| \, \|\mathbf{A}\mathbf{y}\| = |\alpha| \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y}\| = |\alpha| \, \|\mathbf{A}\|$

(iii) $\|\mathbf{A} + \mathbf{B}\| = \max_{\mathbf{y}=1} \|(\mathbf{A} + \mathbf{B})\mathbf{y}\| = \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{y}\|$
$\leq \max_{\mathbf{y}=1} (\|\mathbf{A}\mathbf{y}\| + \|\mathbf{B}\mathbf{y}\|) \leq \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y}\| + \max_{\mathbf{y}=1} \|\mathbf{A}\mathbf{y}\| = \|\mathbf{A}\| + \|\mathbf{B}\|$
This implies that $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.

(iv) $\|\mathbf{A}\,\mathbf{B}\| = \max_{\mathbf{y}=1} \|\mathbf{A}\,\mathbf{B}(\mathbf{y})\| = \max_{\mathbf{y}=1} \|\mathbf{A}\,(\mathbf{B}\mathbf{y})\| \leq \max_{\mathbf{y}=1} (\|\mathbf{A}\| \, \|\mathbf{B}\mathbf{y}\|)$
That is $\|\mathbf{A}\,\mathbf{B}\| \leq \max_{\mathbf{y}=1} (\|\mathbf{A}\| \, \|\mathbf{B}\| \|\mathbf{y}\|)$
Hence, $\|\mathbf{A}\,\mathbf{B}\| \leq \|\mathbf{A}\| \, \|\mathbf{B}\|$.

This shows that Equation 1.16 defines a matrix norm.

Examples of Matrix Norms

The most commonly used norms on an $n \times n$ matrix $\mathbf{A}$ are

(a) $\|\mathbf{A}\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|$ (max column-sum).

(b) $\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$ (max row-sum).

(c) $\|\mathbf{A}\|_2 = \sqrt{\max \text{eigenvalue}(\mathrm{A^T A})}$.

(d) $\|\mathbf{A}\|_\mathrm{F} = \left(\sum_{\mathrm{j}=1}^{\mathrm{n}} \sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{a}_{\mathrm{ij}}^2 \right)^{1/2}$ (Frobenius norm).

Example 1.3. Let $\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ -1 & 1 \end{bmatrix}$

Then,

•$\|\mathbf{A}\|_1 = 4$,

•$\|\mathbf{A}\|_\infty = 3$,

•$\|\mathbf{A}\|_F = \sqrt{1^2 + (1)^2 + 1^2 + 2^2 + (-1)^2 + 1^2} = 3$ and

$\|\mathbf{A}\|_2 = \sqrt{\max \text{eigenvalue}(\mathbf{A^T A})} = \sqrt{\max(7,\ 2)} = \|\mathrm{A}\|_2 = \sqrt{7}$.

Since the eigenvalues of $\mathrm{A^T A}$ are 7 and 2.

Note that the Frobenius norm is analogous to the Euclidean norm

$$\|\mathbf{x}\|_2 = \left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{x}_{\mathrm{i}}^2 \right)^{1/2} \qquad \text{for} \quad \mathbf{x} \in \mathbb{R}^{\mathrm{n}}.$$

Theorem 1.3. Let $\mathbf{A}$ be an $n \times n$ matrix. Then

$$\|\mathbf{A}\|_2 = \sqrt{\rho(\mathbf{A^T A})}$$

Proof. Let $\mathbf{x} = [\mathrm{x}_1, \mathrm{x}_2, \cdots, \mathrm{x}_\mathrm{n}]^\mathrm{T}$. Then

$$\|\mathbf{x}\|_2^2 = \mathrm{x}_1^2 + \mathrm{x}_2^2 + \cdots + \mathrm{x}_\mathrm{n}^2. \tag{1.17}$$

Also

$$\|\mathbf{A\,x}\|_2^2 = (\mathbf{A\,x})^\mathrm{T}\ (\mathbf{A\,x}) = \mathbf{x}^\mathrm{T} \left(\mathbf{A}^\mathrm{T}\,\mathbf{A}\right) \mathbf{x} \geq 0. \tag{1.18}$$

Since $\mathbf{A}^T\,\mathbf{A}$ is symmetric, it has an orthonormal set of eigenvectors,

$\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$; that is $\mathbf{v}_i^T \mathbf{v}_j = \delta_{ij}$ and

$$\left(\mathbf{A}^\mathrm{T}\,\mathbf{A}\right) \mathbf{v}_\mathrm{i} = \lambda_\mathrm{i} \mathbf{v}_\mathrm{i}, \tag{1.19}$$

where λ_i is an eigenvalue of $\mathbf{A}^T\,\mathbf{A}$ corresponding to the eigenvector $\mathbf{v}_i$.

Multiplying both sides of 1.19, on the left by $\mathbf{v}_i^T$ gives

$$\mathbf{v}_{\mathrm{i}}^{\mathrm{T}}\,\left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)\,\mathbf{v}_{\mathrm{i}} = \lambda_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}^{\mathrm{T}}\mathbf{v}_{\mathrm{i}} = \lambda_{\mathrm{i}} \geq 0.$$

Express the vector $\mathbf{x}$ as linear combination of the $\mathbf{v}_i$'s:

$$\mathbf{x} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}, \qquad (\alpha_{\mathrm{i}}'\mathrm{s}\,\mathrm{constants}) \tag{1.20}$$

Substituting 1.20 into 1.18 gives

$$\begin{aligned}
\|\mathbf{A}\,\mathbf{x}\|_2^2 &= (\mathbf{A}\,\mathbf{x})^{\mathrm{T}}\,(\mathbf{A}\,\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\,\mathbf{A}^{\mathrm{T}}\,\mathbf{A}\,\mathbf{x} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}}\,\mathbf{A})\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}, \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}^{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{n}}(\mathbf{A}^{\mathrm{T}}\,\mathbf{A})\alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}^{\mathrm{T}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}\lambda_{\mathrm{i}}\mathbf{v}_{\mathrm{i}}, \\
&= \sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}^2\lambda_{\mathrm{i}} \;\leq\; \max_{\mathrm{i}}\,|\lambda_{\mathrm{i}}|\sum_{\mathrm{i}=1}^{\mathrm{n}} \alpha_{\mathrm{i}}^2, \\
&= \max_{\mathrm{i}}\,|\lambda_{\mathrm{i}}| \qquad (\text{since } \|\mathbf{x}\|_2^2 = 1).
\end{aligned}$$

Thus

$$\|\mathbf{A}\|_2^2 = \max_{\|\mathbf{x}\|=1}\|\mathbf{A}\mathbf{x}\|_2 \;\leq\; \sqrt{\max_{\mathrm{i}}\,|\lambda_{\mathrm{i}}|}. \tag{1.21}$$

Now if the vector $\mathbf{x} = \mathbf{v}_k$ corresponds to

$$\max_{\mathrm{k}}\,\lambda_{\mathrm{k}},$$

then

$$\|\mathbf{A}\,\mathbf{x}\|_2^2 = \mathbf{v}_{\mathrm{k}}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}\,\mathbf{A}\mathbf{v}_{\mathrm{k}} = \mathbf{v}^{\mathrm{T}}\lambda_{\mathrm{k}}\mathbf{v}_{\mathrm{k}} = \max_{\mathrm{i}}\,\lambda_{\mathrm{i}}.$$

This shows that equality is attained in (1.21). Hence,

$$\|\mathbf{A}\|_2 = \sqrt{\max_{\mathrm{i}}\;\lambda_{\mathrm{i}}} = \sqrt{\rho(\mathbf{A}^{\mathrm{T}}\mathbf{A})}.$$

Literature Review

Studies into Regularization Techniques for ill-posed problems have been researched into by many researchers, and many contributions have been made by them.

In the 19^{th} century, Hadamard introduced the concepts of ill-conditioned or ill-posed problems, but his submission that they were "artificial problems" have been proved wrong. In fact, many problems in engineering and science such as image processing, signal processing and biomedical engineering are ill-posed problems. Regularization methods are therefore needed in order to obtain meaningful solutions to these problems.

In recent times, some work has been done in this field of study as a modification of what already is in existence. In a book published by Hansen(1997), a critical look at ill-posed problems were discussed. He focused on two main issues, the first is reliability and efficiency, both of which are important as the size and complexity of the computational problems grow and the demand for advanced real-time processing increases. The second issue is to characterize the regularizing effect of various methods. Hansen stipulated in his book that the development of robust and efficient implementation of numerical regularization methods and algorithms lies outside the scope of his research. In numerical regularization, one cannot expect to deal satisfactorily with ill-conditioned problems without both theoretical and numerical insight. Moreover, this topic has been extensively studied during the last few decades, and an impressive results have been achieved in many publications. A typical example is the book by Goncharisky and Bakushinsky(1994), titled Ill-posed problems: Theory and applications. In their study, Goncharisky and Bakushinsky made constructive iterative regularization procedures for solving both linear and non linear ill-posed problems.

However, other researchers have also contributed in this area of study. Benyah and Jennings(1998), have shown that optimal control computations are naturally ill-conditioned. These computations, can be stabilize by Regularization methods, Benyah and Jennings(1998). Nair and Pereverzev(2006), also used "A collocation method" to regularized an ill-posed problems. Thus

in short, appreciable work has been done in this field of study.

Outline of the Thesis

Chapter one of the thesis talks about the background of the study, the purpose or objective of the Study, some mathematical concepts necessary for the study and a detailed literature review.

Chapter two deals with preliminary concepts of linear algebra. Here, a critical look at conditioning of a problem and practical situations where they occur is considered.

Chapter three discusses least squares solutions of linear systems and singular value decomposition(SVD) as a numerical tool and establish the fact that the computed solution is usually meaningless when approximated to the exact solution.

In Chapter four, we study regularization methods for ill-conditioned linear systems. We looked at Truncated Singular Value Decomposition, followed by Preconditioning, and then Tikhonov Regularization method. The Tikhonov method turns out to be a good method for improving the accuracy of the solution of ill-conditioned system.

In Chapter five, we present a new approach for solving the Fredholm integral equation of the first kind using regularization methods coupled with the parameter-choice methods. A numerical analysis for this study shows that a better solution is possible when the correct regularization method is applied. Finally, we examine the new approach with an example to determine how best the approach works.

Chapter six talks about the summary of the work and observations that came out of the study. Some of the observations were discussed and appropriate conclusions drawn based on the result of the study.

CHAPTER TWO

PRELIMINARY CONCEPTS

Conditioning

Conditioning refers to the sensitivity of the solution of a problem to *small* changes in the input data. Let P(x) denote the value of a problem corresponding to input data x and δx denotes a small perturbation in x, then P is said to be **ill-conditioned**, if the relative error in the solution is much larger than the relative error in the data. That is :

$$\frac{|\mathrm{P}(\mathrm{x}+\delta\mathrm{x})-\mathrm{P}(\mathrm{x})|}{|\mathrm{P}(\mathrm{x})|} \gg \frac{|\delta\mathrm{x}|}{|\mathrm{x}|}.$$

For many problems, a **condition number** can be defined. If the condition number is *large*, then the problem is said to be **ill-conditioned**. On the other hand, if the condition number is *small*, then the problem is said to be **well-conditioned**. Consider the problem of computing a function $\mathbf{y} = \mathrm{f}(\mathbf{x})$, where

$$\mathrm{f} : \mathrm{X} \longrightarrow \mathrm{Y}, \tag{2.1}$$

is a function from a normed vector space X to a normed vector space Y. Here X represents the input to the problem (the data), f the problem itself, and Y its solution. Suppose we are interested in the effects on $\mathrm{y} \in \mathrm{Y}$ when a given $\mathrm{x} \in \mathrm{X}$ is perturbed slightly by a *small* amount δx, then the relative size of the perturbation in x is $\dfrac{|\delta\mathrm{x}|}{|\mathrm{x}|}$, and it's corresponding relative size of the perturbation in f(x) can be written as

$$\frac{|\mathrm{f}(\mathrm{x}+\delta\mathrm{x})-\mathrm{f}(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|} \approx \frac{|\delta\mathrm{x}\ \mathrm{f}'(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|} = \frac{|\mathrm{x}\ \mathrm{f}'(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|} \times \frac{|\delta\mathrm{x}|}{|\mathrm{x}|}.$$

The quantity

$$\kappa = \frac{|\ \mathrm{x}\ \mathrm{f}'(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|} \tag{2.2}$$

is called the **condition number** for the problem. If the quantity is large ($\kappa >> 1$), the problem is **ill-conditioned**; on the other hand , if it is small ($\kappa \approx 1$) , the problem is **well-conditioned**.

Example 2.1. Consider the problem of computing $\mathrm{f}(\mathrm{x}) = \sqrt{\mathrm{x}}$ for $\mathrm{x} > 0$.

Using the relation for a condition number of a problem defined above,

$$\text{we have} \quad \kappa = \frac{|\mathrm{x}\ \mathrm{f}'(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|} = \frac{1/(2\sqrt{\mathrm{x}})}{\sqrt{\mathrm{x}}/\mathrm{x}} = \frac{1}{2} < 1\ .$$

This problem is well-conditioned, since the condition number is very small.

Conditioning of a Matrix-Vector Multiplication

Consider the case where the function f in Equation 2.2 is a linear function, that is $\mathrm{f} : \mathrm{x} \longrightarrow \mathrm{Ax}$. From 2.2

$$\kappa = \frac{|\ \mathrm{x}\ \mathrm{f}'(\mathrm{x})|}{|\mathrm{f}(\mathrm{x})|}\ ,$$

$$\kappa = \lim_{\delta \mathrm{x} \to 0} \left[\frac{\|\mathrm{A}(\mathrm{x} + \delta \mathrm{x}) - \mathrm{Ax}\|}{\|\mathrm{Ax}\|} \Big/ \frac{\|\delta \mathrm{x}\|}{\|\mathrm{x}\|} \right]. \tag{2.3}$$

Using Taylor's expansion, $\mathrm{A}(\mathrm{x} + \delta \mathrm{x})$ can be simplify as:

$$\mathrm{A}(\mathrm{x} + \delta \mathrm{x}) \cong \mathrm{Ax} + \mathrm{A}\delta \mathrm{x} + 0(\delta \mathrm{x}^2) \cong \mathrm{Ax} + \mathrm{A}\delta \mathrm{x}. \tag{2.4}$$

By substituting 2.4 into 2.3, we obtain

$$\kappa = \lim_{\delta \mathrm{x} \to 0} \left[\frac{\|\mathrm{A}\delta \mathrm{x}\|}{\|\delta \mathrm{x}\|} \Big/ \frac{\|\mathrm{Ax}\|}{\|\mathrm{x}\|} \right] = \|\mathrm{A}\| \frac{\|\mathrm{x}\|}{\|\mathrm{Ax}\|}. \tag{2.5}$$

But

$$\|\mathrm{x}\| = \|\mathrm{Ix}\| = \|\mathrm{AA}^{-1}\mathrm{x}\| = \|\mathrm{A}^{-1}\mathrm{Ax}\| \leq \|\mathrm{A}^{-1}\|\|\mathrm{Ax}\|.$$

That is

$$\|\mathrm{x}\| \leq \|\mathrm{A}^{-1}\|\|\mathrm{Ax}\|\ ,$$

or further

$$\frac{\|\mathrm{x}\|}{\|\mathrm{Ax}\|} \leq \|\mathrm{A}^{-1}\|. \tag{2.6}$$

Combining Equations 2.6 and 2.5 we obtain

$$\kappa \leq \|\mathrm{A}\|\|\mathrm{A}^{-1}\|.$$

But for certain value of α, it can be deduced that

$$\kappa = \alpha\|\mathrm{A}\|\|\mathrm{A}^{-1}\| \ ,$$

where α is a proportionality constant.

Condition Number of a Matrix

Definition 2.1. Let **A** be an n × n non-singular matrix, then the condition number of **A** denoted by $\kappa(A)$ is defined as the product of the norm of **A** and the norm of the inverse of **A**. That is

$$\kappa(A) = \|A\|\|A^{-1}\|.$$

Here the assumed value of α is one. For any n × n non-singular matrix **A** and the natural norm $\|\cdot\|$,

$$1 \ = \ \|I_n\| \ = \ \|\mathbf{A}\cdot\mathbf{A}^{-1}\| \ \leq \ \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \ = \ \kappa(\mathbf{A}).$$

If $\kappa(\boldsymbol{A})$ is *small* (close to 1) then the matrix is said to be **well-conditioned**. On the other hand, if $\kappa(\mathbf{A})$ is **large**, that is, if it is significantly larger than one, then it is said to be **ill-conditioned**.

Example 2.2. Consider the linear system $\mathbf{Ax} = \mathbf{b}$ with A given by

$$A = \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix}.$$

Using infinity norm, the condition number $\kappa(A) = \|A\|\|A^{-1}\| = 5.0001\times10^4$. In this example, $\kappa(A)$ is large since $\kappa(A) >> 1$. Hence, A is ill-conditioned.

Existence and Uniqueness of Solutions of Linear Systems

The fundamental theorem of existence and uniqueness of solutions of linear systems answers the following questions: Does a solution exist, and is the solution unique. If at least one solution can be determined for a given problem, then a solution to that problem is said to exist. Most often, we try to prove the existence of solutions by means of the "so-called" existence theorem and then investigate their uniqueness by means of uniqueness theorem.

Definition 2.2. Consider the linear system

$$\mathbf{Ax} = \mathbf{b}. \tag{2.7}$$

If $\mathbf{A}$ is an m $\times$ n matrix and $\mathbf{b}$ a right hand vector with rank of $\mathbf{A}$ defined as the dimension of the largest square non-singular sub-matrix, then $\mathbf{x}$ is said to exist, if $\mathbf{A}$ is invertible or if m = n, and the rank of $\mathbf{A}$ is equal to n. Thus, if $\mathbf{x}$ exist then the linear system $\mathbf{Ax} = \mathbf{b}$ has a solution and the solution must be unique.

On the other hand, if A is not invertible or the rank of $\mathbf{A} \leq n$, then $\mathbf{x}$ cannot have a unique solution. The unique solution must remain valid through a given interval. If the solution is not unique or not valid for a given system of equations, there is the need to verify the error in the systems and correct it where possible.

Error Analysis of Linear Systems

Let $\mathbf{A}$ be an n $\times$ n nonsingular matrix, and let $\hat{\mathbf{x}}$ be the computed solution to the linear system

$$\mathbf{Ax} = \mathbf{b}. \tag{2.8}$$

The **error** vector is given by $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$. If $\|\cdot\|$ is a norm on $\mathbb{R}^n$, then $\|\mathbf{e}\|$ is a measure of the **absolute error**, and $\|\mathbf{e}\|/\|\mathbf{x}\|$ is a measure of the **relative error**.

Generally, we have no way of determining the exact value of $\|\mathbf{e}\|$ and $\|\mathbf{e}\|/\|\mathbf{x}\|$, since in most practical problems, the exact solution is not known. One way of testing the accuracy of the computed solution $\hat{\mathbf{x}}$ is to compute the **residual vector**, $\mathbf{r} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}}$ and see how small the **relative residual** $\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = \dfrac{\|\mathbf{b}-\mathbf{A}\hat{\mathbf{x}}\|}{\|\mathbf{b}\|}$ is. Unfortunately, a small residual does not always guarantee the accuracy of the solution, as the following example shows.

Example 2.3. Consider the linear system $\mathbf{Ax} = \mathbf{b}$ given by

$$\begin{bmatrix} 1 & 4 \\ 1.0001 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 5.0001 \end{bmatrix} .$$

If the exact solution of the system is $[1,\ 1]^T$, and asume its computed solution is $\hat{\mathbf{x}} = [5,\ 0]^T$, then residual vector is given by

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\hat{\mathbf{x}} = \begin{bmatrix} 5 \\ 5.0001 \end{bmatrix} - \begin{bmatrix} 1 & 4 \\ 1.0001 & 4 \end{bmatrix} \begin{bmatrix} 5 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -0.0004 \end{bmatrix} .$$

The relative residual is

$$\frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{\|-0.0004\|}{\|5.0001\|} = \frac{0.0004}{5.0001} = 7.99984 \times 10^{-4} .$$

This is very small even though the solution $\hat{\mathbf{x}} = [5,\ 0]^T$ is no where near the exact solution $\mathbf{x} = [1,\ 1]^T$. The above phenomena can be explained by the following Theorem.

Theorem 2.1 (The Residual Theorem)**.** Let $\mathbf{x}'$ be the computed solution to the linear system $\mathbf{Ax} = \mathbf{b}$. Then $\dfrac{\|\mathbf{x}' - \mathbf{x}\|}{\|\mathbf{x}\|} \ \leq\ \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$

Proof.

$$\text{From } \mathbf{r} = \mathbf{b} - \mathbf{Ax}' = \mathbf{Ax} - \mathbf{Ax}' = \mathbf{A}(\mathbf{x} - \mathbf{x}'),$$

we have

$$\mathbf{x} - \mathbf{x}' = \mathbf{A}^{-1}\mathbf{r} \qquad \text{(since } \mathbf{A} \text{ is nonsingular).}$$

Taking norms give,

$$\|\mathbf{x} - \mathbf{x}'\| \leq \|\mathbf{A}^{-1}\| \, \|\mathbf{r}\| \tag{2.9}$$

Also from $\mathbf{b} = \mathbf{A}\mathbf{x}$, we have $\|\mathbf{b}\| \leq \|\mathbf{A}\| \, \|\mathbf{x}\|$, that is,

$$\frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}.$$

This combines with Equation 2.9 to give

$$\frac{\|\mathbf{x} - \mathbf{x}'\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}. \tag{2.10}$$

The above Theorem tells that the relative error in the computed solution $\hat{\mathbf{x}}$ depends not only on the relative residual, but also on the quantity $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$. A computed solution can be guaranteed to be accurate only when the product $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|}$ is small.

Example 2.4. In Example 2.2, $\kappa(\mathbf{A}) = 5.0001 \times 10^4$ and the relative residual

$\|\mathbf{r}\|/\|\mathbf{b}\| = 0.000066664$. The inequality in 2.10 becomes
$\dfrac{\|\mathbf{x} - \mathbf{x}'\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\dfrac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = (5.0001 \times 10^4) \cdot (0.000066664) = 3.3333.$

Example 2.5. Consider the system of equations $\mathbf{A}\mathbf{x} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 2.02 \\ 1.98 \end{bmatrix}.$$

It is obvious that the exact solution $\mathbf{x}$, of the system is $[2,\ 0]^T$. Suppose we perturb the right hand vector $\mathbf{b}$ slightly to $\mathbf{b}' = [2,\ 2]^T$, the linear system $\mathbf{A}\mathbf{x}' = \mathbf{b}'$ has solution $\mathbf{x}' = [1,\ 1]^T$. The relative error in $\mathbf{b}$ using the infinity norm is given by

$$\frac{\|\mathbf{b} - \mathbf{b}'\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{0.02}{2.02} = 0.01 \ .$$

The relative error in the solution is also given by

$$\frac{\|\mathbf{x}-\mathbf{x}'\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{1}{1} = 1 \ ,$$

where

$$\|\mathbf{A}\|_\infty = 2, \|\mathbf{A}^{-1}\|_\infty = 0.50, \|\mathbf{r}\|_\infty = 0.02, \text{ and } \|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_\infty = 1.0 \ .$$

The residual vector is

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}' = \begin{bmatrix} 2.02 \\ 1.98 \end{bmatrix} - \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.02 \\ -0.02 \end{bmatrix} \ ,$$

and its relative residual is given by

$$\frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{0.02}{2.02} = 0.01 \ .$$

Hence the inequality

$$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\|\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|} = (1.0)(0.01) = 0.01 \ .$$

From above, the value of the relative error in $\mathbf{x}$ is much more smaller than the value of the relative residual. This indicates that the computed solution is accurate and close to the exact solution.

Ill- Conditioning in Linear Systems

Consider the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, it is observed that some linear systems give *good* solutions even under round-off(scaling) or co-efficient inaccuracies, whereas others give *bad* solutions under round-off. These inaccuracies affect the solution strongly. The extent to which it affects the solution is very important to this study.

Generally, if the coefficient matrix $\mathbf{A}$ is **ill-conditioned**, the relative residual may be much smaller than the relative error. On the other hand, if a matrix is **well-conditioned**, the relative residual and the relative error will be very close.

Definition 2.3. The linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a matrix and $\mathbf{b}$ is a right-hand vector is said to be **ill-conditioned** if a small change in the entries of matrix $\mathbf{A}$ or a small change in the right-hand vector $\mathbf{b}$ results in a large change in the vector solution of $\mathbf{x}$

Example 2.6. Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 1.1353 & 0.1859 \\ 0.7237 & 0.1185 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 1.3212 \\ 0.8422 \end{bmatrix}.$$

The exact solution of the system is $\mathbf{x} = [1, 1]^T$.

If $\mathbf{b}$ is perturbed slightly to $\mathbf{b}' = \begin{bmatrix} 1.3210 \\ 0.8420 \end{bmatrix}$, the system $\mathbf{Ax}' = \mathbf{b}'$ has solution $\mathbf{x}' = [-3.8489,\ 30.6115]^T$.

The residual vector $\mathbf{r} = \mathbf{b} - \mathbf{Ax}' = [0.0002, 0.0002]^T$, and the relative residual using the infinity norm is

$$\frac{\|\mathbf{r}\|_\infty}{\|\mathbf{b}\|_\infty} = \frac{0.0002}{1.3212} = 0.00015\ .$$

The relative error in the solution using infinity norm is given by

$$\frac{\|\mathbf{e}\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{\|\mathbf{x} - \mathbf{x}'\|_\infty}{\|\mathbf{x}\|_\infty} = \frac{29.6115}{1} = 29.6115$$

Hence, the relative error is more than 190000 times the relative residual. Also the condition number $\kappa(\mathbf{A}) = 1.244559e^{10}$, is large indicating that the matrix is ill-conditioned.

Perturbation Analysis of Linear Systems

Consider the linear system $\mathbf{Ax} = \mathbf{b}$. The entries of the coefficient matrix $\mathbf{A}$ and the right hand vector $\mathbf{b}$ of the linear system are assumed accurately represented. In practice, the entries contains small errors due to limitations in the accuracy of the data. Even if there are no errors in either of $\mathbf{A}$ or $\mathbf{b}$, round-off errors will occur when their entries are translated into the finite

precision number system of the computer. Thus, we generally expect that the coefficient matrix and the right hand vector will contain some errors. The system the computer solves is a slightly perturbed version of the original system. If the system is very sensitive, the solution could differ greatly from the solution of the perturbed system.

Example 2.7. Consider the following linear system

$$\begin{aligned} x_1 \;+\; 3x_2 &= 4 \\ 2x_1 \;+\; 5.999x_2 &= 7.999\ , \end{aligned} \tag{2.11}$$

has the exact solution $\mathbf{x} = [1,\ 1]^T$. Now, perturb the right-hand side to obtain the system

$$\begin{aligned} x_1 \;+\; 3x_2 &= 4 \\ 2x_1 \;+\; 5.999x_2 &= 8\ . \end{aligned} \tag{2.12}$$

Then, the solution to $\mathbf{Ax'} = \mathbf{b'}$ with $\mathbf{b'} =]4,\ 8]^T$, using Gaussian elimination with partial pivoting (considered to be a stable algorithm) is $\mathbf{x'} = [4,\ 0]^T$, which is no where near the true solution $\mathbf{x} = [1,\ 1]^T$.

Example 2.8. Consider the linear system

$$\begin{aligned} x_1 \;-\; x_2 &= 1 \\ x_1 \;-\; 1.01x_2 &= 0\ , \end{aligned} \tag{2.13}$$

which has the exact solution $\mathrm{x}_1 = 101, \quad$ and $\mathrm{x}_2 = 100$. Now, the slightly modified system

$$\begin{aligned} x_1 \;-\; x_2 &= 1 \\ x_1 \;-\; 0.99x_2 &= 0\ , \end{aligned} \tag{2.14}$$

has exact solution $\mathrm{x}_1 = -99 \quad$ and $\mathrm{x}_2 = -100$. The systems in 2.13 and 2.14 agree except for the small variation in the entry a_{22} of the matrix $\mathbf{A}$, but the solutions are completely different.

Example 2.9. Consider the system $A\mathbf{x} = \mathbf{b}$ where,

$$\mathbf{A} = \begin{bmatrix} 3 & 0.999 \\ 3.999 & 2.002 \end{bmatrix} \quad and \ \mathbf{b} = \begin{bmatrix} 3.999 \\ 6.001 \end{bmatrix}$$

The exact solution to $A\mathbf{x} = \mathbf{b}$ is $\mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. Now let $\mathbf{b}' = \mathbf{b} + \delta\mathbf{b} = \begin{bmatrix} 4 \\ 6 \end{bmatrix}$.

The solution to $A\mathbf{x}' = \mathbf{b}'$ is $\mathbf{x}' = \begin{bmatrix} 1.00149 \\ 0.99652 \end{bmatrix} \approx \mathbf{x} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$.

Definition 2.4. A linear system $\mathbf{Ax} = \mathbf{b}$ is said to be **ill-conditioned** if relatively small changes in the entries of the augmented matrix $[\mathbf{A} \ : \mathbf{b}]$ causes relatively large changes in the solution. $\mathbf{A}$ is said to be **well-conditioned** if relatively small changes in the entries of $\mathbf{A}$ and/or $\mathbf{b}$ results in relatively small changes in the solutions to $\mathbf{Ax} = \mathbf{b}$.

If the system is ill-conditioned, the computed solution to $\mathbf{Ax} = \mathbf{b}$ will generally not be very accurate. Even if the entries of $\mathbf{A}$ can be represented exactly as floating-point numbers, small round-off errors occurring in the reduction process can have very drastic effect on the computed solution. The systems in Example 2.7 and 2.8 are ill-conditioned systems. On the other hand, the system in Equation 2.9 is a well-conditioned system.

Effect of Perturbation in the Right-Hand Vector b

Consider the linear system $\mathbf{Ax} = \mathbf{b}$, suppose the error in the computed solution of the system when the data is perturbed slightly, are error in $\mathbf{b}$ but not the matrix $\mathbf{A}$, then we can assess the degree of error in the system, by the theorem below.

Theorem 2.2. (Right Perturbation Theorem) Let $\delta\mathbf{b}$ be the perturbation in $\mathbf{b}$ and $\delta\mathbf{x}$ be the resulting perturbation in the solution $\mathbf{x}$ of the linear

system $\mathbf{Ax} = \mathbf{b}$. If $\mathbf{A}$ is assumed to be non-singular and $\mathbf{b} \neq 0$, then

$$\frac{1}{\|\kappa(\mathbf{A})\|} \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \ .$$

Proof. Let

$$\mathbf{Ax} = \mathbf{b} \ , \tag{2.15}$$

then for a small perturbation in $\mathbf{x}$, there exist a corresponding perturbation in $\mathbf{b}$, such that $\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}$.

Using Taylor expansion and ignoring $0(\delta\mathbf{x}^2)$, we obtain

$$\mathbf{A}\delta\mathbf{x} = \delta\mathbf{b} \quad \text{or} \quad \delta\mathbf{x} = \mathbf{A}^{-1}\delta\mathbf{b}.$$

Taking norm of equation above, we have

$$\|\delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{b}\|. \tag{2.16}$$

Also from Equation 2.15, we have

$$\|\mathbf{b}\| \leq \|\mathbf{A}\|\|\mathbf{x}\|. \tag{2.17}$$

Using the Equations 2.16 and 2.17 above, we obtain

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \ . \tag{2.18}$$

On the other hand, from $\mathbf{A}\delta\mathbf{x} = \delta\mathbf{b}$ we get

$$\|\delta\mathbf{x}\| \geq \frac{\|\delta\mathbf{b}\|}{\|\mathbf{A}\|} \ . \tag{2.19}$$

Similarly, from $\mathbf{Ax} = \mathbf{b}$, it can be deduced that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, from which we get

$$\|\mathbf{A}^{-1}\|\|\mathbf{b}\| \geq \|\mathbf{x}\| \ . \tag{2.20}$$

Combining the Equations 2.19 and 2.20 above, we obtain

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \geq \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \times \frac{1}{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|}. \tag{2.21}$$

From the Equations 2.18 and 2.21 above we get the result

$$\frac{1}{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|} \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{2.22}$$

But

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|.$$

Therefore,

$$\frac{1}{\kappa(\mathbf{A})} \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$$

Remark 2.1. If $\kappa(\mathbf{A})$ is large, then a small perturbation in $\mathbf{b}$ can change the solution drastically, making the problem an ill-conditioned one. On the other hand, if $\kappa(\mathbf{A})$ is small (close to 1), then a small perturbation in $\mathbf{b}$ will have little effect on the solution. This is said to be well-conditioned.

Example 2.10. Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where

$$\mathbf{A} = \begin{bmatrix} 1.01 & 0.99 \\ 0.99 & 1.01 \end{bmatrix} \text{ and } \mathbf{b} = \begin{bmatrix} 2.0 \\ 2.0 \end{bmatrix}.$$

The exact solution $\mathbf{x}$ of the system $\mathbf{Ax} = \mathbf{b}$ is $[1,\ 1]^T$.

However, if $\mathbf{b}$ is perturbed slightly to $\mathbf{b}' = [2.02, 1.98]^T$, it has solution $\mathbf{x}' = [2,\ 0]^T$. The relative error in both $\mathbf{b}$ and $\mathbf{x}$ are respectively

$$\frac{\|\mathbf{b} - \mathbf{b}'\|_\infty}{\|\mathbf{b}\|_\infty} = 0.1 \tag{2.23}$$

and

$$\frac{\|\mathbf{x} - \mathbf{x}'\|_\infty}{\|\mathbf{x}\|_\infty} = 1\ . \tag{2.24}$$

Thus using the infinity norm on $\mathbf{A}$, we have

$$\|\mathbf{A}\|_\infty = 2.0\ ,\ \|\mathbf{A}^{-1}\|_\infty = 50\ ,$$

and

$$\|\mathbf{A}\|_\infty\|\mathbf{A}^{-1}\|_\infty = [2.0][50] = 100. \tag{2.25}$$

Hence by the right perturbation theorem we find that

$$\frac{1}{\|\mathbf{A}\|\|\mathbf{A}^{-1}\|} \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \leq \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| \times \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \ . \tag{2.26}$$

Substituting Equations 2.23, 2.24, and 2.25 , into 2.26 , we obtain

$$0.0001 \leq \frac{\|\mathbf{x} - \mathbf{x}'\|}{\|\mathbf{x}\|} \leq 10.$$

Remark 2.2. The actual relative change in $\mathbf{x}$ is equal to 1, which lies in the above interval. However, the relative error is bounded above by 10, but there is no precision in the accuracy of the computed solution.

Effect of Perturbation in the Matrix A

Consider the linear system $\mathbf{Ax} = \mathbf{b}$, suppose there are errors in the matrix $\mathbf{A}$ only and that $\mathbf{b}$ is exact, then we can assess the degree of error in the system by the theorem below.

Theorem 2.3. (Left Perturbation Theorem) Let $\delta\mathbf{A}$ be the perturbation in $\mathbf{A}$ and $\delta\mathbf{x}$ be the resulting perturbation in $\mathbf{x}$, in the solution of the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is non-singular and $\mathbf{b}$ is not equal to zero. Then

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A}) \times \dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \kappa(\mathbf{A}) \times \dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}} \ .$$

Proof. Given the linear system $\mathbf{Ax} = \mathbf{b}$, for small perturbations in $\mathbf{A}$, there is a corresponding perturbation in $\mathbf{x}$ such that $(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$. Subtracting $\mathbf{Ax} = \mathbf{b}$ from $(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$, we obtain

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}).$$

Taking norms of both side, we get

$$\|\delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|.(\|\mathbf{x}\| + \|\delta\mathbf{x}\|) = \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}.(\|\mathbf{x}\| + \|\delta\mathbf{x}\|),$$

which can be simplified as

$$\left(1 - \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right)\|\delta\mathbf{x}\| \leq \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\|\mathbf{x}\|.$$

Dividing through by the expression in parentheses we have

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}},$$

or further

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{\kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}.$$

Remark 2.3. The denominator of the equation above is less than 1. Thus even if $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ is small, there could be a drastic change in the solution if $\kappa(\mathbf{A})$ is large.

Example 2.11. Consider the linear systems of the equations below:

$$2.0012x_1 + 2.0000x_2 = 12.0060$$

$$1.0000x_1 + 1.0000x_2 = 6.0000$$

If we solve the system above using five digit decimal floating point arithmetic, the exact solution is $\mathbf{x} = [5,\ 1]^T$. On the other hand if we perturb the digit to four digit floating point arithmetic, the system of equations changes to

$$2.001x_1 + 2.000x_2 = 12.006 \text{ and}$$

$$1.000x_1 + 1.000x_2 = 6.000.$$

The computed solution then becomes $\mathbf{x}' = [6,\ 0]^T$ and the relative error in the solution is

$$\frac{\|\mathbf{x}' - \mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = 0.2.$$

The coefficient matrices $\mathbf{A}$ and $\mathbf{A}'$ of the systems are respectively

$$\mathbf{A} = \begin{bmatrix} 2.0012 & 2.0000 \\ 1.0000 & 1.0000 \end{bmatrix} \text{ and } \mathbf{A}' = \begin{bmatrix} 2.001 & 2.000 \\ 1.000 & 1,000 \end{bmatrix}.$$

Hence

$$\delta\mathbf{A} = \mathbf{A}' - \mathbf{A} = \begin{bmatrix} -0.0002 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}.$$

The computed $\mathbf{A}^{-1}$ of $\mathbf{A}$ is

$$\mathbf{A}^{-1} = \begin{bmatrix} 833\frac{1}{3} & -1666\frac{2}{3} \\ -833\frac{1}{3} & 1667\frac{2}{3} \end{bmatrix},$$

and its $\|\mathbf{A}^{-1}\|_\infty = 2501$.

Also

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|_\infty \|\mathbf{A}^{-1}\|_\infty = (2501)(4.0012) = 10007.0012, \|\delta\mathbf{A}\|_\infty = 0.0002,$$

with

$$\|\mathbf{A}\|_\infty = 4.0012$$

and

$$\frac{\|\delta\mathbf{A}\|_\infty}{\|\mathbf{A}\|_\infty} = 0.0000499850.$$

Thus

$$\|\delta\mathbf{A}\|_\infty \|\mathbf{A}^{-1}\|_\infty = (0.0002)(4.0012) = 0.0008,$$

which is less than one. Hence using the left perturbation theorem, that is

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}},$$

we have

$$\frac{\|\mathbf{x} - \hat{\mathbf{x}}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}{1 - \kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}} = \frac{0.5002}{1 - 0.5002} = 1.0.$$

Effect of perturbation in the Matrix A and the Vector b

Here, we assume there are errors in both the matrix $\mathbf{A}$ and the right-hand vector $\mathbf{b}$. Using the left and right perturbation theorems, we can establish the theorem below.

Theorem 2.4. Let $\delta\mathbf{A}$ and $\delta\mathbf{b}$ be small perturbation in $\mathbf{A}$ and $\mathbf{b}$ respectively, and let $\delta\mathbf{x}$ be the resulting perturbation in $\mathbf{x}$. Assume that $\mathbf{A}$ is nonsingular, $\mathbf{b} \neq 0$ and

$$\|\delta\mathbf{A}\| < \frac{1}{\|\mathbf{A}^{-1}\|},$$

then

$$\frac{\|\delta\mathbf{x}\|}{\|x\|} \leq \left(\frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}\right)\left(\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right).$$

Proof. Let

$$\mathbf{A}\mathbf{x} = \mathbf{b}. \tag{2.27}$$

For small perturbation in $\mathbf{A}$ and $\mathbf{b}$, we have

$$(\mathbf{A} + \delta\mathbf{A})(x + \delta\mathbf{x}) = (\mathbf{b} + \delta\mathbf{b}). \tag{2.28}$$

Subtracting Equation 2.27 from 2.28, we get

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

which can further be simplified as

$$\mathbf{A}[\mathrm{I} + \mathbf{A}^{-1}(\delta\mathbf{A})]\delta\mathbf{x} + (\delta\mathbf{A})\mathbf{x} = \delta\mathbf{b},$$

or

$$\mathbf{A}[\mathrm{I} - \mathbf{A}^{-1}(-\delta\mathbf{A})]\delta\mathbf{x} = \delta\mathbf{b} - (\delta\mathbf{A})\mathbf{x}. \tag{2.29}$$

Now, suppose we let $G = \mathbf{A}^{-1}(-\delta\mathbf{A})$. Then we can say that

$$\|\mathrm{G}\| = \|\mathbf{A}^{-1}(-\delta\mathbf{A})\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\| < 1.$$

Since $\|\mathrm{G}\| < 1$, it implies G is invertible and

$$\|(\mathrm{I}-\mathrm{G})^{-1}\| \leq \frac{1}{1-\|\mathrm{G}\|}, \quad \text{with} \quad \|\mathrm{I}\| = 1.$$

From 2.29, we have

$$\mathbf{A}[\mathrm{I}-\mathrm{G}]\delta\mathbf{x} = \delta\mathbf{b} - (\delta\mathbf{A})\mathbf{x},$$

or

$$\delta\mathbf{x} = (\mathrm{I}-\mathrm{G})^{-1}\mathbf{A}^{-1}(\delta\mathbf{b} - \delta\mathbf{A}\mathbf{x}). \tag{2.30}$$

If we take the norm of both sides of Equation 2.30, and simplify, we obtain

$$\|\delta\mathbf{x}\| \leq \frac{\|\mathbf{A}^{-1}}{1-\|\mathrm{G}\|}\left(\|\delta\mathbf{b}\| + \|\delta\mathbf{A}\|\|\mathbf{x}\|\right) ,$$

or further

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}}{1-\|\mathrm{G}\|}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{x}\|} + \|\delta\mathbf{A}\|\right) .$$

Using the fact that $\frac{1}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}$, the inequality above becomes

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}}{1-\|\mathrm{G}\|}\left(\frac{\|\delta\mathbf{b}\|\|\mathbf{A}\|}{\|\mathbf{b}\|} + \|\delta\mathbf{A}\|\right).$$

Upon simplification, it gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|}{1-\|\mathrm{G}\|}\left(\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right). \tag{2.31}$$

But,

$$\|\mathrm{G}\| = \|\mathbf{A}^{-1}(-\delta\mathbf{A})\| \leq \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\| = \frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

or

$$\|\mathrm{G}\| = \kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \tag{2.32}$$

Substituting Equation 2.31 into Equation 2.32, we get

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left(\frac{\|\mathbf{A}^{-1}\|\|\mathbf{A}\|}{1-\kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}\right)\left(\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right),$$

further

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left(\frac{\kappa(\mathbf{A})}{1-\kappa(\mathbf{A})\dfrac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}\right)\left(\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} + \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}\right). \tag{2.33}$$

Remark 2.4. The inequality 2.33 stipulate that if the matrix $\mathbf{A}$ is **well-conditioned**, *small* changes in $\mathbf{A}$ and $\mathbf{b}$ will produce a corresponding *small* changes in the solution $\mathbf{x}$.

On the other hand, if the matrix is **ill-conditioned**, then small changes in $\mathbf{A}$ and $\mathbf{b}$ may produce a large change in $\mathbf{x}$. It must be noted that if the perturbations $\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}$ and $\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}$ are small, there will be a drastic change in the solution if $k(\mathbf{A})$ is large.

Corollary 2.1. Suppose the linear system $\mathbf{Ax} = \mathbf{b}$ has a relative error of 10^{-d}, then it can be inferred that the perturbation

$$\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} = \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = 10^{-d}.$$

Thus from Equation 2.33, we have

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \left[\frac{\kappa(\mathbf{A})}{1 - \kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}}\right] \times 2(10^{-d}). \tag{2.34}$$

If $\kappa(\mathbf{A}) \approx 10^k$, then the quantity $\kappa(\mathbf{A})\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}$ becomes very *small* and approximately equal to zero.

Hence from Equation 2.34 we get

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq 2 \times \kappa(\mathbf{A}) \times 10^{-\kappa} \approx 10^{\kappa-d} = 10^{-(d-\kappa)}.$$

This shows that if the entries of $\mathbf{A}$ and $\mathbf{b}$ are accurate to about d-significant digits, and if the condition number of $\mathbf{A}$ is approximately 10^k, then the computed solution $\mathbf{x}$ is accurate to about $(d - k)$ significant digits.

Examples of Ill-Conditioned Systems

Ill-conditioned systems have many applications. A typical one is seen in polynomial data fitting using **Vandermonde System** and in least squares polynomial approximation using **Hilbert Matrix**.

Polynomial Data Fitting:Vandermonde System

Let $\mathrm{P_n(x)}$ be an interpolating n^{th} degree polynomial with $(n+1)$ data points $(\mathrm{x_i}, \mathrm{y_i})$. The interpolating polynomial $\mathrm{P_n(x)}$ can be obtained by solving the $(\mathrm{n}+1)\times(\mathrm{n}+1)$ linear system

$\mathrm{y_i} = \mathrm{a_0} + \mathrm{a_1 x_i} + \mathrm{a_2 x_i^2} + \mathrm{a_3 x_i^3} + ... + \mathrm{a_n x_i^n}$, where $\mathrm{i} = 0, 1, ..., \mathrm{n}$ for $\mathrm{a_o}, \mathrm{a_1}, ..., \mathrm{a_n}$.

In matrix form we have

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}.$$

The coefficient matrix for the linear system above is called Vandermonde matrix and it is non-singular, since $\mathbf{x}_i's$ are distinct. Therefore the system has unique solution and the $a_i's$ can be solved uniquely. However the numerical solution for small n can be solved easily without any difficulty, but for large n, the system becomes increasingly difficult to solve, that is becomes increasingly ill-conditioned.

Example 2.12. Let the $\mathbf{x}_i's$ be the $(n+1)$ equally spaced points in the interval [0 ,1]. For example the Vandermonde matrix for $n = 5$ is

$$\mathbf{V_5} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & (1/5) & (1/5)^2 & (1/5)^3 & (1/5)^4 & (1/5)^5 \\ 1 & (2/5) & (2/5)^2 & (2/5)^3 & (2/5)^4 & (2/5)^5 \\ 1 & (3/5) & (3/5)^2 & (3/5)^3 & (3/5)^4 & (3/5)^5 \\ 1 & (4/5) & (4/5)^2 & (4/5)^3 & (4/5)^4 & (4/5)^5 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}.$$

The condition number for Vandermonde matrix of order five is given by $\mathrm{cond(V_5)} = 4.92\mathrm{e}^{+3}$. As n increases from five onward, the condition numbers

increases and the matrix becomes increasingly ill-conditioned. The table below shows the condition numbers of the Vandermonde matrix for $n = 5$ *to* 12.

Table 1: Condition Numbers of Vandermonde Matrices

n	Cond(V_n)
5	$4.9244 \times 10^{+3}$
6	$3.6061 \times 10^{+4}$
7	$2.6782 \times 10^{+5}$
8	$2.0094 \times 10^{+6}$
9	$1.5193 \times 10^{+7}$
10	$1.1558 \times 10^{+8}$
11	$8.8348 \times 10^{+8}$
12	$6.67806 \times 10^{+9}$

The Hilbert Matrix

The n × n matrix H_n with entries

$$h_{ij} = \frac{1}{i + j - 1}, \quad \text{where} \quad 1 \leq i \leq n, \text{ and } 1 \leq j \leq n$$

is called the **Hilbert Matrix** of order n.

The Hilbert matrix arises in least squares polynomial approximation of continuous functions on the interval [0, 1], using the basis $1, x, x^n ..., x^n$ for P^n. Suppose a continuous function f(x) is defined on the interval [0, 1] and is to be approximated by a polynomial of degree $(n - 1)$, then

$$P_{n-1}(x) = \sum_{i=1}^{n} (a_i x^{i-1}),$$

such that the error(E) defined as

$$E = \|P_{n-1} - f\|_2^2 = \int_0^1 (\sum_{i=1}^{n} a_i x^{i-1} - f(x))^2 dx$$

is minimized.

The coefficient a_i of the polynomial are determined by equating the partial derivative of E with respect to a_i to zero, to obtain the normal equation.

$$\frac{\partial \mathrm{E}}{\partial \mathrm{a_i}} = 2\int_0^1 (\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{a_i x^{j-1}} - \mathrm{f(x)})\mathrm{x^{i-1}dx},\ \mathrm{i} = 1, 2, ..., \mathrm{n}.$$

If $\dfrac{\partial \mathrm{E}}{\partial \mathrm{a_i}} = 0$, for all $i = 0, 1, ...n$,
we obtain

$$\sum_{\mathrm{j=1}}^{\mathrm{n}} (\mathrm{a_i}) \int_0^1 \mathrm{x^{i+j-2}dx} = \int_0^1 \mathrm{f(x)x^{i-1}dx},\ \mathrm{i} = 1, 2, ..., \mathrm{n}.$$

With

$$\mathrm{h_{ij}} = \int_0^1 \mathrm{x^{i+j-2}dx} \quad \text{and} \quad \mathrm{b_i} = \int_0^1 \mathrm{dx},\ (\mathrm{i} = 1, 2, ..., \mathrm{n}).$$

we obtain

$$\sum_{\mathrm{j=1}}^{\mathrm{n}} \mathrm{h_{ij} a_j} = \mathrm{b_i}, \quad \text{for} \quad \mathrm{i, j} = 1, 2, ..., \mathrm{n}.$$

This is exactly the linear system $\mathrm{H_n a} = \mathrm{b}$, given by

$$\begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix} \times \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix},$$

with

$$\mathrm{H_n} = [\mathrm{h_{ij}}],\ \mathrm{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} \text{ and } \mathrm{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}.$$

Thus the matrix $\mathrm{H} = [\mathrm{h_{ij}}]$ is easily identified as the Hilbert matrix since

$$\mathrm{h_{ij}} = \int_0^1 \mathrm{x^{i+j-2}dx} = \frac{1}{\mathrm{i+j-1}}.$$

Example 2.13. Consider the Hilbert matrix of order six H_6, defined below

$$\mathbf{H_6} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 \\ 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 \end{bmatrix},$$

with inverse H_6^{-1} given by

$$\mathbf{H_6^{-1}} = \begin{bmatrix} 36 & -630 & 3360 & -750 & 7560 & -2772 \\ -630 & 14700 & -88200 & 211680 & -220500 & 83160 \\ 3360 & -88200 & 564480 & -1411200 & 1512000 & -582120 \\ -7560 & 211680 & -1411200 & 3628800 & -396900 & 1552320 \\ 7560 & -220500 & 1512000 & -3969000 & 4410000 & -1746360 \\ -2772 & 83160 & -582120 & 1552320 & -1746360 & 698544 \end{bmatrix}.$$

It is observed that the entries of H_6^{-1} are large compared with the entries of H_6.

When the right-hand vector $\mathbf{b}$ is multiplied by H_6^{-1}, the entries become magnified.

The condition number of H_6 is $\text{cond}(H_6) = 1.495106e^{+7}$, which increases rapidly with n, thus making the matrix increasingly ill-conditioned.
The relation $k_\infty(H_n) = ce^{3.5n}$ can be used to determine the condition number of the Hilbert matrix. It increases rapidly with n, due to the ill-conditioned nature of the Hilbert matrix. The table below gives the condition numbers for $n = 5, 6...12$.

Table 2: Condition Numbers of Hilbert Matrices

n	cond(H_n)
5	$4.766072 \times 10^{+05}$
6	$1.495106 \times 10^{+07}$
7	$4.753674 \times 10^{+08}$
8	$1.525758 \times 10^{+10}$
9	$4.931544 \times 10^{+11}$
10	$1.602529 \times 10^{+13}$
11	$5.223945 \times 10^{+14}$
12	$1.794510 \times 10^{+16}$

Example 2.14. Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 12×12 Hilbert matrix, and $\mathbf{b}$ chosen such that the linear system has the exact solution $\mathbf{x}_i = i^2$, for i = 1 , 2 , 3 , ..., 12. If the system is solved, using the back slash method $\hat{\mathbf{x}} = \mathbf{A}/\mathbf{b}$, we realized that for small $\mathbf{n}$, the computed solution $\hat{\mathbf{x}}$ is reasonably accurate.

However as n increases the precision degenerates very rapidly and has no significance. The Table 3 below illustrate the computed values for relative error, condition number and the number of digits lost for n = 5 up to 12.

We realize from Table 3 that the machine precision is about 16 digit of accuracy. This signifies that the exact solution of the system is accurate to about 16 significant digits. For $n = 5, 6, 7, 8, 9, 10$ and 11, we have respectively $11, 9, 8, 6, 5, 3$ and 2 digits of accuracy.

When $n \geq 12$, the computed solution $\hat{\mathbf{x}}$ has no significant digit of accuracy. This can be explain by the increasingly nature of the condition number of the matrix $\mathbf{A}$ from n = 5 to 12.

Table 3: Accuracy of a Hilbert Systems

n	$\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_\infty}{\|\mathbf{x}\|_\infty}$	Cond($\mathbf{A}$)	No. of Digits Lost
5	$7.135919719e^{-12}$	$4.766072502e^{+05}$	4
6	$1.540944388e^{-11}$	$1.495105864e^{+07}$	5
7	$3.016719939e^{-10}$	$4.753673562e^{+08}$	6
8	$4.623435736e^{-08}$	$1.525757542e^{+10}$	8
9	$5.409071818e^{-06}$	$4.931538214e^{+11}$	10
11	$1.80367381e^{-04}$	$1.602515828e^{+12}$	12
11	$6.28747466e^{-03}$	$5.221040338e^{+14}$	13
12	$7.42237815e^{-04}$	$1.794510255e^{+16}$	16

Examining Accuracy of Hilbert System Using Different Methods

Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 12×12 Hilbert matrix H_{12} with

$$\mathbf{b}_i = \sum_{j=1}^{n} (h_{ij}),$$

and has exact solution $\mathbf{x} = [1, 1, ..., 1]^T$. If the linear system $\mathbf{Ax} = \mathbf{b}$ is solved using standard methods for solving linear systems such as the QR-factorization(QR), LU-factorization(LU) and Choleskey factorization for positive index, we obtain various solutions, as shown below.

From Table 4, we observe that the relative error increases as n increases in all the methods. The number of digits lost in each case also increases relative to the error. For instance, the QR-factorization method is a typical case. Here, the errors increases with n from $E = 12$, to $E = 01$ respectively. The number of digits lost from n = 5 to 12 are $4, 6, 7, 8, 10, 12, 13$ and 15 respectively. The condition numbers (Cond(n)) in all the methods also show a lot of resemblance.

Table 4: Various Solutions of the Hilbert System

n	QR	Choleskey	LU	Cond(n)
5	$4.92922205e^{-12}$	$1.12608245e^{-11}$	1.17773406^{-11}	$4.76607250e^{+05}$
6	$2.75982022e^{-10}$	$1.20268228e^{-10}$	$2.18522312e^{-10}$	$1.49510584e^{+07}$
7	$7.79173750e^{-09}$	$1.28701202e^{-09}$	$1.85017132e^{-10}$	$4.75367356e^{+08}$
8	$5.11039341e^{-08}$	$1.04032552e^{-07}$	$7.39361953e^{-08}$	$1.52575754e^{+10}$
9	$4.03744625e^{-06}$	$4.50884240e^{-06}$	$3.76765864e^{-06}$	$4.93153832e^{+11}$
10	$2.07679375e^{-04}$	$6.71064175e^{-06}$	$8.85489699e^{-05}$	$1.60251582e^{+13}$
11	$1.30527625e^{-03}$	$2.45715014e^{-03}$	$1.06459286e^{-03}$	$5.22104933e^{+14}$
12	$1.41364200e^{-01}$	$3.67288989e^{-02}$	$1.73584356e^{-02}$	$1.79451025e^{+16}$

It can be inferred that if the condition number of a matrix $\mathbf{A}$ is to the order of 10^k, that is about k significant figures, and the entries of $\mathbf{A}$ and $\mathbf{b}$ are accurate to d-significant digits, then the computed solution $\hat{\mathbf{x}}$ is accurate to about $(d-k)$ significant digits. This confirms the corollary 2.1. Thus in effect, the errors in the computed solutions can be attributed to the ill-conditioned nature of the matrix but not the vector $\mathbf{b}$. It is therefore necessary to decompose the matrix to investigate the errors in the solution.

CHAPTER THREE

Singular Value Decomposition (SVD)

The **singular value decomposition** (SVD) is perhaps, the most numerically effective approach for solving linear systems, and least-squares problems. The decomposition also enables us to determine the condition number of a matrix. Thus, every square symmetric matrix $\mathbf{A}$ can be factored as $\mathbf{A} = \mathbf{PDP}^T$, where $\mathbf{P}$ is an orthogonal matrix and $\mathbf{D}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$. If $\mathbf{A}$ is not symmetric, then such a factorization is not always possible, but we may still be able to factorize $\mathbf{A}$ as $\mathbf{A} = \mathbf{PDP}^{-1}$, where $\mathbf{D}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$, and $\mathbf{P}$ is just an invertible matrix.

However, not every square matrix can be diagonalized, but every $\mathrm{m} \times \mathrm{n}$ matrix has a factorization of the form $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and Σ is an $\mathrm{m} \times \mathrm{n}$ "diagonal" matrix whose entries are the **singular values** of $\mathbf{A}$. This factorization is called the **singular value decomposition** of $\mathbf{A}$.

The Singular Values of a Matrix

To obtain the SVD of the matrix $\mathbf{A}$, we need the eigenvalues and the eigenvectors of the symmetric matrices $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$. These eigenvalues are non-negative. If we denote the eigenvalues of $\mathbf{A}$ by $\lambda_\mathrm{i} \geq 0,\ \mathrm{i} = 1, \cdots \mathrm{n}$, then their square roots, $\sqrt{\lambda_\mathrm{i}} = \sigma_\mathrm{i},\ \mathrm{i} = 1, \cdots, \mathrm{n}$ are called
the **singular values** of $\mathbf{A}$.

Theorem 3.1. *:* Let $\mathbf{A}$ be an m $\times$ n matrix with m $\geq$ n then

1. The eigenvalues of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are real and non-negative .

2. If λ is a non-zero eigenvalue of $\mathbf{A}^T\mathbf{A}$ corresponding to the eigenvector $\mathbf{x}$, then λ is also eigenvalue of $\mathbf{A}\mathbf{A}^T$ with corresponding eigenvector $\mathbf{A}\mathbf{x}$. In other words, $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have the same non-zero eigenvalues.

Proof. 1. $(\mathbf{A}^{\mathrm{T}}\mathbf{A})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}}\mathbf{A}$, and so $\mathbf{A}^T\mathbf{A}$ is symmetric. Similarly for $\mathbf{A}\mathbf{A}^T$

2. Let $\mathbf{x}$ be an eigenvector of $\mathbf{A}^T\mathbf{A}$ corresponding to a non-zero eigenvalue λ. Then

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{3.1}$$

Multiplying through equation above on the left by $\mathbf{x}^T$ yields

$$\mathbf{x}^T\mathbf{A}^T\mathbf{A}\mathbf{x} = \lambda\mathbf{x}^T\mathbf{x},$$

$$(\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{x}) = \lambda\|\mathbf{x}\|_2^2,$$

$$\|\mathbf{A}\mathbf{x}\|_2^2 = \lambda\|\mathbf{x}\|_2^2 \geq 0.$$

Hence

$$\lambda = \frac{\|\mathbf{A}\mathbf{x}\|_2^2}{\|\mathbf{x}\|_2^2} \geq 0 \ .$$

Similarly, applying the same steps to $\mathbf{A}\mathbf{A}^T$, it can be deduced that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ are real and non-zero.

3. Using the Equation 3.1 above, and pre-multiplying through the equation by A, we have

$$\mathbf{A}(\mathbf{A}^T\mathbf{A})\mathbf{x} = \mathbf{A}(\lambda\mathbf{x}).$$

Further $\mathbf{A}\mathbf{A}^T(\mathbf{A}\mathbf{x}) = \lambda(\mathbf{A}\mathbf{x})$. This implies that $\mathbf{A}\mathbf{x}$ is an eigenvector of $\mathbf{A}\mathbf{A}^T$ corresponding to the eigenvalue λ , where $\mathbf{x}$ is an eigenvector of $\mathbf{A}^T\mathbf{A}$ corresponding to the eigenvalue λ.

Example 3.1. Given the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix},$$

we obtain the eigenvalues of the matrices (i) $\mathbf{A}^T\mathbf{A}$ (ii) $\mathbf{A}\mathbf{A}^T$ as below

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix}.$$

For eigenvalues, we use

$$\det(\mathbf{A}^{\mathrm{T}}\mathbf{A} - \lambda \mathrm{I}) = 0.$$

That is

$$\begin{bmatrix} 9-\lambda & -9 \\ -9 & 9-\lambda \end{bmatrix} = 0,$$

or

$$(9-\lambda)(9-\lambda) - 81 = 0$$

$$\lambda^2 - 18\lambda = 0.$$

Solving we find that the eigenvalues of A^TA are $\lambda_1 = 0$ or $\lambda_2 = 18$.

Similarly,

$$\mathbf{A}\mathbf{A}^{\mathrm{T}} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} 1 & -2 & 2 \\ -1 & 2 & -2 \end{bmatrix} = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix}.$$

Hence, the determinant for $\mathbf{A}\mathbf{A}^{\mathrm{T}} = \det(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \lambda \mathrm{I})$ is

$$\begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix} = 0,$$

which implies

$$(9-\lambda)(9-\lambda) - 81 = 0$$

$$\lambda^2 - 18\lambda = 0$$

$$\lambda 1 = 0 \quad \text{or} \quad \lambda_2 = 18.$$

Remark 3.1. This confirms Theorem 1.1, which stipulate that $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$ have the same non-zero eigenvalues.

Construction of SVD of a Matrix

Given a matrix $\mathbf{A}$, we can construct the SVD of $\mathbf{A}$ by going through the following steps:

1. Find an orthogonal diagonalization of $\mathbf{A}^T\mathbf{A}$.

2. Set up V and Σ.

3. Construct U, the normalization vector obtained from $\mathbf{A}V_n$, where n = 1, 2, 3, ..., r.

Example 3.2. Find the singular value decomposition of the matrix $\mathbf{A}$ in Example 3.1.

From Example 3.1 above,

$$\mathbf{A}^T\mathbf{A} = \begin{bmatrix} 9 & -9 \\ -9 & 9 \end{bmatrix}.$$

The eigenvalues of A^TA are

$$\lambda_1 = 18 \text{ and } \lambda_2 = 0,$$

with corresponding unit eigenvectors

$$\mathrm{V}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} \quad \text{and} \quad \mathrm{V}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}.$$

These unit vectors form the columns of V as

$$\mathrm{V} = [\mathrm{v}_1, \mathrm{v}_2] = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}.$$

The singular values are $\sigma_1 = 3\sqrt{2} \quad \text{and} \quad \sigma_2 = 0$. Since there is only one non-zero singular value, the diagonal matrix D may be written as a single number in terms of σ_1. The matrix Σ has the same size as $\mathbf{A}$, with D in its upper left corner as shown below

$$\Sigma = \begin{bmatrix} D & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Next we construct U, but since U depends on $\mathbf{A}$V, it's necessary to find first $\mathbf{A}\mathrm{V}_1$ and $\mathbf{A}\mathrm{V}_2$. Thus

$$\mathbf{A}\mathrm{V}_1 = \begin{bmatrix} \frac{2}{\sqrt{2}} \\ \frac{-4}{\sqrt{2}} \\ \frac{4}{\sqrt{2}} \end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathrm{V}_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

But $\|\mathbf{A}\mathrm{V}_1\| = \sigma_1 = 3\sqrt{2}$ and $\|\mathbf{A}\mathrm{V}_2\| = \sigma_2 = 0$.

From $\mathbf{A}\mathrm{V}_\mathrm{i} = \sigma_\mathrm{i}\mathrm{U}_\mathrm{i}$, where $\mathrm{i} = 1, 2, ..., \mathrm{r}$, it can can deduced that

$$\mathrm{U}_1 = \frac{\mathbf{A}\mathrm{V}_1}{\sigma_1} = \frac{1}{3\sqrt{2}} \times \mathbf{A}\mathrm{V}_1 = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix}.$$

U_1 is the only column for U since the value of σ_2 is zero. But other column of U can be deduced by extending the set U_1 to an orthonormal basis for $\mathbb{R}^3$. The two orthonormal unit vectors of U_2 and U_3 which are orthogonal to U_1 are found such that $\mathrm{U}_1^{\mathrm{T}} = 0$, and have basis

$$\mathbf{x}_1 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathbf{x}_2 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}.$$

Applying the Gram-Schmidt process with normalization to $\mathbf{x}_1$ and $\mathbf{x}_2$, gives

$$\mathrm{U}_2 = \begin{bmatrix} 2/\sqrt{5} \\ 1/\sqrt{5} \\ 0 \end{bmatrix} \text{ and } \mathrm{U}_3 = \begin{bmatrix} -2/\sqrt{45} \\ 4\sqrt{45} \\ 5\sqrt{45} \end{bmatrix}.$$

If we set $\mathrm{U} = [\mathrm{U}_1, \mathrm{U}_2, \mathrm{U}_3]$, then the 3by3 orthogonal matrix is given by

$$\mathrm{U} = \begin{bmatrix} 1/3 & 2/\sqrt{5} & -2\sqrt{45} \\ -2/3 & 1\sqrt{5} & 4/\sqrt{45} \\ 2/3 & 0 & 5/\sqrt{45} \end{bmatrix}.$$

It can be verified that

$$\mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \begin{bmatrix} 1/3 & 2/\sqrt{5} & -2/\sqrt{45} \\ -2/\sqrt{3} & 1/\sqrt{5} & 4/\sqrt{45} \\ 2/3 & 0 & 5\sqrt{45} \end{bmatrix} \begin{bmatrix} 3/\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{} \end{bmatrix}.$$

$$\text{Therefore } \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} = \mathbf{A},$$

where U and V are the left and right singular vectors of $\mathbf{A}$ respectively.

The SVD and the Structure of a Matrix

The structure of a matrix and its associated properties such as the rank, the 2-norm, F-norm, the infinity-norm, the condition number, as well as the orthonormal basis for the null space and the range of a matrix can be determine using SVD.

Theorem 3.2. Let $\sigma_1 \geq \sigma_2 \geq, ..., \geq \sigma_{\mathrm{n}}$ be the singular values of an $\mathrm{m} \times \mathrm{n}$ matrix $\mathbf{A}$, then

1. $\|\mathbf{A}\|_2 = \sigma_1 = \sigma_{max}$.

2. If $\mathbf{A}$ is n $\times$ n and nonsingular, then $\|\mathbf{A}^{-1}\|_2 = \dfrac{1}{\sigma_n} = \dfrac{1}{\sigma_{min}}$.

3. $\mathrm{Cond}_2(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \dfrac{\sigma_1}{\sigma_\mathrm{n}} = \dfrac{\sigma_{\max}}{\sigma_{\min}}$.

4. $\|\mathbf{A}\|_\mathrm{F} = (\sigma_1^2 + \sigma_2^2 + ... + \sigma_\mathrm{n}^2)$.

Proof. :

Given that the SVD of $\mathbf{A} = \mathrm{U}\Sigma\mathrm{V}^\mathrm{T}$, since $\|.\|_2$ and $\|.\|_\mathrm{F}$ are invariant under orthogonal transformation, we have

1. $\|\mathbf{A}\|_2 = \|\mathrm{U}\Sigma\mathrm{V}^\mathrm{T}\|_2 = \|\Sigma\|_2 = \max\sigma_\mathrm{i} = \sigma_1$.

2. Since A is invertible, its smallest singular value $\sigma_n \neq 0$. From the SVD of $\mathbf{A}^{-1}$, the largest singular value of $\mathbf{A}^{-1}$ is $\dfrac{1}{\sigma_n}$. Hence $\|A^{-1}\|_2 = \dfrac{1}{\sigma_n}$.

3. The result follows from (1) and (2) .

4. $\mathrm{F} = (\sigma_1^2 + \sigma_2^2 + ... + \sigma_\mathrm{n}^2)^{1/2}$.

Remark 3.2. For an m $\times$ n matrix $\mathbf{A}$, we define $\mathrm{cond}(\mathbf{A}) = \dfrac{\sigma_{\max}}{\sigma_{\min}}$. However, when $\mathbf{A}$ is rank deficient, then $\sigma_{min} = 0$. In this case, we say that $\mathrm{cond}(\mathbf{A})$ is infinite.

The Linear Least Square Problem

The linear system $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is singular or not necessarily a square matrix have infinitely many solutions or the solution does not exist at all. For an m $\times$ n matrix $\mathbf{A}$, if m $>$ n, the linear system of equations is said to be an overdetermined linear system of m equations in n unknowns, and have no solution. An overdetermined system have the number of equations greater than the number of unknowns.

Similarly, if m $<$ n, the system is said to be under-determined and have infinite number of solutions. In situations where the linear system $\mathbf{Ax} = \mathbf{b}$

has no solution, one tries to find a value for $\mathbf{x}$ for which the residual vector $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$ is as small as possible.

If r = 0, it can be inferred that $\mathbf{A}\mathbf{x} = \mathbf{b}$, but for a general overdetermined system r $\neq 0$. To minimize $\|\mathbf{b} - \mathbf{A}\mathbf{x}\|$ for a vector solution $\mathbf{x}$, we make use of the Euclidean norm . If we choose the 2-norm, the problem can be stated as $\min\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2$ and its solution is called the **linear Least Square solution** of the overdetermined system $\mathbf{A}\mathbf{x} = \mathbf{b}$, Hansen(1987).

Thus, the least squares solution $\hat{\mathbf{x}}$, for the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is obtained by minimizing the 2-norm square of the residual of $\mathbf{A}\mathbf{x} = \mathbf{b}$. The 2-norm residual factor $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ can be expressed as

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\mathbf{x} - \mathbf{b})^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \tag{3.2}$$

and further be written as

$$\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 = (\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{x}) - \mathbf{b}^T\mathbf{A}\mathbf{x} - (\mathbf{A}\mathbf{x})^T\mathbf{b} - \mathbf{b}^T\mathbf{b} \tag{3.3}$$

The terms $\mathbf{b}^T(\mathbf{A}\mathbf{x})$ and $(\mathbf{A}\mathbf{x})^T\mathbf{b}$ are equal, and the derivative with respect to $\mathbf{x}$ at zero is minimum. Thus the equation reduce to

$$2\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} - 2\mathbf{A}^T\mathbf{b} = 0 \ ,$$

or

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}.$$

Hence the least square solution is

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

Definition 3.1. Let $\mathbf{A}$ be an m $\times$ n matrix and $\mathbf{b} \in \mathbb{R}^m$, then a solution $\mathbf{x} \in \mathbb{R}^n$ of the least square problem can only exist if the norm of the residual vector $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ is as small as possible and the solution of the least square problem with dimension at most n is

$$\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}.$$

Reduced Singular Value Decomposition and the Pseudoinverse

Let $\mathbf{A}$ be an m $\times$ n matrix, then the singular value decomposition of $\mathbf{A}$ is given by $\mathbf{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}}$. When Σ contains rows or columns of zeros, a more compact decomposition of $\mathbf{A}$ is possible.

Suppose rank($\mathbf{A}$) = r < m, then U and V can be partitioned into sub-matrices whose first blocks contain r columns as

$$\mathrm{U} = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \text{ and } \mathrm{V} = \begin{bmatrix} V_r & V_{n-r} \end{bmatrix},$$

where $\mathrm{U_r} = \mathrm{U}_1, \mathrm{U}_2, ..., \mathrm{U_r}$, $\mathrm{V_r} = \mathrm{V}_1, \mathrm{V}_2, ..., \mathrm{V_r}$ and $\mathrm{U_r}, \mathrm{V_r}$ are m $\times$ r and n $\times$ r sub-matrices respectively. The partition matrix multiplication of $\mathbf{A}$ becomes

$$\mathbf{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \begin{bmatrix} U_r & U_{m-r} \end{bmatrix} \begin{bmatrix} D_r & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_r^T \\ V_{n-r} \end{bmatrix} = \mathrm{U_r D_r V_r^T}. \tag{3.4}$$

A factorization of $\mathbf{A}$ in 3.4 above by Mathews(1992), is called a reduced singular value decomposition of $\mathbf{A}$. The diagonal entries of D_r are nonzero and D_r is invertible. The Pseudoinverse ($\mathbf{A}^+$) of reduced SVD of $\mathbf{A}$ is given by $\mathbf{A}^+ = \mathrm{V_r D_r^{-1} U_r^T}$.

Example 3.3. For example, consider the problem of finding the Pseudoinverse of the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

The eigenvalues of the matrix $\mathbf{A}^T\mathbf{A}$ above are $\lambda_1 = 18$ and $\lambda_2 = 0$, and its singular values are $\sigma_1 = \sqrt{18} = 3\sqrt{2}$, and $\sigma_2 = 0$ respectively. The corresponding eigenvectors are also given by

$$\mathrm{V}_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \end{bmatrix} \text{ and } \mathrm{V}_2 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}.$$

The 2×2 orthogonal matrix V and the 3×2 diagonal matrix Σ are given by

$$\mathrm{V} = [\mathrm{V}_1, \mathrm{V}_2] = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} ,$$

and

$$\Sigma = \begin{bmatrix} 3\sqrt{2} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} .$$

To construct the 3×3 orthogonal matrix U, we first construct

$$\mathrm{U}_1 = \frac{1}{\sigma_1} \mathbf{A}\mathrm{V}_1 = \frac{1}{3\sqrt{2}} \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix} .$$

The second and third columns of U are obtained from the orthogonal complement of U_1, by applying the Gram-Schmidt process to the linearly independent set

$$\mathrm{U}_1 = \begin{bmatrix} 1/3 \\ -2/3 \\ 2/3 \end{bmatrix} \text{ with basis } \mathrm{e}_2 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix} \text{ and } \mathrm{e}_3 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} ,$$

to give

$$\mathrm{U}_2 = \begin{bmatrix} \frac{2}{\sqrt{5}} \\ \frac{1}{\sqrt{5}} \\ 0 \end{bmatrix} \text{ and } \mathrm{U}_3 = \begin{bmatrix} \frac{-6}{\sqrt{5}} \\ \frac{12}{\sqrt{5}} \\ \frac{15}{\sqrt{5}} \end{bmatrix} \text{ respectively.}$$

Thus

$$\mathrm{U} = [\mathrm{U}_1, \mathrm{U}_2, \mathrm{U}_3] = \begin{bmatrix} 1/3 & 2/\sqrt{5} & -6/\sqrt{5} \\ -2/3 & 1/\sqrt{5} & 12/\sqrt{5} \\ 2/3 & 0 & 15/\sqrt{5} \end{bmatrix} .$$

With rank $\mathbf{A} = r = 1$, the reduced SVD of

$$\mathbf{A} = \mathrm{U_r D_r V_r^T} = \mathrm{U_1 D_1 V_1^T} = \begin{bmatrix} 1/3 \\ -2/3 \\ -2/3 \end{bmatrix} [3\sqrt{2}] \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{2} \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -2 & 2 \\ 2 & -2 \end{bmatrix}.$$

The Pseudoinverse of $\mathbf{A}$ for $r = 1$ is

$$\mathbf{A}^+ = \mathrm{V_r D_r^{-1} U_r^T} = \mathrm{V_1 D_1 U_1^T} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{3\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1/3 & -2/3 & 2/3 \end{bmatrix}.$$

Hence,

$$\mathbf{A}^+ = \begin{bmatrix} 1/18 & -1/9 & 1/9 \\ -1/18 & 1/9 & -1/9 \end{bmatrix}.$$

Singular Value Decomposition and Linear Systems

Consider the least squares solution of overdetermined linear system of equation $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{\mathrm{m \times n}},\ \mathbf{x} \in \mathbb{R}^{\mathrm{n}},\ \mathbf{b} \in \mathbb{R}^{\mathrm{m}},$ and $\mathrm{m} \geq \mathrm{n}$. Suppose the singular values are arranged such that $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\mathrm{n}} > 0$ and the rank of A is equal to n, then the least-square solution $\hat{\mathbf{x}}$ to the problem $\mathbf{Ax} = \mathbf{b}$ is given by $\hat{\mathbf{x}} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}$, or

$$\mathbf{A}^T\mathbf{A}\hat{\mathbf{x}} = \mathbf{A}^T\mathbf{b}. \tag{3.5}$$

But the singular value decomposition of $\mathbf{A} = \mathrm{U\Sigma V^T}$, where U is an $\mathrm{m \times m}$ orthogonal matrix, V is an $n \times n$ orthogonal matrix, and Σ is an $\mathrm{n \times n}$ diagonal matrix, whose diagonal entries are the singular values defined by $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{\mathrm{n}} > 0$. Substituting $\mathbf{A} = \mathrm{U\Sigma V^T}$ into Equation 3.5 we get

$$(\mathrm{U\Sigma V^T})^{\mathrm{T}}\mathrm{U\Sigma V^T}\hat{\mathbf{x}} = \mathrm{U\Sigma V^T})^{\mathrm{T}}\mathbf{b},$$

$$\mathrm{V\Sigma^T U^T U\Sigma V^T}\hat{\mathbf{x}} = \mathrm{V\Sigma^T U^T}\mathbf{b}.$$

By orthogonality,

$$\mathrm{U^T U} = \mathrm{I} \text{ and } \mathrm{VV^T} = \mathrm{I}.$$

Hence,

$$\mathrm{V}\Sigma^{\mathrm{T}}\Sigma\mathrm{V}^{\mathrm{T}}\hat{\mathbf{x}} = \mathrm{V}\Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathbf{b}.$$

If we pre-multiple both side by V^T, we obtain

$$\mathrm{V}^{\mathrm{T}}\mathrm{V}\Sigma^{\mathrm{T}}\Sigma\mathrm{V}^{\mathrm{T}}\hat{\mathbf{x}} = \mathrm{V}^{\mathrm{T}}\mathrm{V}\Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathbf{b} \ ,$$

$$\Sigma^{\mathrm{T}}\Sigma\mathrm{V}^{\mathrm{T}}\hat{\mathbf{x}} = \Sigma^{\mathrm{T}}\mathrm{U}^{\mathrm{T}}\mathbf{b}.$$

Dividing through by Σ^T, the equations become

$$\Sigma\mathrm{V}^{\mathrm{T}}\hat{\mathbf{x}} = \mathrm{U}^{\mathrm{T}}\mathbf{b} \text{ or } \mathrm{V}^{\mathrm{T}}\hat{\mathbf{x}} = \Sigma^{-1}\mathrm{U}^{\mathrm{T}}\mathbf{b},$$

which can be simplified as

$$\hat{\mathbf{x}} = \mathrm{V}\Sigma^{-1}\mathrm{U}^{\mathrm{T}}\mathbf{b} \ ,$$

and further as

$$\hat{\mathbf{x}} = \frac{\mathrm{U}^{\mathrm{T}}\mathbf{b}}{\Sigma}\mathrm{V}.$$

Therefore, for each $\mathrm{u_i} \in \mathrm{U}, \mathrm{v_i} \in \mathrm{V}$ and $\sigma_\mathrm{i} \in \Sigma$

$$\hat{\mathbf{x}} = \frac{\mathrm{u_i^T}\mathbf{b}}{\sigma_\mathrm{i}}\mathrm{v_i}. \tag{3.6}$$

Thus,

$$\hat{\mathbf{x}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{u_i^T}\mathbf{b}}{\alpha_\mathrm{i}}\mathrm{v_i}.$$

Equation 3.6 shows that small singular values can considerably magnify round-off errors, resulting in large errors in the solution.

Example 3.4. Use the singular value decomposition of the matrix $\mathbf{A}$ to find the least-squares solution of the system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is the Hilbert

matrix of order six defined as

$$\mathbf{A} = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 & 1/10 \\ 1/6 & 1/7 & 1/8 & 1/9 & 1/10 & 1/11 \end{bmatrix},$$

with $\mathbf{b}$ chosen such that the exact solution $\mathbf{x}$ is given by $\mathbf{x} = [1\ 1\ 1\ 1\ 1\ 1]^T$ and

$\mathbf{b} = [2.4500\ 1.5929\ 1.2179\ 0.99563\ 0.84563\ 0.73654]^{\mathrm{T}}$.

The singular values and the rank of $\mathbf{A}$ are : $\sigma_1 = 1.61890$, $\sigma_2 = 0.24236$, $\sigma_3 = 0.01632$, $\sigma_4 = 0.00062$, $\sigma_5 = 0.00001$, $\sigma_6 = 0.00000$, and rank(A) = 6 respectively. If the system $\mathbf{Ax} = \mathbf{b}$ is solved using singular value decomposition, the computed solution obtained is far from the exact solution. The m-file for implementing the least-squares solution for a system using SVD is shown below:

$$\begin{aligned} \text{function} &= [\mathbf{A}, \mathbf{b}, \mathbf{x}]; \\ \text{svd}(\mathbf{A}) &= [USV]; \\ \mathbf{b} * p &= U(:, i)' * \mathbf{b}; \\ \text{n} &= \text{length}(\mathbf{b}); \\ \text{s} &= \text{diag}(S); \\ \text{for } i &= 1 : n; \\ \mathbf{x}(:, \text{i}) &= \text{bp(i)/S(i, i)} * \text{V(:, i)}; \\ \text{Least Squares Solution} &= \text{sum}(\mathbf{x}')'; \\ \text{end;} & \end{aligned} \tag{3.7}$$

Remark 3.3. The computed solution $\hat{\mathbf{x}}$ is:

$$\hat{\mathbf{x}} = \begin{bmatrix} 1.1281 & -2.4402 & 23.25212 & -54.9612 & 61.1856 & -22.2314 \end{bmatrix}^{\mathrm{T}}.$$

The exact solution of $\mathbf{x} = \text{ones}(6, 1)$, an indication of a wide disparity between the two solutions. The error in the computed solution is attributed to the small singular values. To reduce the drastic effect of the small singular values on the solution, we choose a cut-off level α , such that $\sigma_{\mathrm{i}} < \alpha$. This eliminate the small singular values on the solution. The cutting off of the small singular values to obtain a better solution is called **Truncated Singular Value Decomposition**.

SVD and Stability of a Computed Solution

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be a rectangular or square matrix for the system $\mathbf{Ax} = \mathbf{b}$ and assume for ease of presentation that $\mathrm{m} \geq \mathrm{n}$. Then the SVD of $\mathbf{A}$ is a decomposition of the form

$$\mathbf{A} = \mathrm{U}\Sigma\mathrm{V}^{\mathrm{T}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{U_i}\Sigma_{\mathrm{i}}\mathrm{V_i^T}.$$

If $\mathbf{A}$ is invertible, then it's inverse is given by

$$\mathbf{A} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{v_i}\sigma_{\mathrm{i}}^{-1}\mathrm{u_i^T},$$

where

$$\mathrm{U} = [\mathrm{u}_1, ..., \mathrm{u_n}],\ \mathrm{V} = [\mathrm{v}_1, ..., \mathrm{v_n}],\ \text{and } \Sigma = \mathrm{diag}[\sigma_1, ..., \sigma_{\mathrm{n}}].$$

Therefore the solution to $\mathbf{Ax} = \mathbf{b}$ can be defined as

$$\mathbf{x} = \sum_{\mathrm{i}}^{\mathrm{n}} \sigma_{\mathrm{i}}^{-1}(\mathrm{u_i^T}\mathbf{b})\mathrm{v_i}. \tag{3.8}$$

The pseudo-inverse ($\mathbf{A}^+$) is also given by

$$\mathbf{A}^+ = \sum_{\mathrm{i}=1}^{\mathrm{rank(A)}} \mathrm{v_i}\sigma_{\mathrm{i}}^{-1}\mathrm{u_i^T}, \tag{3.9}$$

and the least squares solution $\hat{\mathbf{x}}$ to the least squares problem is given by

$$\hat{\mathbf{x}}_{\mathrm{ls}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{u}_{\mathrm{i}}^{\mathrm{T}}\mathbf{b}}{\sigma_{\mathrm{i}}}\mathrm{v}_{\mathrm{i}}. \tag{3.10}$$

The sensitivity or stability of the solution $\mathbf{x}$ and $\mathbf{x}_{ls}$ to perturbations of $\mathbf{A}$ and $\mathbf{b}$ can be measured by the 2-norm condition Number of $\mathbf{A}$, and is defined by

$$\mathrm{cond}(\mathbf{A}) = \|\mathbf{A}\|_2\|\mathbf{A}^{-1}\|_2 = \frac{\sigma_{\mathrm{l}}}{\sigma_{\mathrm{s}}}. \tag{3.11}$$

Thus, the condition number of $\mathbf{A}$ can be defined as the ratio of the largest singular values(σ_l) and the smallest singular values(σ_s) of $\mathbf{A}$. For Hilbert matrix of order 12, given it's singular values as shown in table below, the condition number can be determined.

Table 5: Singular Values for Hilbert Matrix of Order Twelve

Number	SVD-Values
1	$1.79537206e^{+}000$
2	$3.80275246e^{-}001$
3	$4.47385488e^{-}002$
4	$3.72231222e^{-}003$
5	$2.33089089e^{-}004$
6	$1.11633574e^{-}005$
7	$4.08237611e^{-}007$
8	$1.12286107e^{-}008$
9	$2.25196452e^{-}010$
10	$3.11134562e^{-}012$
11	$2.64930826e^{-}014$
12	$1.03967207e^{-}016$

The condition number from table 2.1, is $\mathrm{cond}(\mathbf{A}) = \frac{\sigma_{\mathrm{l}}}{\sigma_{\mathrm{s}}}$=1.726864e^{16}. This shows how ill-conditioned (or unstable) the solution is and the need

to rectify it. In order to improve the solution, we try to truncate (cut off) the small singular values for a better solution.

CHAPTER FOUR

Regularization Methods for Linear Ill-posed Problems

Regularization techniques are used to obtain meaningful estimates for discrete ill-posed problems or rank-deficient linear problems. In cases where some parameters are ill-determined either by least-square methods or in situations where the number of parameter is larger than the number of available measurements, it is necessary to stabilize the system by using regularization methods. Some of these methods are:

1. Truncated Singular Value Decomposition(TSVD).
2. Preconditioning .
3. Tikhonov Regularization Method.

Truncated Singular Value Decomposition(TSVD)

The idea behind truncated singular value decomposition is to replace all nonzero singular value less than certain threshold say, α with exact zeros. Thus, if σ_k is the smallest singular value greater than or equal to α, then $\sigma_{k+1} = \sigma_{k+2} = \sigma_{k+3} = ... = 0$.

Let α be the cut-off level, then all singular values smaller than α will be replaced by zero. The solution of a truncated problem at certain cut-off level say α is denoted by $\mathbf{x}_\alpha$ and defined as

$$\mathbf{x}_\alpha = \sum_{i=1}^{k} \frac{u_i^T \mathbf{b}}{\sigma_i} v_i.$$

Example 4.1. Consider the linear system $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 12×12 Hilbert matrix H_{12} and $\mathbf{b}$ defined as

$$\mathbf{b} = \sum_{\mathrm{j}=1}^{\mathrm{n}} \mathrm{h}_{\mathrm{ij}} \mathbf{x}_{\mathrm{j}}.$$

This linear system has the exact solution

$\mathbf{x} = [1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1]^T$,

with 16 digits of accuracy. The computed solution using

$$\hat{\mathbf{x}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{\mathrm{u}_{\mathrm{i}}^{\mathrm{T}} \mathbf{b}}{\sigma_{\mathrm{i}}} \mathrm{v}_{\mathrm{i}} \ ,$$

is shown below.

Table 6: Accuracy of a Computed Solution.

(k)	Exact Solution($\mathbf{x}$)	Computed Solution($\hat{\mathbf{x}}$)
1	1.0000000000000000	0.999999984903960
2	1.0000000000000000	1.000001882544423
3	1.0000000000000000	0.999941439900766
4	1.0000000000000000	1.000791780820418
5	1.0000000000000000	1.000791780820418
6	1.0000000000000000	1.025258190023834
7	1.0000000000000000	0.929842667541179
8	1.0000000000000000	1.126690208270141
9	1.0000000000000000	0.851749330982105
10	1.0000000000000000	1.108415121437981
11	1.0000000000000000	0.954977579488462
12	1.0000000000000000	1.008104137234545

The maximum error is

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_\infty = 1.322341312564\mathrm{e}^{00}.$$

This shows there is no precision. The error in the solution is associated with the smaller singular values in the solutions .

However, if we truncate (cut-off) the smaller singular values (σ) to reduce the drastic effect on the solution at various cut-off levels (α), we obtain

Table 7: Truncated Singular Values

i	σ_i	α	Rank r	max. error
1	1.7954×10^{0}		1	
2	3.8028×10^{-01}		2	
3	4.4739×10^{-02}		3	
4	3.7223×10^{-03}		4	
5	2.3309×10^{-04}		5	
6	1.1163×10^{-05}	1.1163×10^{-05}	6	1.09×10^{-3}
7	4.0824×10^{-07}	4.0824×10^{-07}	7	1.63×10^{-4}
8	1.1229×10^{-08}	1.1229×10^{-08}	8	2.91×10^{-5}
9	2.2520×10^{-10}	2.2520×10^{-10}	9	3.57×10^{-6}
10	3.1113×10^{-12}	3.1113×10^{-12}	10	4.51×10^{-5}
11	2.6492×10^{-14}	2.6492×10^{-14}	11	1.19×10^{-3}
12	1.0772×10^{-16}	1.0772×10^{-16}	12	3.31×10^{-1}

From Table 7, it is observe that the solution of the Hilbert matrix of order 12 has no significant digit of accuracy. If we truncate it to $\alpha = 2.6492 \times 10^{-14}$, we gain some significant digits of accuracy. This significance increases to five for $\alpha = 3.1113 \times 10^{-12}$ and six for $\alpha = 2.2520 \times 10^{-10}$. As we truncate further to $\alpha = 2.2520 \times 10^{-10}$ and beyond, the solution deteriorates again. The minimum error in the solutions occurs at the cut-off level $\alpha = 2.2520 \times 10^{-10}$. The optimal solution(Sol) for $\alpha = 2.2520 \times 10^{-10}$ is shown in the Table 8 below.

Table 8: Optimal Truncated Solution

Exact solution	Sol. with r = 12, $\alpha = 0$	Sol. with r = 9, $\alpha = 2.25e^{-10}$
1.0000000000000e+00	9.9999996705141e-01	1.0000000000328e+00
1.0000000000000e+00	1.0000041370137e+00	9.9999999700270e-01
1.0000000000000e+00	9.9987071398844e-01	1.0000000607111e+00
1.0000000000000e+00	1.0017537108519e+00	9.9999951417830e-01
1.0000000000000e+00	9.8718532993339e-01	1.0000018473071e+00
1.0000000000000e+00	1.0561684056900e+00	9.9999665243828e-01
1.0000000000000e+00	8.4379161569956e-01	1.0000018848108e+00
1.0000000000000e+00	1.2823440776789e+00	1.0000021835864e+00
1.0000000000000e+00	6.6937445118911e-01	9.9999768413185e-01
1.0000000000000e+00	1.2419130000125e+00	9.9999787257888e-01
1.0000000000000e+00	8.9949871252080e-01	1.0000035737227e+00
1.0000000000000e+00	1.0180958999351e+00	9.9999872943484e-01

Preconditioning

The use of iterative methods for solving symmetric positive definite systems of linear equations, require some form of preconditioning M to improve the convergence of the solution by manipulating the spectrum of the coefficient matrix. Given the linear system $\mathbf{Ax} = \mathbf{b}$, we transform it into an equivalent system of the form $\mathbf{MAx} = \mathbf{Mb}$, such that the conditioned number of $\mathbf{MA}$ is far less than that of $\mathbf{A}$, that is $\kappa(\mathbf{MA}) << \kappa(\mathbf{A})$. Basically, there are two types of preconditioner's, the left and the right preconditioner's, but for this study we will restrict ourself to only the left preconditioner.

The Jacobi and Gauss-Seidel Preconditioner

The method of jacobi and Gauss-Seidel for solving $\mathbf{Ax} = \mathbf{b}$, split the matrix $\mathbf{A}$ into $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, where $\mathbf{L}$ is a lower triangular matrix, $\mathbf{U}$ is an upper triangular matrix and $\mathbf{D}$ a diagonal matrix. The Jacobi scheme is given by

$$\mathbf{x}^{k+1} = -\mathbf{D}^{-1}(\mathbf{U} + \mathbf{L})\mathbf{x}^k + \mathbf{D}^{-1}\mathbf{b}, \quad k = 0, 1, 2, \cdots$$

and the Gauss-Seidel scheme is also given by

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

Both scheme converge to the solution if the matrix is *strictly diagonally dominant*.

In the Jacobi iteration, the matrix $\mathbf{D}^{-1}$ is used to rescale all the non-diagonal entries of the matrix $\mathbf{A}$, to obtain a good preconditioner, known as the **Jacobi Preconditioner**. Thus, if $\mathbf{A}$ is ill-conditioned then $\mathbf{M} = \mathbf{D}^{-1}\mathbf{A}$ is better conditioned than $\mathbf{A}$.

A Gauss-Seidel preconditioner can be used to solve the same problem. Here, the preconditioning matrix is lower triangular, and is defined as $\mathbf{M} = (\mathbf{L} + \mathbf{D})^{-1}$, from the Gauss-Seidel iteration. Solving a linear system with a Gauss-Seidel preconditioner is computationally expensive, but at times yield a better result than the Jacobi Preconditioner. Both the Jacobi and the Gauss-Seidel schemes are used for linear systems where the coefficient matrix is *sparse*, consist mainly of zeros. Such matrices occur in the numerical solution of boundary-value problems. The following examples illustrate the use of these preconditioner's.

Example 4.2. Consider the boundary-value problem

$$\mathrm{u_{xx}} + \mathrm{u_{yy}} = 0, \tag{4.1}$$

in the rectangle

$\mathrm{R} = \{(\mathrm{x}, \mathrm{y}) : \ 0 \leq \mathrm{x} \leq 4, \quad 0 \leq \mathrm{y} \leq 4\}$,

where $u(x, y)$ denotes the temperature at the point $u(x, y)$, with boundary values

$$\mathrm{u(x, 0)} = 180, \quad \mathrm{u(x, 4)} = 20, \quad \text{for} \quad 0 < \mathrm{x} < 4$$

$$\mathrm{u(0, y)} = 80, \quad \mathrm{u(4, y)} = 0, \quad \text{for} \quad 0 < \mathrm{y} < 4.$$

Central-difference approximation for $\mathrm{u_{xx}}$ and $\mathrm{u_{yy}}$ are given by

$$\mathrm{u_{xx}} = \frac{\mathrm{u(x + h, y) - 2u(x, y) + u(x - h, y)}}{\mathrm{h^2}} \tag{4.2}$$

$$\mathrm{u_{yy}} = \frac{\mathrm{u(x, y + h) - 2u(x, y) + u(x, y - h)}}{\mathrm{h^2}}. \tag{4.3}$$

Then,

$$\begin{aligned} u_{xx}(x, y) + u_{yy}(x, y) \quad &\approx \quad u(x + h, y) + u(x - h, y) + u(x, y + g) + \\ &\qquad u(x, y - g) - 4u(x, y) = 0. \end{aligned}$$

Choosing $\mathrm{h = g = 1}$, corresponds to approximating the temperature $\mathrm{u(x, y)}$ at nine interior points in the rectangle shown in Figure 4.

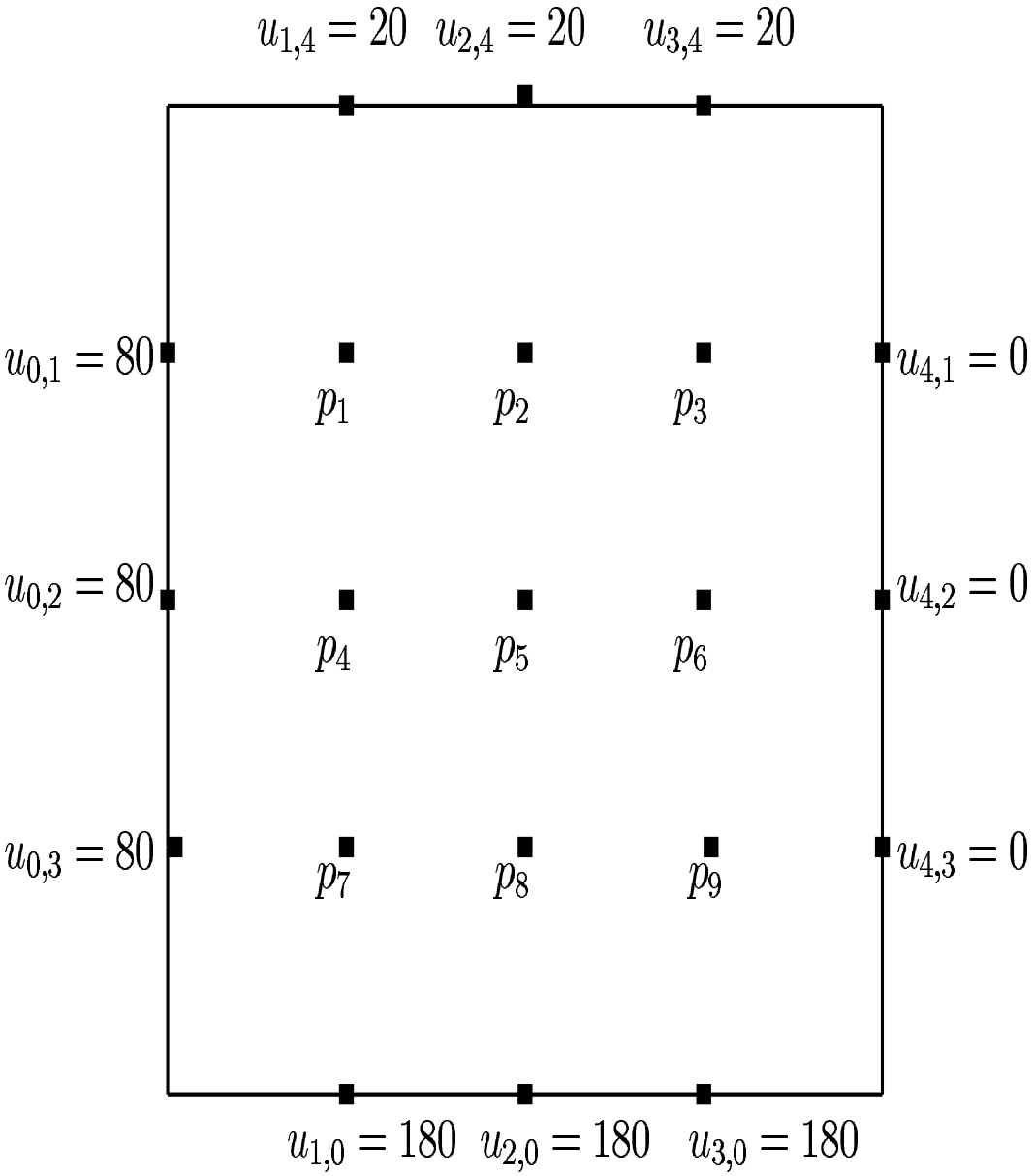

Figure 1: The grid for Example (4.2)

Using the central difference formula for second derivative, to approximate the partial derivatives at each of the nine interior points, results in a series of linear systems. Thus u_{xx} and u_{yy} together with the *five-point formula* gives

the following linear systems,

$$\begin{array}{rrrrrrrrrl}
-4p_1 & +p_2 & & +p_4 & & & & & & = -100 \\
p_1 & -4p_2 & +p_3 & & +p_5 & & & & & = -20 \\
& p_2 & -4p_3 & & & +p_6 & & & & = -20 \\
p_1 & & & -4p_4 & +p_5 & & +p_7 & & & = -80 \\
& p_2 & & +p_4 & -4p_5 & +p_6 & & +p_8 & & = 0 \\
& & p_3 & & +p_5 & -4p_6 & & & +p_9 & = 0 \\
& & & p_4 & & & -4p_7 & +p_8 & & = -260 \\
& & & & p_5 & & p_7 & -4p_8 & +p_9 & = -180 \\
& & & & & p_6 & & +p_8 & -4p_9 & = -180
\end{array}$$

where $\mathrm{P}_1, \mathrm{P}_2, \cdots, \mathrm{p}_9$ are approximations of the temperature in the interior of the rectangle. In matrix form, we have $\mathbf{Ap} = \mathbf{b}$, where $\mathbf{A}$ is a 9×9 symmetric positive definite matrix.

$$\mathbf{A} = \begin{bmatrix}
-4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & -4 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 1 & -4 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 0 & -4 & 1 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 1 & -4 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & -4 & 0 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 & 0 & -4 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & -4 & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & -4
\end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix}
-100 \\ -20 \\ -20 \\ -80 \\ 0 \\ 0 \\ -260 \\ -180 \\ -180
\end{bmatrix} \quad \text{and}$$

$$\mathbf{p} = \begin{bmatrix} p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 \end{bmatrix}^T.$$

The solution of the above system gives a crude approximation to the temperature distribution $\mathrm{u}(\mathrm{x_i}, \mathrm{y_j})$ in the interior of the rectangular region. A more accurate estimate of the temperature distribution can be obtain by using smaller step size.

For instance with h = g = 1/2, we obtain a linear system $\mathbf{Ap} = \mathbf{b}$, where $\mathbf{A}$ is a 49×49 symmetric positive definite matrix. Similarly, using h = g = 1/4 gives a linear system $\mathbf{Ap} = \mathbf{b}$, where $\mathbf{A}$ is a 225×225 symmetric positive definite matrix. The coefficient matrix $\mathbf{A}$ becomes increasingly ill-conditioned as the step size becomes smaller, and the dimension of the coefficient matrix becomes larger. The system is sparse and the coefficient matrix is strictly diagonally dominant, and so the Jacobi and the Gauss-Seidel scheme's can be used to solve the systems. For relatively small values of the matrix $\mathbf{A}$ the solution vector $\mathbf{p}$ can be obtained relatively easily.

We solve the linear system $\mathbf{Ap} = \mathbf{b}$ using the Gauss-Seidel and the Jacobi method without preconditioning. The solutions obtained shows some convergence to the exact solutions. If the system is regularized using the Gauss-Seidel and Jacobi preconditioner's, the optimal solutions approximate accurately to the exact solutions. The optimal solutions for the Jacobi Preconditioner (X_{Jpred}) and the Gauss-Seidel preconditioner (X_{GSpred}) are shown in the Table 9 below.

Table 9: Optimal Solutions For Jacobi and Gauss-Seidel Preconditioner's.

X_J	X_G	X_{Jpred}	X_{GSpred}	Exact
55.6630161963	55.7141486236	55.7142857142	55.7142857142	55.7143
43.1459263712	43.2141486236	43.2142857142	43.2142857142	43.2143
27.0915876515	27.1427885975	27.1428571428	27.1428571428	27.1429
79.5744977518	79.6427200521	79.6428571428	79.6428571428	79.6429
69.8974609375	69.9998629093	70.0000000000	70.0000000000	70.0000
45.2887834981	45.3570743118	45.3571428571	45.3571428571	45.3571
112.805873285	112.857074311	112.857142857	112.857142857	112.857
111.717354878	111.785645740	111.785714285	111.785714285	111.786
84.2344447411	84.2856800130	84.2857142857	84.2857142857	84.2857

From Table 9, we can conclude that the solution from preconditioning gives better accuracy. There is no much difference in the accuracy obtained from either preconditioner's. This explains why the cheaper Jacobi preconditioner is more popular.

Tikhonov Regularization Method

Regularization methods for least square problems are the most commonly used method for obtaining stable and smooth solution to rank deficient and ill-posed problems, Hansen(1992). In solving such problems, it is necessary to incorporate additional information as smoothness, continuity and the size of the residual to obtain the desired solution for $\mathbf{x}$. Such additional information is then used as a constraint to control the smoothness of the solution. The side constraint is usually of the form

$$\mathrm{C}_\lambda(\mathbf{x}) = \lambda \|\mathrm{L}\mathbf{x}\|_2^2, \tag{4.4}$$

where L is the identity matrix ($\mathrm{I_n}$) or an $(\mathrm{n}-\mathrm{p}) \times \mathrm{n}$ discrete approximation of the p^{th} derivative operator. The side constraint gives a fair balance between minimizing $\mathrm{C}_\lambda(\mathbf{x})$ and minimizing the residual norm $\|\mathbf{Ax}-\mathbf{b}\|_2^2$ instead of giving us the solution of $\mathbf{Ax} = \mathbf{b}$. The basic idea is that a regularized solution $\mathbf{x}$ should give a small residual and also be small in $2-$ norm to give a desired solution. One of the most important form of regularization of ill-posed least squares problems is the **Tikhonov regularization.** This method is often used to regularize ill-posed problems. It involves obtaining the exact or least squares solution of linear systems by minimizing the function

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{Ax}-\mathbf{b}\|_2^2, \tag{4.5}$$

subject to the side constraint $\|\mathrm{L}\mathbf{x}\|$, where the matrix L is either an identity matrix, a diagonal weighting matrix or an $(\mathrm{n}-\mathrm{p}) \times \mathrm{n}$ discrete approximation of the p^{th} derivative operator.

Since standard Algorithms normally fails to provide suitable solution to stabilize the system, the regularized solution $\mathbf{x}_\lambda$, is defined as the minimizer of the weighted combination of the residual and the side constraint according to Wang and Linz(2003). The minimized expression is

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathrm{L}\mathbf{x}\|_2^2, \tag{4.6}$$

where λ is greater than zero and is called a regularization parameter.

Regularization of Order Zero

If we let $\mathrm{L} = \mathrm{I_n}$, the minimizing function becomes

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2. \tag{4.7}$$

This signifies that it can be expressed as a balance between the quantities, $\|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$ and $\|\mathbf{x}\|$. Here, the regularization parameter controls the weights given to minimization of the side constraint relative to the minimization of the residual norm. But we can show that the minimizing solution($\mathbf{x}_\lambda$) is given by the non-singular linear system as $(\mathbf{A}'\mathbf{A} + \lambda\mathrm{I})\mathbf{x}_\lambda = \mathbf{A}'\mathbf{b}$. From

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda\|\mathbf{x}\|_2^2, \tag{4.8}$$

it follows that

$$\phi_\lambda = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x}) - (\mathbf{A}\mathbf{x})'\mathbf{b} - \mathbf{b}'(\mathbf{A}\mathbf{x}) + \mathbf{b}'\mathbf{b} + \lambda\mathbf{x}'\mathbf{x}. \tag{4.9}$$

But $(\mathbf{A}\mathbf{x})'\mathbf{b}$ and $\mathbf{b}'(\mathbf{A}\mathbf{x})$ are equal since they are scalars. So

$$\phi_\lambda = (\mathbf{A}\mathbf{x})'(\mathbf{A}\mathbf{x}) - 2(\mathbf{A}\mathbf{x})'\mathbf{b} + \lambda\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b}, \tag{4.10}$$

$$\phi_\lambda = \mathbf{x}'(\mathbf{A}'\mathbf{A})'\mathbf{x} - 2\mathbf{A}'\mathbf{x}'\mathbf{b} + \lambda\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b}, \tag{4.11}$$

Differentiating the function $\phi_\lambda(\mathbf{x})$ for the minimizing solution $\mathbf{x}_\lambda$, we obtain

$$\frac{\partial\phi_\lambda(\mathbf{x})}{\partial\mathbf{x}}|_{\mathbf{x}=\mathbf{x}_\lambda} = 0. \tag{4.12}$$

It implies that

$$\frac{\partial \phi_\lambda(\mathbf{x})}{\partial \mathbf{x}}|_{\mathbf{x}=\mathbf{x}_\lambda} = 2\mathbf{A}'\mathbf{A}\mathbf{x}_\lambda - 2\mathbf{A}'\mathbf{b} + 2\lambda\mathbf{x}_\lambda = 0, \tag{4.13}$$

or

$$(\mathbf{A}'\mathbf{A} + \lambda \mathrm{I})\mathbf{x}_\lambda = \mathbf{A}'\mathbf{b}, \tag{4.14}$$

that is

$$\mathbf{x}_\lambda = (\mathbf{A}'\mathbf{A} + \lambda \mathrm{I})^{-1}(\mathbf{A}'\mathbf{b}). \tag{4.15}$$

This regularization method above penalizes large components in the solution and is called regularization of order Zero. Using singular value decomposition, we can simplify further the minimizer $\mathbf{x}_\lambda$.

From $\mathbf{x}_\lambda = (\mathbf{A}'\mathbf{A} + \lambda \mathrm{I})^{-1}(\mathbf{A}'\mathbf{b})$, substituting

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}} = \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathbf{u}_\mathrm{i}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}$$

into the solution($\mathbf{x}_\lambda$), we obtain

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [(\mathbf{u}_\mathrm{i}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}})^{\mathrm{T}}(\mathbf{u}_\mathrm{i}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}}) + \lambda \mathrm{I}]^{-1}(\mathbf{u}_\mathrm{i}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}})^{\mathrm{T}}\mathbf{b}, \tag{4.16}$$

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}^{\mathrm{T}}\mathbf{u}_\mathrm{i}^{\mathrm{T}}\mathbf{u}_\mathrm{i}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}}) + \lambda \mathrm{I}]^{-1}(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}^{\mathrm{T}}\mathbf{u}_\mathrm{i}^{\mathrm{T}})\mathbf{b}. \tag{4.17}$$

But $\mathbf{u}_\mathrm{i}\mathbf{u}_\mathrm{i}^{\mathrm{T}} = \mathbf{u}_\mathrm{i}^{\mathrm{T}}\mathbf{u}_\mathrm{i} = \mathrm{I}_\mathrm{n}$ and $\mathbf{v}_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}} = \mathbf{v}_\mathrm{i}^{\mathrm{T}}\mathbf{v}_\mathrm{i} = \mathrm{I}_\mathrm{n}$. Hence

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}^{\mathrm{T}}\mathrm{I}\sigma_\mathrm{i}\mathbf{v}_\mathrm{i}^{\mathrm{T}}) + \lambda \mathrm{I}]^{-1}(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}^{\mathrm{T}}\mathbf{u}_\mathrm{i}^{\mathrm{T}})\mathbf{b}. \tag{4.18}$$

Also σ_i and $\sigma_\mathrm{i}^{\mathrm{T}}$ are scalars and equal. So we have

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}^2\mathbf{v}_\mathrm{i}^{\mathrm{T}}) + \lambda \mathrm{I}]^{-1}(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}\mathbf{u}_\mathrm{i}^{\mathrm{T}})\mathbf{b}, \tag{4.19}$$

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [\sigma_\mathrm{i}^2 + \lambda]^{-1}(\mathbf{v}_\mathrm{i}\sigma_\mathrm{i}\mathbf{u}_\mathrm{i}^{\mathrm{T}})\mathbf{b}, \tag{4.20}$$

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} [\frac{\sigma_\mathrm{i}}{\sigma_\mathrm{i}^2 + \lambda}](\mathbf{v}_\mathrm{i}\mathbf{u}_\mathrm{i}^{\mathrm{T}})\mathbf{b}, \tag{4.21}$$

$$\mathbf{x}_\lambda = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\frac{\sigma_\mathrm{i}^2}{\sigma_\mathrm{i}^2 + \lambda})(\frac{\mathbf{u}_\mathrm{i}^{\mathrm{T}}\mathbf{b}}{\sigma_\mathrm{i}})\mathbf{v}_\mathrm{i}. \tag{4.22}$$

where $\sigma_1 \geq \sigma_2 \geq \sigma_3 \geq ... \geq \sigma_n > 0$ are the singular values of A, and $\mathbf{u}, \mathbf{v}$ are respectively the left and right singular vectors of $\mathbf{A}$. Thus the effect of the addition of $\lambda\|\mathbf{x}\|_2^2$ to the equation $\phi_\lambda(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\|_2^2$ is to dampen the contributions of the terms involving small singular values so that instead of cutting them off totally we modify the methods to reduce its impact. A small λ has very little effect on the component associated with large singular values. If λ is far smaller than σ_i^2, then

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \cong \frac{1}{\sigma_i}, \tag{4.23}$$

and Equation 4.22 becomes

$$\mathbf{x}_\lambda = \sum_{i=1}^{n} (\frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i}) \mathbf{v}_i. \tag{4.24}$$

On the other hand, if σ_i^2 is much more smaller than λ, then

$$\frac{\sigma_i}{\sigma_i^2 + \lambda} \cong \frac{\sigma_i}{\lambda} << \frac{1}{\sigma_i}, \tag{4.25}$$

such that the magnification of the components associated with small singular values are reduced. With a good choice of λ one can then hope to get a relatively smooth solution which is a good approximation to the true solution. This is called SVD with damping.

Regularization of Linear Systems

We have realized that if the coefficient matrix $\mathbf{Ab}$ of the linear system $\mathbf{Abx} = \mathbf{b}$ is ill-conditioned, then the computed solution $\hat{\mathbf{x}}$, is usually a meaningless approximation to $\mathbf{x}$. Regularization methods are often used to obtain stable and smooth solutions to ill-conditioned problems. To regularize the solution, we solve instead the perturbed system

$$\mathbf{Mby} = \mathbf{b}, \tag{4.26}$$

where $\mathbf{Mb} = (\mathbf{Ab} + k\mathbf{Ib}_n)$, and k is a small positive number.

Since **Ab** is positive-definite, the eigenvalues of **Ab** must satisfy

$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n = \lambda_{\min} > 0.$

The condition number of **Ab** is given by

$$\kappa(\mathbf{Ab}) = \frac{\lambda_{\max}}{\lambda_{\min}} \ .$$

Similarly

$$\kappa(\mathbf{Mb}) = \frac{\lambda_{\max} + k}{\lambda_{\min} + k} \quad \text{for all } k > 0 \ .$$

Theorem 4.1. For all positive values of k, the matrix **Mb** in Equation 4.26 is better conditioned than **Ab**, in the sense that

$$\kappa(\mathbf{Mb}) \ < \ \kappa(\mathbf{Ab}).$$

Proof. For any $k > 0$, $\quad k\lambda_{\max} \quad \geq \quad k\lambda_{\min}$, and so

$$\begin{aligned} \lambda_{max}\lambda_{min} + k\lambda_{max} \quad &\geq \quad \lambda_{max}\lambda_{min} + k\lambda_{mim} \\ \lambda_{max}\left(\lambda_{min} + k\right) \quad &\geq \quad \lambda_{min}\left(\lambda_{max} + k\right) \\ \frac{\lambda_{max}}{\lambda_{min}} \quad &\geq \quad \frac{\lambda_{max} + k}{\lambda_{min} + k} \end{aligned}$$

This shows that $\kappa(\mathbf{Ab}) \quad \geq \quad \kappa(\mathbf{Mb})$.

Theorem 4.2. $\kappa(\mathbf{Mb})$ is a decreasing function of k, that is

$$k_1 < k_2 \Rightarrow \qquad \frac{\lambda_{\max} + k_2}{\lambda_{\min} + k_2} \quad \leq \quad \frac{\lambda_{\max} + k_1}{\lambda_{\min} + k_1}$$

The Table 10 below illustrates Theorem (4.2).

Table 10: Decreasing Condition Number with Parameter k

Regularization Parameter (k)	$\kappa(\mathbf{Mb})$
1.00000000000000e-15	1.62732922269604e+15
1.00000000000000e-14	1.77544770319155e+14
1.00000000000000e-13	1.79341052915694e+13
1.00000000000000e-12	1.79517106478186e+12
1.00000000000000e-11	1.79535290832870e+11
1.00000000000000e-10	1.79537017493371e+10
1.00000000000000e-09	1.79537186699643e+09
1.00000000000000e-08	1.79537205075106e+08
1.00000000000000e-07	1.79537215756213e+07
1.00000000000000e-06	1.79537305937215e+06
1.00000000000000e-05	1.79538205954236e+05
1.00000000000000e-04	1.79547205956003e+04
1.00000000000000e-03	1.79637205956180e+03
1.00000000000000e-02	1.80537205956198e+02
1.00000000000000e-01	1.89537205956199e+01
1.00000000000000e+00	2.79537205956199e+00
1.00000000000000e+01	1.17953720595620e+00
1.00000000000000e+02	1.01795372059562e+00
1.00000000000000e+03	1.00179537205956e+00
1.00000000000000e+04	1.00017953720596e+00

Regularization of Order One

Regularization of Equation 4.7 in most cases dampens components that are large in magnitudes, since the component in a solution oscillate with moderate amplitudes. Such component are undesirable and may need a penalty term that is large for rapid change in the solution. This penalty term result in another form of regularization called "Order One". For this reason the penalty term is added to Equation 4.7 to obtain

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \sum_{\mathrm{i}=2}^{\mathrm{n}} (|\mathbf{x}|_{\mathrm{i}} - |\mathbf{x}|_{\mathrm{i}-1})^2. \tag{4.27}$$

The above expression is minimized by the solution of $(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda \mathrm{L}_1^{\mathrm{T}}\mathrm{L}_1)\mathbf{x}_\lambda = \mathbf{A}^{\mathrm{T}}\mathbf{b}$, where L_1 is an $(\mathrm{n}-1) \times \mathrm{n}$ first derivative operator defined as

$$\mathrm{L}_1 = \begin{bmatrix} 1 & -1 & 0 & \cdots & o \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & 1 & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}.$$

The m-file for implementing L_1 is shown below:

function $\mathrm{M} = \mathrm{L}_1(\mathrm{n})$;

create an m-file for computing an $(\mathrm{n}-1) \times \mathrm{n}$ first derivative operator M.

$\mathrm{M} = \mathrm{zeros}(\mathrm{n}-1, \mathrm{n})$;

for $\mathrm{i} = 1 : \mathrm{n}-1$

for $\mathrm{j} = 1 : \mathrm{n}$

if $\mathrm{i} == \mathrm{j}, \mathrm{M}(\mathrm{i},\mathrm{j}) = 1$;

else if $\mathrm{i} == \mathrm{j}-1,\ \mathrm{M}(\mathrm{i},\mathrm{j}) = -1$;

else $\mathrm{M}(\mathrm{i},\mathrm{j}) = 0$;

end.

The first derivative operator L_1 helps us to compute $\mathbf{x}_\lambda$ for order one regularization using matlab or octave script files for regularization.

Regularization of Order Two

This type of regularization is not far from order one regularization. Here, the penalty term is much stronger than that in order one. Thus regularization of order two is based on minimizing

$$\phi_\lambda(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 + \lambda \sum_{\mathrm{i}=2}^{\mathrm{n}} (|\mathbf{x}|_{\mathrm{i}+1} - 2|\mathbf{x}| + |\mathbf{x}|_{\mathrm{i}-1})^2, \tag{4.28}$$

which lead to the system $(\mathbf{A}^T\mathbf{A} + \lambda L_1^T L_1)\mathbf{x}_\lambda = \mathbf{A}^T\mathbf{b}$. Here L_2 is an $(\mathrm{n}-2) \times \mathrm{n}$ second derivative operator defined as

$$\mathrm{L}_2 = \begin{bmatrix} 1 & -2 & 1 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -2 & 1 \end{bmatrix}.$$

The second derivative operator L_2 also helps us to compute $\mathbf{x}_\lambda$ for order two regularization using matlab or octave script files for regularization.

Example 4.3. Consider the system $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A}$ is a 12×12 Hilbert system and $\mathbf{b}$ chosen such that the system has exact solution $\mathbf{x} = \mathrm{ones}(12, 1)$.

If the linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ is solved using any of the standard methods (LU-factorization, QR-factorization, Cholesky and Singular Value Decomposition), we realize that the computed solution $\hat{\mathbf{x}}$ shows a little resemblance to the exact solution. The exact solution, and the computed solution are shown in the Table 11 below. The maximum error in the solution using infinity norm is $||\mathbf{x} - \hat{\mathbf{x}}||_\infty = 1.89056261650000\mathrm{e} - 004$.

However, if the same linear system is solved using regularization methods of order zero, one or two for a given range of values of the regularization parameter λ , we obtain convergent solutions for various orders which approximate the exact solution. The detailed regularized solutions of order zero, one and two are shown in appendix B.

Table 11: Comparison of Solutions

Exact Solution($\mathbf{x}$)	Computed Solution($\hat{\mathbf{x}}$)
1.000000000000000	1.000000000361744
1.000000000000000	0.999999967356196
1.000000000000000	1.000000750743523
1.000000000000000	0.999992519342171
1.000000000000000	1.000039112701456
1.000000000000000	0.999884579842439
1.000000000000000	1.000189056261650
1.000000000000000	0.999864632593107
1.000000000000000	0.999946095450604
1.000000000000000	1.000177724682646
1.000000000000000	0.999874569344744
1.000000000000000	1.000030991570177

The regularized solutions for some selected values of λ for order one, which clearly shows the convergence to the exact solution, is shown in the Table 12 below. From Table 12, the regularized solutions for the parameter λ, from 10^{-10} onward, shows a tremendous improvement in the regularized solutions compared to the unregularized. For $\lambda = 10^3$, the regularized solution, converge close to the exact solution. The maximum error at each regularization parameter helps us to identity the optimal solution of the system. The Table 13 below, illustrate the maximum error and their corresponding regularization parameter. The optimal solution of the system correspond to the regularization parameter with the least maximum error. From Table 13, the least error is $\lambda = 6.66133814775094\text{e}^{-16}$. Hence the optimal solution occurs at $\lambda = 10^0$ with the solution shown in Table 12.

Table 12: Some Selected Solutions and their Regularization Parameters.

Unregularized Solution($\hat{\mathbf{x}}$)	$\mathbf{x}$ at $\lambda = 10^{-15}$	$\mathbf{x}$ at $\lambda = 10^{-10}$
1.000000000361744	0.999966246960022	0.999999987662394
0.999999967356196	1.001226252624526	1.000000242617692
1.000000750743523	0.990021092332578	0.999998962191316
0.999992519342171	1.026750479592941	1.000001181241874
1.000039112701456	0.986324452717069	1.000000467702928
0.999884579842439	0.976656405745774	0.999999383875175
1.000189056261650	0.972026382861084	0.999999042356027
0.999864632593107	1.086991975538807	1.000000432616368
0.999946095450604	1.018763248686092	1.000000375736882
1.000177724682646	0.919011410428752	0.999999748504047
0.999874569344744	0.995178142675670	1.000000228535360
1.000030991570177	1.027115874077475	0.999999953867416
$\mathbf{x}$ at $\lambda = 10^{-5}$	$\mathbf{x}$ at $\lambda = 10^{0}$	$\mathbf{x}$ at $\lambda = 10^{2}$
0.999999999998463	1.000000000000001	0.999999999999999
1.000000000010832	1.000000000000001	0.999999999999999
0.999999999983163	1.000000000000000	0.999999999999999
1.000000000007078	1.000000000000000	0.999999999999998
1.000000000006437	0.999999999999999	0.999999999999998
0.999999999995342	0.999999999999999	0.999999999999998
0.999999999987873	0.999999999999999	0.999999999999998
1.000000000001512	1.000000000000000	0.999999999999998
1.000000000001211	1.000000000000000	0.999999999999998
0.999999999997879	1.000000000000000	0.999999999999998
1.000000000006480	1.000000000000000	0.999999999999997
1.000000000005793	1.000000000000000	0.999999999999997

Table 13: Regularization Parameter for Optimal Solution

Unregularization Parameter (λ)	Maximum error
10e-16	2.98722262889516e-01
10e-15	8.69919755388067e-02
10e-14	8.65490837008687e-03
10e-13	9.35832178475593e-04
10e-12	1.02242916088624e-04
10e-11	1.2481417730488oe-05
10e-10	1.18124187364899e-06
10e-09	1.06160033430669e-07
10e-08	1.58048437064400e-08
10e-07	1.50154999545293e-09
10e-06	1.94635751959993e-10
10e-05	1.68371983022553e-11
10e-04	1.74105174727170e-12
10e-03	2.18491891246231e-13
10e-02	4.90718576884319e-14
10e-01	5.99520433297585e-15
10e+00	6.66133814775094e-16
10e+01	2.44249065417534e-15
10e+02	2.77555756156289e-15
10e+03	1.3766765053519e-14

Parameter-Choice Methods

In regularization, various algorithms are used for computing a regularized solution, but these algorithms have their advantages and disadvantages in terms of implementation issues, filter properties and others.

However no regularization method is complete without a method for choosing the regularization parameters. In all cases either the continuous parameter λ or the discrete parameter k must be chosen. Most of these parameter choice methods are based on residual norms. Basically, some of these parameter choice methods for regularization are the L-Curve method and Discrepancy principle, by Hansen(1989) and Benyah(2005). A good regularization parameter should yield a fair balance between the perturbation error and the regularization error in the regularized solution.
These parameter choice methods can be roughly divided into two classes depending on their assumptions about error and the norm of the perturbation at the right-hand side. These two classes can be characterized as follows:

1. Methods based on Knowledge on good estimate of the error norms.

2. Methods that do not require error norms, but instead seek to extract the necessary information from the given right hand side.

For many of these methods, the convergence rate for the solution is very necessary.

Choosing the Regularization Parameter for Tikhonov

For Tikhonov regularization,we have to make two choices :

1. The order of the regularization.

2. The regularization parameter λ .

The regularization parameter λ controls the weight given to the minimization of $\|\mathrm{L}\mathbf{x}\|_2^2$ relative to the minimization of the residual norm $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$. A large λ favors a small solution norm at the expense of a large residual norm, while a small λ favors a large solution norm at the expense of a small residual norm. Normally when the solution have a great deal of smoothness, second order regularization is preferred. This is because it improves the solution than the lower orders.

However, when the exact solution have points where it's value or it's derivative changes rapidly, second order regularization has the tendency to smooth the solution too much. In such cases, a lower order regularization is better, although in some cases none of the orders may give a very good result.

Here, a number of strategies can be used to select λ such that the regularized solution $\mathbf{x}_\lambda$ is both acceptable and plausible. A simple way of choosing the regularization parameter is to compute the results for various λ and inspect the results. If λ decreases, the residual of $\mathbf{x}_\lambda$ decreases, while if λ increases, the residual of $\mathbf{x}_\lambda$ increases.

Suppose we find λ for which the residual is sufficiently small and satisfies our intuitive idea of an acceptable solution, then we can accept that solution. In many cases, the behavior of $\|\mathbf{A}\mathbf{x}_\lambda-\mathbf{b}\|$ gives a good indication of a proper choices of λ. If $\|\mathbf{A}\mathbf{x}_\lambda-\mathbf{b}\|$ increase or decrease steadily with λ, until some critical value λ_0 after which further increase or decrease in λ have little effect on the residual norm, then λ_0 can be chosen as the optimal regularization parameter. The solution corresponding to λ_0 becomes the acceptable or the optimal solution $\mathbf{x}_\lambda$. Also a decreasing λ will make $\|\mathrm{L}\mathbf{x}_\lambda\|$ larger, that is, reduce the plausibility of the solution and vice versa. This method is by inspection.

The L-Curve Method

The L-curve is a graphical plot for all valid regularization parameters of the discrete smoothing norm. It's the most convenient graphical tool for analysis of discrete ill-posed problems, Hansen(1997). For example norm $\|\mathrm{L}\mathbf{x}\|_2^2$ of the regularized solution of a problem versus it's corresponding residual norm $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$. Thus the L-curve clearly displays the relationship between minimization of the quantities $\|\mathrm{L}\mathbf{x}\|_2^2$ and $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2$, which is the basic idea for any regularization method. For discrete ill-posed problems, it turns out that the L-curve has the characteristic L-shaped appearance, hence it's name. This shape has a distinct corner separating the vertical and the horizontal parts of the curve.

The Tikhonov L-curve for regularization plays an important role in regularization of discrete ill-posed problems. It's divides the first quadrant into two regions. Thus, with the Tikhonov L-curve, any regularized solution must lie on or above the curve. The L-curve basically consist of a vertical part and an adjacent horizontal part. The horizontal part correspond to over-smoothed solution where the regularization parameter is too large and the solution is dominated by regularization errors. The vertical part corresponds to under-regularized solutions where the regularization parameter is too small and the solution is dominated by perturbation errors.

It is important to plot the L-curve in log-log scale in order to emphasize the two different parts of the curve. The purpose of this, is to show the behavior of the L-curve which is more easily seen in such a log-log scale. In addition, the log-log scale emphasizes "flat" parts of the L-curve where the variation in $\|\mathbf{A}\mathbf{x}_{\mathrm{reg}}-\mathbf{b}\|$ is small compared to the variation in other variables. Also the log-log scale helps in scaling of $\mathbf{x}$ and $\mathbf{b}$ which simply shift the L-curve vertically and horizontally.

The Discrepancy Principle

The Discrepancy Principle, attributed to Morozov by Wing and Zahart(1991), is the most widespread error based method for regularizing ill-conditioned linear systems. Thus, if a computational and data error of an ill-posed problem is represented by a term of size ϵ, then the solution $\mathbf{x}$ satisfies the relation

$$\|\mathbf{Ax} - \mathbf{b}\| \leq \epsilon, \tag{4.29}$$

which is very close to the exact solution and therefore acceptable.

Here, the idea is simple, to choose the regularization parameter such that the residual norm is equal to the upper bound(ϵ). Normally it's not too difficult to find an acceptable solution to the problem. But if the problem is ill-conditioned, it becomes necessary to choose a regularization matrix $\mathbf{B}$, such that $\|\mathbf{Bx}\|$ is very large for all undesirable $\mathbf{x}$. We formalize this by making sure that a solution $\mathbf{x}$ is plausible only if $\|\mathbf{Bx}\| \leq \mathbf{M}$, where $\mathbf{M}$ is a chosen positive number. We then select a regularization parameter (λ) such that the regularized solution ($\mathbf{x}_\lambda$) is both acceptable and plausible.

Generally this requirement is not enough to give a unique solution, and some more choices are necessary. One of such choices is to minimize the residual

$$\rho(\mathbf{x}) = \|\mathbf{Ax} - \mathbf{b}\| \ , \tag{4.30}$$

subject to

$$\|\mathbf{Bx}\| \leq \mathbf{M}. \tag{4.31}$$

This is a constrained optimization problem, whose solution is the best plausible solution. Alternatively, we can find a solution $\mathbf{x}$ that minimizes the functional

$$\phi_{\mathbf{B}} = \|\mathbf{Bx}\|, \tag{4.32}$$

subject to

$$\|\mathbf{Ax} - \mathbf{b}\| \leq \epsilon. \tag{4.33}$$

This is referred to as the most plausible acceptable solution. The extremum of the functional to be minimized is unique and occurs on the boundary defined by the constraints. The minimization of the residual in 4.30 subject to 4.31 can be replaced by finding λ and $\mathbf{x}_\lambda$ such that

$$(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{B}^T\mathbf{B})\mathbf{x}_\lambda = \mathbf{A}^T\mathbf{b} \ , \tag{4.34}$$

with

$$\|\mathbf{Bx}_\lambda\| = \mathbf{M}. \tag{4.35}$$

Also the minimization problem in Equation 4.34 and 4.35 is equivalent to solving

$$(\mathbf{A}^T\mathbf{A} + \lambda\mathbf{B}^T\mathbf{B})\mathbf{x}_\lambda = \mathbf{A}^T\mathbf{b}, \tag{4.36}$$

subject to

$$\|\mathbf{Ax}_\lambda - \mathbf{b}\| = \epsilon. \tag{4.37}$$

Thus in practice, if we have a value for $\mathbf{M}$, we can solve Equations 4.34 repeatedly with various values of the regularization parameter (λ) until we find a suitable value for λ such that Equation 4.35 is satisfied. This process is repeated for Equations 4.36 and 4.37, but starting with a given value of ϵ which is easier to obtain than $\mathbf{M}$. The second method is often preferred to the first, and for a given λ satisfied by Equation 4.37, is referred to as the **Morozov Discrepancy Principle.**

CHAPTER FIVE

APPLICATION TO THE SOLUTION OF FREDHOLM INTEGRAL EQUATION OF THE FIRST KIND

Integral Equations

An integral equation is an equation for an unknown function f , where f appears under the integral sign. The regular occurrence of various types of these equations that exhibit ill-conditioned nature analytically is sufficient to explore them to find a solution to it. Integral equations basically occurs in two forms, namely Linear and Non-Linear integral equations, by Hobson(1998).

However, under this study we will restrict ourselves to only Linear integral equations, which has the general form:

$$g(x)f(x) = y(x) + \lambda \int_a^b k(x,z)f(z)dz. \tag{5.1}$$

The function f($\mathbf{x}$) is an unknown function, while the function y(x) , g(x) and k(x, z) are assumed known. k(x, z) is called the kernel of the integral equation. The integral limits a and b are also assumed known and constants, with λ being a known constant or parameter.

Types of Integral Equations

The Linear integral Equation in 5.1 generate various kinds of equations if the functions g and f are altered. If g(x) = 0, the unknown function f appears only under the integral sign and the equation obtained is called

the Linear integral equation of the first kind. Alternatively, if $g(x) = 1$, f appears twice, one inside the integral sign and the other outside. The integral equation obtained here is also called integral equation of the second kind. If $y(x) = 0$, the equation is called homogeneous and if $y(x) \neq 0$, it is called inhomogeneous. Integral equations of the first and second kind can further be distinguished by the form of the integration limits "a" and "b". If the limits are fixed constant, then the equation is called Fredholm equation.

However, if the upper limit $b = x$, that is a variable, the equation is called Volterra equation. The Volterra equation is analogous to one with fixed limits, but for which $k(x, z) = 0$ for $z > x$. Also for cases where either or both limits of integration are infinite, or for which $k(x, z)$ is infinite in a given range of integration, the equation is called Singular Integral Equation.

Examples of Linear Integral equations where g is a given function and f the function to be determined, are illustrated below.

1. Linear Fredholm Integral equation of the First kind.

$$g(x) = \int_a^b k(x, z)y(z)dz \text{ for } x \in [a, b] \tag{5.2}$$

2. Linear Fredholm Integral equation of the Second kind.

$$f(x) = g(x) + \int_a^b k(x, z)y(z)dz \text{ for } x \in [a, b] \tag{5.3}$$

3. Linear Volterra Integral equation of the First kind.

$$g(x) = \int_a^x k(x, z)y(z)dz \text{ for } x > a \tag{5.4}$$

4. Linear Volterra Integral equation of the Second kind.

$$f(x) = g(x) + \int_a^x k(x, z)y(z)dz \text{ for } x > a \tag{5.5}$$

Fredholm Integral Equation as an Ill-Posed or Inverse Problem

Fredholm Integral equations of the first and second kind are typical examples of ill-posed or inverse problems. These problems arises naturally in determining the natural structure of the physical systems, from the system's measured behavior, or in determining the unknown input that gives rise to a measured output signal.

Suppose a linear inverse problem is to be formulated to compute an output signal, given the input signal and a mathematical description of the problem, then we can formulate it as :

$$\int_{\Gamma} \text{Input} \times \text{System} = \text{Output}.$$

In order to be able to solve such problem, it is appropriate to discretized the system into a linear equation .

Example 5.1. Consider the integral equation

$$\int_0^1 \sin(\mathrm{nt})\mathrm{dt} = \frac{1}{-\mathrm{n}}\cos(\mathrm{nt}).$$

This is an example of Fredholm integral equation of the first kind with
$\mathrm{k(x,y)} = 1$, $\mathrm{f(y)} = \sin(\mathrm{nt})$, $\mathrm{a} = \mathrm{o}$ and $\mathrm{b} = 1$.
The function f(y) of order unity produces a function g(x) of order $\frac{1}{-n}$. This is true for large n, since a small change on the right hand side of the integral equation $\mathrm{g(x)} = \int_{\mathrm{a}}^{\mathrm{b}} \mathrm{k(x,y)f(y)dy}$
for $x \in [a, b]$, can create an undesirable effect on the left hand side, thus making the system ill-conditioned.

Example 5.2. Consider also the Fredholm integral equation of second kind given by

$$\mathrm{y(x)} = \mathrm{x} + \lambda \int_0^1 (\mathrm{xz} + \mathrm{z}^2)\mathrm{y(z)dz}.$$

The kernel for the equation is $\mathrm{k(x,z)} = \mathrm{x(z+z^2)}$, which is clearly separable. Using Fredholm integral equation of the second kind with separable or degenerate kernel of the form

$$\mathrm{k(x,y)} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \phi_{\mathrm{i}}(\mathrm{x})\psi_{\mathrm{i}}(\mathrm{z}), \quad \text{where} \quad \phi_1(\mathrm{x}) = \mathrm{x},\ \phi_2(\mathrm{x}) = 1,\ \psi_1(\mathrm{x}) = \mathrm{z},$$

$\psi_2(\mathrm{x}) = \mathrm{z}^2$, and writing the kernel in its separated form, the function $\phi_{\mathrm{i}}(\mathrm{x})$ may be taken outside the integral over z to obtain

$$\mathrm{y(x)} = \mathrm{f(x)} + \lambda \sum_{\mathrm{i=1}}^{\mathrm{n}} \phi_{\mathrm{i}}\mathrm{x} \int_{\mathrm{a}}^{\mathrm{b}} \psi_{\mathrm{i}}(\mathrm{z})\mathrm{y(z)dz} \quad \text{for} \quad \mathrm{x} \in [\mathrm{a,b}].$$

Since the integration limits a and b are constant for a Fredholm equation, the integral over z in each term of the sum is a constant. Denoting the constant by

$$\mathrm{c_i} = \int_{\mathrm{a}}^{\mathrm{b}} \psi_{\mathrm{i}}(\mathrm{z})\mathrm{y(z)dz},$$

the solution to the equation now becomes

$$\mathrm{y(x)} = \mathrm{f(x)} + \lambda \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{c_i}\phi_{\mathrm{i}}(\mathrm{x}).$$

Hence the solution to the problem

$\mathrm{y(x)} = \mathrm{x} + \lambda \int_0^1 (\mathrm{xz} + \mathrm{z}^2)\mathrm{y(z)dz}$ is

$$\mathrm{y(x)} = \mathrm{x} + \lambda(\mathrm{c_1 x} + \mathrm{c_2}), \tag{5.6}$$

with c_1 and c_2 defined as

$$\mathrm{c_1} = \int_0^1 \mathrm{z}[\mathrm{z} + \lambda(\mathrm{c_1 z} + \mathrm{c_2})]\mathrm{dz} = \frac{1}{3} + \frac{1}{3}\lambda \mathrm{c_1} + \frac{1}{2}\lambda \mathrm{c_2},$$

and

$$\mathrm{c_2} = \int_0^1 \mathrm{z}^2[\mathrm{z} + \lambda(\mathrm{c_1 z} + \mathrm{c_2})]\mathrm{dz} = \frac{1}{4} + \frac{1}{4}\lambda \mathrm{c_1} + \frac{1}{3}\lambda \mathrm{c_2}.$$

That is

$$\mathrm{c_1} = \frac{1}{3} + \frac{1}{3}\lambda \mathrm{c_1} + \frac{1}{2}\lambda \mathrm{c_2}, \tag{5.7}$$

and

$$c_2 = \frac{1}{4} + \frac{1}{4}\lambda c_1 + \frac{1}{3}\lambda c_2. \tag{5.8}$$

Solving the Equations 5.7 and 5.8 simultaneously for c_1 and c_2, we have

$$c_1 = \frac{24 + \lambda}{72 - 48\lambda - \lambda^2} \text{ and } c_2 = \frac{18}{72 - 48\lambda - \lambda^2} \ .$$

Substituting c_1 and c_2 into Equation 5.6, the solution becomes

$$y(x) = \frac{(72 - 24\lambda)x + 18}{72 - 48\lambda - \lambda^2} \ .$$

Hence the solution y(x) has a finite unique solution when λ is not equal to either roots. If λ is equal to either roots of the quadratic in the denominator, the solution becomes increasingly ill-conditioned.

Numerical Solution of Fredholm Integral Equation of the First Kind

When a rank-deficient or ill-posed problem is discretized, the inherent problem in the coefficient matrix decay gradually to zero. Thus in practice, it is important to discretize integral equations in other to solve them numerically. Although, there are many ways of discretizing integral equations or ill-posed problems, two main methods for discretizing integral equations are the Quadrature methods and the Galerkin methods, by Hamming(1989) and Hansen(1998). Both methods compute an approximation f to f. In the Quadrature method, a quadrature rule with abscissas $t_1, t_2, t_3, ..., t_n$, and corresponding weights $w_1, w_2, w_3, ..., w_n$, can be used to approximate an integral equation. The approximated integral equation is

$$\int_0^1 \phi(t)dt = \sum_{j=1}^{n} w_j \phi(t_j). \tag{5.9}$$

If the approximation is applied to an integral equation with m distinct point values $s_1, s_2, s_3, ...s_m$, we obtain an m × n matrix $\mathbf{A}$ given by $a_{ij} = w_j k(s_i, t_j)$

and a right hand side vector $\mathbf{b}$ also by $b_i = g(s_i)$. The equation $\mathbf{Ax} = \mathbf{b}$ becomes:

$$g(s_i) = \sum_{j=1}^{n} w_j k(s_i, t_j) x(t_j). \tag{5.10}$$

To make this equation a finite solvable system, we modify the equation by satisfying it exactly at the quadrature point to obtain the new equation.

$$g(t_i) = \sum_{j=1}^{n} k(t_i, t_j) x_j, \tag{5.11}$$

where $x_1, x_2, x_3, ..., x_n$ are unknown parameters for $i = 1, 2, 3...n$.
In the Galerkin method, we chose two sets of basis function ϕ_i and ψ_j such that the matrix $\mathbf{A}$ and the vector $\mathbf{b}$ can be defined as

$$a_{ij} = \int_0^1 k(s,t)\phi_i(s)\psi_j ds dt, \quad \text{and} \quad b_i = \int_0^1 g(s)\phi_i(s) ds.$$

Solving the equation $\mathbf{Ax} = \mathbf{b}$, for the vector $\mathbf{x}$, we obtain the solution

$$f(t) = \sum_{j=1}^{n} \mathbf{x}_i \psi_i(t) \ ,$$

which satisfy the integral Equation 5.9. If the kernel k is symmetric and the two set of basis functions are equal, then the matrix $\mathbf{A}$ is symmetric and the Galerkin's method is called Rayleigh Ritz Method.

Example 5.3. Consider the integral equation

$$\int_0^1 e^{(s+1)t} x(t) dt = \frac{e^{s+1} - 1}{s+1}. \tag{5.12}$$

To solve the equation numerically, we discretized the equation into a linear system

$$\sum w_j k(s_i, t_j) x_j = g(s_i),$$

by using Equation 5.10, with an n-point composite trapezoidal rule and with uniformly spaced quadrature points.
From 5.10, the kernel of the integral equation is

$k(t_i, t_j) = e^{(t_i+1)t_j}$, with it's weight function as w_j, the computed solution as $\mathbf{x}_j$, and the function $g(t_i)$ also given by

$$g(t_i) = \frac{e^{(t_i+1)} - 1}{t_i + 1} \ .$$

Where $s = t_i$, $t = t_j$ for $i = 1, 2, 3, ..., n$ and $j = 1, 2, 3, ..., n$.

Thus, for a five point composite trapezoidal rule, we divide the integral limits from 0 to 1 into a number of strip's of size $n = 4$, to obtain the five points. The Equation 5.10 is further expressed as:

$$\begin{aligned}
y(t_1) &= w_1k(t_1,t_1)x_1 + w_2k(t_1,t_2)x_2 + w_3k(t_1,t_3)x_3 + \\
&\quad w_4k(t_1,t_4)x_4 + w_5k(t_1,t_5)x_5 \\
y(t_2) &= w_1k(t_2,t_1)x_1 + w_2k(t_2,t_2)x_2 + w_3k(t_2,t_3)x_3 + \\
&\quad w_4k(t_2,t_4)x_4 + w_5k(t_2,t_5)x_5 \\
y(t_3) &= w_1k(t_3,t_1)x_1 + w_2k(t_3,t_2)x_2 + w_3k(t_3,t_3)x_3 + \\
&\quad w_4k(t_3,t_4)x_4 + w_5k(t_3,t_5)x_5 \\
y(t_4) &= w_1k(t_4,t_1)x_1 + w_2k(t_4,t_2)x_2 + w_3k(t_4,t_3)x_3 + \\
&\quad w_4k(t_4,t_4)x_4 + w_5k(t_4,t_5)x_5 \\
y(t_5) &= w_1k(t_5,t_1)x_1 + w_2k(t_5,t_2)x_2 + w_3k(t_5,t_3)x_3 + \\
&\quad w_4k(t_5,t_4)x_4 + w_5k(t_5,t_5)x_5.
\end{aligned}$$

Using the Composite Trapezoidal rule with it's weight functions defined as $w_1 = w_5 = 1/8$ and $w_2 = w_3 = w_4 = 1/4$, at the quadrature points $t_1 = 0,\ t_2 = 0.25,\ t_3 = 0.50,\ t_4 = 0.75,$ and $t_5 = 1.0$ respectively, the five equations represented by $y(t_1),\ y(t_2),\ y(t_3),\ y(t_4)$ and $y(t_5)$ further becomes:

$$\frac{1}{8}x_1 + \frac{1}{4}x_2e^{1/4} + \frac{1}{4}x_3e^{1/2} + \frac{1}{4}x_4e^{3/4} + \frac{1}{8}x_5e^1 = e^1 - 1, \qquad (5.13)$$

$$\frac{1}{8}x_1 + \frac{1}{4}x_2 e^{5/16} + \frac{1}{4}x_3 e^{5/8} + \frac{1}{4}x_4 e^{5/16} + \frac{1}{8}x_5 e^{5/4} = \frac{e^{1.25}-1}{1.25}, \tag{5.14}$$

$$\frac{1}{8}x_1 + \frac{1}{4}x_2 e^{3/8} + \frac{1}{4}x_3 e^{3/4} + \frac{1}{4}x_4 e^{9/8} + \frac{1}{8}x_5 e^{3/2} = \frac{e^{1.5}-1}{1.5}, \tag{5.15}$$

$$\frac{1}{8}x_1 + \frac{1}{4}x_2 e^{7/16} + \frac{1}{4}x_3 e^{7/8} + \frac{1}{4}x_4 e^{21/16} + \frac{1}{8}x_5 e^{7/4} = \frac{e^{1.75}-1}{1.75}, \tag{5.16}$$

$$\frac{1}{8}x_1 + \frac{1}{4}x_2 e^{1/2} + \frac{1}{4}x_3 e^{1} + \frac{1}{4}x_4 e^{3/2} + \frac{1}{8}x_5 e^{2} = \frac{e^{2}-1}{2}. \tag{5.17}$$

Thus discretizing the Equations 5.13 to 5.17 into a linear system $\mathbf{Ax} = \mathbf{b}$, we have

$$\begin{bmatrix} \frac{1}{8} & \frac{1}{4}e^{1/4} & \frac{1}{4}e^{1/2} & \frac{1}{4}e^{3/4} & \frac{1}{8}e^{1} \\ \frac{1}{8} & \frac{1}{4}e^{5/16} & \frac{1}{4}e^{5/8} & \frac{1}{4}e^{15/16} & \frac{1}{8}e^{5/4} \\ \frac{1}{8} & \frac{1}{4}e^{3/8} & \frac{1}{4}e^{3/4} & \frac{1}{4}e^{9/8} & \frac{1}{8}e^{3/2} \\ \frac{1}{8} & \frac{1}{4}e^{7/16} & \frac{1}{4}e^{7/8} & \frac{1}{4}e^{21/16} & \frac{1}{8}e^{7/4} \\ \frac{1}{8} & \frac{1}{4}e^{1/2} & \frac{1}{4}e^{1} & \frac{1}{4}e^{3/2} & \frac{1}{8}e^{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} (e^1-1) \\ \frac{(e^{1.25}-1)}{1.25} \\ \frac{(e^{1.5}-1)}{1.5} \\ \frac{(e^{1.75}-1)}{1.75} \\ \frac{(e^2-1)}{2} \end{bmatrix},$$

which can further be simplified as

$$\begin{bmatrix} 0.125 & 0.321006 & 0.412180 & 0.529250 & 0.339785 \\ 0.125 & 0.341709 & 0.467061 & 0.638397 & 0.436292 \\ 0.125 & 0.363748 & 0.529250 & 0.770054 & 0.560211 \\ 0.125 & 0.387208 & 0.599719 & 0.928863 & 0.719325 \\ 0.125 & 0.412180 & 0.679570 & 1.120422 & 0.923632 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1.718282 \\ 1.992274 \\ 2.321126 \\ 2.716915 \\ 3.194528 \end{bmatrix}.$$

Here $\mathbf{A}$ is a 5 by 5 square matrix and $\mathbf{b}$ a 5 by 1 vector. If the system $\mathbf{Ax} = \mathbf{b}$ is solved using any of the standard methods, we have $x_1 = 0.56015$, $x_2 = 1.51820$, $x_3 = 0.42125$, $x_4 = 1.4807$, and $x_5 = 0.59914$. The computed solution differs from the exact solution but shows a little resemblance.

If the integral equation in Example 5.3 is repeated for different values of n, for instance, for n = 8, we obtain

$$\begin{aligned} e^1 - 1 \;=\; & \frac{1}{16}x_1 + \frac{1}{8}x_2 e^{1/8} + \frac{1}{8}x_3 e^{1/4} + \frac{1}{8}x_4 e^{3/8} + \frac{1}{8}x_5 e^{1/2} + \frac{1}{8}x_6 e^{5/8} + \\ & \frac{1}{8}x_7 e^{3/4} + \frac{1}{8}x_8 e^{7/8} + \frac{1}{16}x_9 e^{1}, \end{aligned}$$

$$\frac{e^{9/8}-1}{9/8} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{9/64}+\frac{1}{8}x_3e^{9/32}+\frac{1}{8}x_4e^{27/64}+\frac{1}{8}x_5e^{9/16}+\frac{1}{8}x_6e^{45/64} +\frac{1}{4}x_7e^{27/32}+\frac{1}{8}x_8e^{63/64}+\frac{1}{16}x_9e^{9/8},$$

$$\frac{e^{5/4}-1}{5/4} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{5/32}+\frac{1}{8}x_3e^{5/16}+\frac{1}{8}x_4e^{15/32}+\frac{1}{8}x_5e^{5/8}+\frac{1}{8}x_6e^{25/32} +\frac{1}{8}x_7e^{15/16}+\frac{1}{8}x_8e^{35/32}+\frac{1}{16}x_9e^{5/4},$$

$$\frac{e^{11/8}-1}{11/8} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{11/64}+\frac{1}{8}x_3e^{11/32}+\frac{1}{8}x_4e^{33/64}+\frac{1}{8}x_5e^{11/16}+\frac{1}{8}x_6e^{35/64} +\frac{1}{8}x_7e^{33/32}+\frac{1}{8}x_8e^{77/64}+\frac{1}{16}x_9e^{11/8},$$

$$\frac{e^{3/2}-1}{3/2} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{3/16}+\frac{1}{8}x_3e^{3/8}+\frac{1}{8}x_4e^{9/16}+\frac{1}{8}x_5e^{3/4}+\frac{1}{8}x_6e^{15/16} +\frac{1}{8}x_7e^{9/8}+\frac{1}{8}x_8e^{21/16}+\frac{1}{16}x_9e^{3/2},$$

$$\frac{e^{13/8}-1}{13/8} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{13/64}+\frac{1}{8}x_3e^{13/32}+\frac{1}{8}x_4e^{39/64}+\frac{1}{8}x_5e^{13/16}+\frac{1}{8}x_6e^{65/64} +\frac{1}{8}x_7e^{39/32}+\frac{1}{8}x_8e^{91/64}+\frac{1}{16}x_9e^{13/8},$$

$$\frac{e^{7/4}-1}{7/4} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{7/32}+\frac{1}{8}x_3e^{7/16}+\frac{1}{8}x_4e^{21/32}+\frac{1}{8}x_5e^{7/8}+\frac{1}{8}x_6e^{35/32} +\frac{1}{8}x_7e^{21/16}+\frac{1}{8}x_8e^{49/32}+\frac{1}{16}x_9e^{7/4},$$

$$\frac{e^{15/8}-1}{15/8} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{15/64}+\frac{1}{8}x_3e^{15/32}+\frac{1}{8}x_4e^{45/64}+\frac{1}{8}x_5e^{15/16}+\frac{1}{8}x_6e^{75/64} +\frac{1}{8}x_7e^{45/32}+\frac{1}{8}x_8e^{105/64}+\frac{1}{16}x_9e^{15/8},$$

$$\frac{e^2-1}{2} = \frac{1}{16}x_1+\frac{1}{8}x_2e^{1/4}+\frac{1}{8}x_3e^{1/2}+\frac{1}{8}x_4e^{3/4}+\frac{1}{8}x_5e^{1}+\frac{1}{8}x_6e^{5/4}+ \frac{1}{8}x_7e^{3/2}+\frac{1}{8}x_8e^{7/4}+\frac{1}{16}x_9e^{2}.$$

Discretizing the nine equations above into a linear system $\mathbf{Ax} = \mathbf{b}$, our matrix $\mathbf{A}$ becomes

$$\begin{bmatrix}
\frac{1}{16} & \frac{1}{8}e^{1/8} & \frac{1}{8}e^{1/4} & \frac{1}{8}e^{3/8} & \frac{1}{8}e^{1/2} & \frac{1}{8}e^{5/8} & \frac{1}{8}e^{3/4} & \frac{1}{8}e^{7/8} & \frac{1}{16}e^{1} \\
\frac{1}{16} & \frac{1}{8}e^{9/64} & \frac{1}{8}e^{9/32} & \frac{1}{8}e^{27/64} & \frac{1}{8}e^{9/16} & \frac{1}{8}e^{45/64} & \frac{1}{4}e^{27/32} & \frac{1}{8}e^{63/64} & \frac{1}{16}e^{9/8} \\
\frac{1}{16} & \frac{1}{8}e^{5/32} & \frac{1}{8}e^{5/16} & \frac{1}{8}e^{15/32} & \frac{1}{8}e^{5/8} & \frac{1}{8}e^{25/32} & \frac{1}{8}e^{15/16} & \frac{1}{8}e^{35/32} & \frac{1}{16}e^{5/4} \\
\frac{1}{16} & \frac{1}{8}e^{11/64} & \frac{1}{8}e^{11/32} & \frac{1}{8}e^{33/64} & \frac{1}{8}e^{11/16} & \frac{1}{8}e^{35/64} & \frac{1}{8}e^{33/32} & \frac{1}{8}e^{77/64} & \frac{1}{16}e^{11/8} \\
\frac{1}{16} & \frac{1}{8}e^{3/16} & \frac{1}{8}e^{3/8} & \frac{1}{8}e^{9/16} & \frac{1}{8}e^{3/4} & \frac{1}{8}e^{15/16} & \frac{1}{8}e^{9/8} & \frac{1}{8}e^{21/16} & \frac{1}{16}e^{3/2} \\
\frac{1}{16} & \frac{1}{8}e^{13/64} & \frac{1}{8}e^{13/32} & \frac{1}{8}e^{39/64} & \frac{1}{8}e^{13/16} & \frac{1}{8}e^{65/64} & \frac{1}{8}e^{39/32} & \frac{1}{8}e^{91/64} & \frac{1}{16}e^{13/8} \\
\frac{1}{16} & \frac{1}{8}e^{7/32} & \frac{1}{8}e^{7/16} & \frac{1}{8}e^{21/32} & \frac{1}{8}e^{7/8} & \frac{1}{8}e^{35/32} & \frac{1}{8}e^{21/16} & \frac{1}{8}e^{49/32} & \frac{1}{16}e^{7/4} \\
\frac{1}{16} & \frac{1}{8}e^{15/64} & \frac{1}{8}e^{15/32} & \frac{1}{8}e^{45/64} & \frac{1}{8}e^{15/16} & \frac{1}{8}e^{75/64} & \frac{1}{8}e^{45/32} & \frac{1}{8}e^{105/64} & \frac{1}{16}e^{15/8} \\
\frac{1}{16} & \frac{1}{8}e^{1/4} & \frac{1}{8}e^{1/2} & \frac{1}{8}e^{3/4} & \frac{1}{8}e^{1} & \frac{1}{8}e^{5/4} & \frac{1}{8}e^{3/2} & \frac{1}{8}e^{7/4} & \frac{1}{16}e^{2}
\end{bmatrix},$$

and that of the right-hand vector $\mathbf{b}$, and the solution vector $\mathbf{x}$ becomes

$$\begin{bmatrix} 1.7183 & 1.8491 & 1.9923 & 2.1492 & 2.3211 & 2.5098 & 2.7169 & 2.9444 & 3.1945 \end{bmatrix}^T$$

and $\begin{bmatrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 & x_8 & x_9 \end{bmatrix}^{\mathrm{T}}$ respectively .

Solving the linear system of equation $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a 9 by 9 square matrix and $\mathbf{b}$ a 9 by 1 vector, we obtain the computed solution $\hat{\mathbf{x}}$. The values of the computed solution is said to deviate the more from the exact solution. The Table 14 below shows the computed and the exact solutions for some selected values of n.

It is observed that the computed solution has contaminated errors compared to that of the exact solution. While for small values of n, the computed solution has some resemblance to the true or exact solution, the error increases rapidly with n, and we cannot get any value of the computed solution that approximates the exact solution with acceptable accuracy. The reason for this anomaly is quite clear when we look at the condition numbers of some selected values of n, shown in the Table 15 below.

Table 14: Computed Solutions(C.S) for n-point Composite Trapezoidal Rule.

Points	C.S	C.S	C.S	Exact Solution
t	$n = 4$	$n = 8$	$n = 16$	Exact
0.0000	0.56015	0.5306	−47.3089	1.0000
0.1250	–	1.7584	−316.2421	1.0000
0.2500	1.51820	−0.5601	102.0745	1.0000
0.3750	–	3.4871	49.3657	1.0000
0.5000	0.42125	−1.8604	−8.5007	1.0000
0.6250	–	3.3715	−24.4726	1.0000
0.7500	1.4807	−0.4432	−33.7697	1.0000
0.8750	–	1.7075	30.1161	1.0000
1.0000	0.59914	0.5480	3.0331	1.0000

Table 15: Condition Numbers for Selected Values of n.

n	Condition Number
4	3.2845×10^5
8	1.7263×10^7
16	5.7636×10^{15}

From the table, as n increases from 4 onward, we expect the discretization error to reduce, that is the accuracy with which Equation 5.12 represents 5.10 gets better. Unfortunately this is not so, the condition numbers rather increases rapidly destroying any gain from the more accurate discretization. This makes it quite clear that any computations with n greater than sixteen will make the solution worse.

Applying Regularization Methods in Solving Problem 5.3

Generally, when we solve ill-posed problems by discretizing the equation, the matrix we get is inherently ill-conditioned, and the condition numbers increases rapidly with n. Thus, there is the need to apply some form of regularization in order to obtain reasonable approximations. Basically, there are two main approaches to solving ill-posed problems. The first approach is to regularize the problem to obtain a well-posed equation and then discretized. This has many theoretical advantages but poses some problem. The second approach is to discretized first, and then apply regularization methods to the resulting finite systems. The second approach is better and produce satisfactory results.

When order zero, one and two was applied to Example 5.3 for $n = 16$, the solutions for order one and two converges approximately to the exact solution, but that of order zero shows little convergence to the exact solution. The detailed solutions for order zero, one and two, for a given range of values of the regularization parameter λ can be seen in Appendix C. These convergent solutions are not the same as the optimal solution. Rather, it gives us a fair idea of the nature of the optimal solution. The respective convergent solutions for order zero, one and two are shown in Table 16 below.

Table 16: Convergent Regularized Solutions for n = 16.

Order Zero	Order One	Order Two	Exact
0.40206	0.99886	0.9984229	1.0000
0.84232	0.99886	0.9984659	1.0000
0.88038	0.99886	0.9985090	1.0000
0.91211	0.99886	0.9985521	1.*oooo*
0.95428	0.99886	0.9985951	1.0000
0.98904	0.99886	0.9986382	1.0000
1.00440	0.99886	0.9986812	1.0000
1.05060	0.99886	0.9987243	1.0000
1.07560	0.99886	0.9987673	1.0000
1.09520	0.99886	0.9988104	1.0000
1.10810	0.99886	0.9988535	1.0000
1.11260	0.99886	0.9988965	1.0000
1.10710	0.99886	0.9989396	1.0000
1.08930	0.99886	0.9989826	1.0000
1.05670	0.99886	0.9990257	1.0000
1.01090	0.99886	0.9990687	1.0000
0.46762	0.99886	0.9991118	1.0000

Determination of Optimal Solution

In determining the optimal regularization parameter λ corresponding to the optimal solution, two methods can be applied. The first is by inspection and the second is the L-curve approach. For the method by inspection, we compute the residual norm, the solution norm and inspect the result for a range of values of the parameter λ . The behaviour of the residual norm gives a good indication of a proper choice of λ. If the residual norm increases or decreases steadily with λ, until a critical value λ_0, where further increases or decreases have no effect on the residual norm, then λ_0 is chosen as the optimal regularization parameter and its corresponding value becomes the optimal solution. But the method by inspection, is a times very difficult to use to locate the optimal regularization parameter.

However, it is more appropriate to use the L-curve method to determine the critical value for the optimal solution, since it helps us to identify clearly all the parameter points. Here, we plot the solution norm against the residual norm to obtain the L-curve. The optimal regularization parameter λ_0 is usually identified at the sharp corner of the L-curve.

Normally, we make use of the scatter plot of the L-curve to identify the points concentrated at the sharp corner, to enable us determine the optimal regularization parameter λ_0 . The optimal solution corresponding to the optimal regularization parameter λ_0 is obtained in column form in Appendix C. The first column correspond to $\lambda_1 = 10^{-16}$, the second $\lambda_2 = 10^{-15}$, and the rest continues in that order. The L-curve for order zero, one and two for the problem 5.3, are shown in the Figures 2, 3 and 4 below.

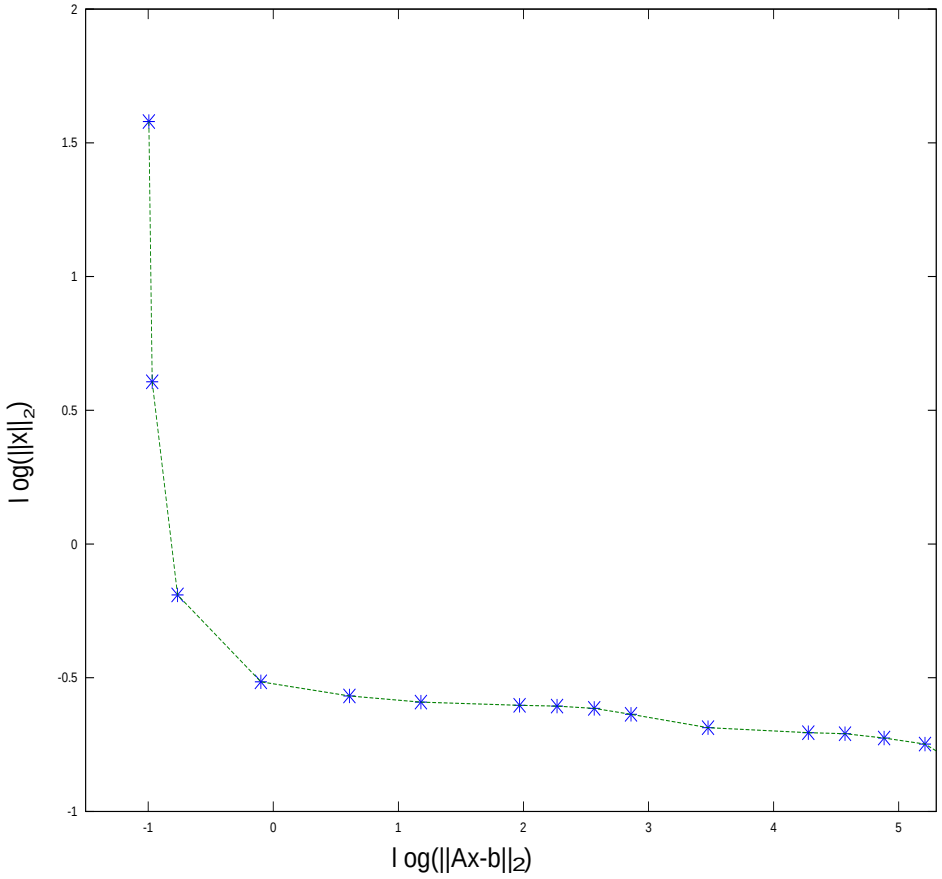

Figure 2: L-Curve for Order Zero Regularization

For order zero regularization, the regularization parameters concentrated at the sharp corner of the L-curve are $\lambda_3 = 10^{-14}$ and $\lambda_4 = 10^{-13}$. The regularized solutions corresponding to these parameters can be seen in Appendix C. These regularized solutions do not approximate in any way to the exact solutions. Therefore, order zero regularization in this case, has no optimal solution.

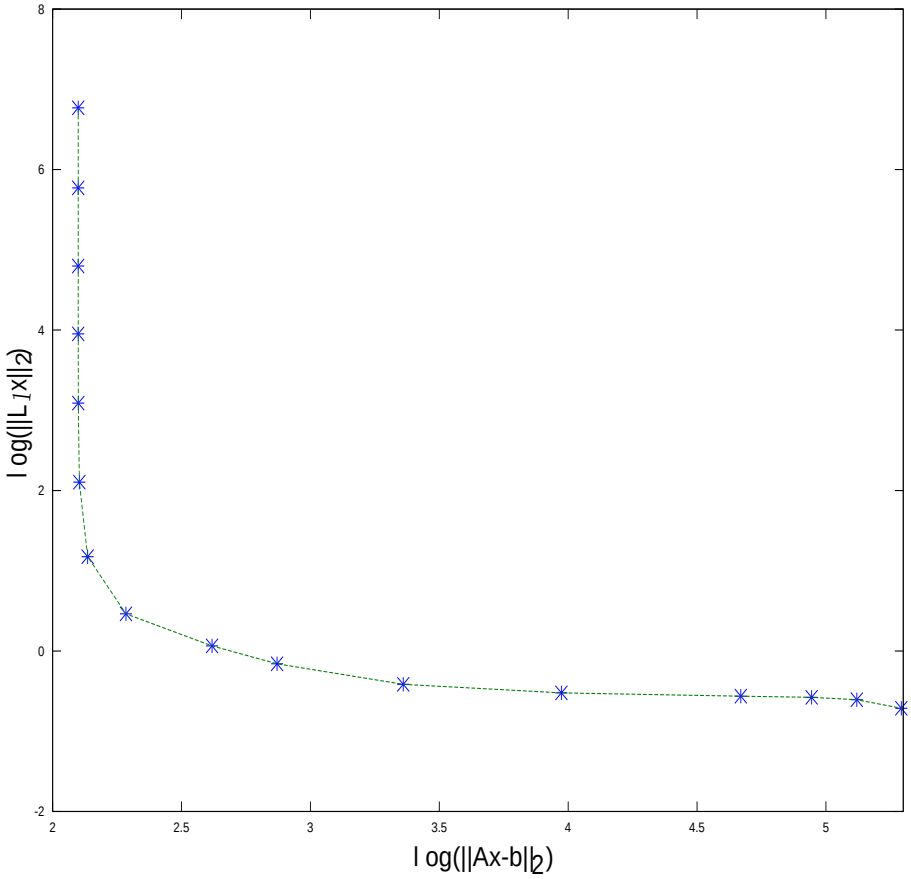

Figure 3: L-Curve for Order One Regularization

For order one regularization, the regularization parameters concentrated at the sharp corner of the L-curve are $\lambda_7 = 10^{-10}$ and $\lambda_8 = 10^{-9}$. The regularized solutions corresponding to these parameters in Appendix C do not approximate to the exact solutions. Hence, order one regularization for this problem has no optimal solution.

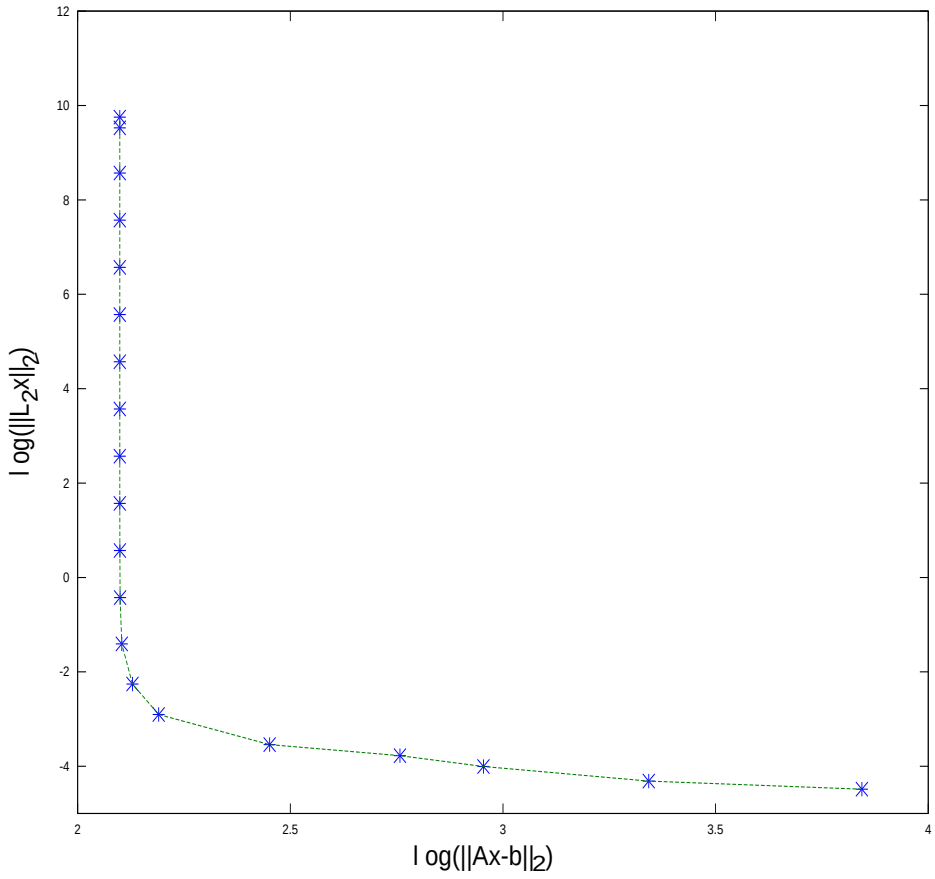

Figure 4: L-Curve for Order Two Regularization

For order two regularization, the regularization parameters concentrated at the corner of the L-curve are $\lambda_{12} = 10^{-5}$, $\lambda_{13} = 10^{-4}$ and $\lambda_{14} = 10^{-3}$ respectively. The regularized solutions corresponding to these parameters are in Appendix C. All the solutions approximate to the exact solution, but the entries of the solution for $\lambda_{13} = 10^{-4}$ are well skewed and better approximate to the exact solution compared to the solutions of $\lambda_{12} = 10^{-5}$ and $\lambda_{14} = 10^{-3}$.

Therefore the optimal regularization parameter is $\lambda_{13} = 10^{-4}$ and the corresponding optimal solution $\mathbf{x}_{\lambda_13}$ is shown in Table 17 below.

Table 17: The Optimal Solution

Exact solution $\mathbf{x}$	$\mathbf{x}_{\lambda_{13}} = 10^{-13}$
1.000000	0.9984226
1.000000	0.9984656
1.000000	0.9985087
1.000000	0.9985517
1.000000	0.9985947
1.000000	0.9986377
1.000000	0.9986807
1.000000	0.9987237
1.000000	0.9987667
1.000000	0.9988097
1.000000	0.9988527
1.000000	0.9988957
1.000000	0.9989387
1.000000	0.9989816
1.000000	0.9990245
1.000000	0.9990675
1.000000	0.9991094

In conclusion, the optimal solution for the problem approximate accurately to the exact solution, compared to the unregularized solution which shows no resemblance to the exact solution. The maximum error in the computed solution is reduced drastically to $1.6001e+01$ in the optimal solution.

CHAPTER SIX

Summary, Discussion, Conclusion and Recommendation

Summary

The focus of the thesis has been on the numerical regularization methods for ill-conditioned linear systems. The study examined the concept of Discrete ill-posed problems. Several observations were made in relation to the effect of perturbation of a linear system and the accuracy of a computed solution. The singular value decomposition is an important tool in numerical linear algebra for solving rank deficient or Discrete ill-posed problems. The singular value decomposition of a matrix shows that, it is the small singular values that contaminate the solution. Based on this analysis of the decomposition, certain numerical methods were employed to solve the problem. Some of the method employed are the Truncated Singular value Decomposition, Preconditioning and Tikhonov Regularization. The Tikhonov regularization methods was then applied to solve the Fredholm integral equation of the First kind. The method of Pre-conditioning was applied successfully to the regularization of the solution to a boundary-value problem. The truncated singular value decomposition was applied to the Hilbert system of order 12. A review of these methods shows that the Tikhonov Regularization method is the method of choice for regularizing rank deficient and discrete ill-posed problems.

Discussion

The goal to give a numerical treatment of efficient and reliable method for regularizing ill-conditioned or ill-posed problems have been thoroughly discussed in this thesis. The numerical solution of the Hilbert and the Vandermonde matrices which are practical examples of an ill-conditioned or ill-posed problems, attest to the fact that the accuracy of the solution of a linear system is very important and depends on the matrix $\mathbf{A}$ or the right hand vector $\mathbf{b}$. Thus in effect, if the entries of $\mathbf{A}$ and $\mathbf{b}$ of the linear system $\mathbf{Ax} = \mathbf{b}$, are accurate to about d-significant digits, and the condition number of $\mathbf{A}$ is approximately 10^{k}, then the computed solution $\hat{\mathbf{x}}$ is accurate to about $(\mathrm{d} - \mathrm{k})$ significant digits. Various standard method (LU-factorization, QR-factorization, Exact inverse, and the Cholesky factorization for positive index) were all applied to solve the problem but there was virtually no improvement in the computed solutions. This compels us to decomposed the matrix using singular value decomposition. The decomposition reveals to us that the error in the computed solution is as a results of the drastic effect of the small singular values. Certain regularization methods (Truncated Singular value Decomposition, Preconditioning and Tikhonov Regularization) for solving ill-posed problems were applied to correct the effect of the small singular values. The motive behind truncation is to replace the small nonzero singular values with exact zero.

Although this gives a better solution, it does not solve the problem totally. As we truncate the small singular values, the error in the solution reduces drastically and a better result obtained. Further truncation was done with the aim of obtaining a desired result, but the solution deteriorated again. We then survey the concepts of preconditioning .

Here we preconditioned the system using the Jacobi or Gauss-Seidel preconditioner. The condition number of the preconditioned matrix was better

than the original matrix. The reduction of the condition number of the system improves the accuracy of the solution and solve the problem to our satisfaction. The Tikhonov regularization method for order zero, one and two were used to regularized the problem. The various orders gave an insight into how stabilizing or regularizing effect of a particular regularization method influences the solution, and how the solution depends on the regularization parameter. To obtain a desired solution , we introduce the Parameter-Choice Methods (the L-curve method and the Discrepancy principle) for determining the optimal solution. The L-curve clearly displays the relationship between the minimization of the quantity $||\mathrm{L}\mathbf{x}||_2$ against $||\mathbf{A}\mathbf{x} - \mathbf{b}||_2$, which is the basic idea for any regularization method. The corner part of the curve helps us to determine the optimal solution.

Finally , we present a numerical result of the Fredholm integral equation of the first kind as an ill-conditioned problem. The integral equation was first discretized into a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ using the quadrature method. The problem was examined using Tikhonov regularization methods couple with the parameter-choice methods. Both methods proved to be very effective in determining the optimal solution.

Conclusion and Recommendation

Ill-conditioned linear systems can be solved using numerical regularization methods despite it's robustness. A classical example considered in this thesis to justify our claim was the Fredholm integral equation of the first kind. The discretization of the problem into a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$ shows how the matrix $\mathbf{A}$ was inherently ill-conditioned. Appropriate regularization method was employed to regularize the problem, after which the parameter-choice methods was used to obtain the optimal solution successfully.

In conclusion, we justify that in the absence of computational errors and

with proper regularization, a convergent discretization error leads to a convergent solution. One hopes that this work could later be extended further to solve practical problems such as: Image restoration problems, Image deblurring problems, Inverse Laplace transformation, inverse heat problems and Deconvolution (discontinuous solutions).

REFERENCES

Benyah , F. and Jennings, L. (1998) . *Regularization in Optimal Control Computations.* International Symposium on Intelligent Automation and Control, World Automation Congress, Anchorage, Alaska. pp. 78-135.

Benyah, F. and Jenning, L. (1998). *The L-Curve in the Regularization of Optimal Control Computation,* J. Austral.Math. Soc., Series BE-40, Sydney, Australia. pp.98-172.

Benyah , F. and Jennings, L .(2001). *A Rreview of Ill-Conditioning and Regularization in Optimal Control Computations, Optimization Methods and Applications,* (Xiaoqi-Yang, Kok Lay, and Lou Caccetta, eds.), vol. E40, Kluwer Academic Publishers, Boston, pp. 23- 44.

Benyah , F. (2005). *Lecture Notes on Numerical Linear Algebra* , university of Western Cape, Cape Town, pp. 70-119.

Goncharsky, A. and Bakushinsky, A. (1994). *Ill-posed problems: Theory and applications.* Kluwer Academic Publishers, Boston, pp. 1-19.

Hadamard, J. (1923). *Lectures on Cauchy Problem in Linear Partial Differential Equations.* Yale University Press, New Haven, pp. 74-105

Hamming, R. (1989). *Numerical methods for scientists and engineers* . Dover Press , New York , pp. 196-225.

Hansen, P. C. (1987). *Truncated Singular Value as a Method for Regularization.* BIT Review 27, Philadelphia, pp. 534-553.

Hansen, P. C. (1992). *Analysis of Discrete Ill-Posed Problems by means of the L-Curve.* SIAM Review 34, Philadelphia, pp. 561-580.

Hansen, P. C. (1997). *Rank-deficient and discrete ill-posed problems:Numerical aspects of linear inversion.* SIAM, Philadelphia, pp. 13-85.

Hobson, M. P. (1998). *Mathematical methods for physics and engineering.* Cambridge University Press, Cambridge , pp. 758-801.

Mathews, J. H. (1992). *Numerical methods for scientist and engineers.* Englewood Cliffs, Pentice Hall, New Jersey, pp. 10-112

Nair , T. M. and Pereverzev , V. S. (2006). *Regularized Collocation Method for Fredholm Integral equations of the first kind.* Complexity, Elsevier, pp. 1-14.

Wang, R. and Linz, P. (2003). *Exploring numerical methods: Introduction to scientific computing using matlab.* Jones and Bartlett Publishers, Sadbury, MA, pp. 402-405.

Wing, G. M. and Zahrt. J. D. (1991). *A prime on integral equations of the first kind.* SIAM, Philadelphia, pp. 191-204.

APPENDIX A

OCTAVE CODE FOR TRUNCATED SINGULAR VALUE DECOMPOSITION, JACOBI AND GAUSS-SEIDEL PRECONDITIONER'S

TSVD:

Function $\mathrm{x_{svd}} = \mathrm{TSVD(A, b, n)}$.

Format long;

$A =$ Hibert matrix of any order;

$[m, n] = \mathrm{size}(A)$;

$\mathrm{x} = \mathrm{ones}(n, 1)$;

$\mathrm{b} = \mathrm{A} * \mathrm{x}$

$[U, S, V] = \mathrm{svd}(A)$;

$\mathrm{s} = \mathrm{diag}(S)$;

for $i = 1 : n$

$\mathrm{x_{svd}(:, i)} = (\mathrm{U(:, i)}' * \mathrm{b/s(i)}) * \mathrm{V(:, i)}$;

endfunction.

$\mathrm{x_{svd}}$ = The Truncated Solution.

JACOBI:

Function $\mathrm{x_J} = \mathrm{precond}(A, b, n)$.

Format long;

$A =$ Sparse Matrix of a defined order;

$[m, n] = \mathrm{size}(A)$;

$\mathrm{x} = \mathrm{ones}(n, 1)$;

$\mathrm{b} = \mathrm{A} * \mathrm{x}$

$\mathrm{D} = \mathrm{diag(diag(A))}$;

$\mathrm{b1} = \mathrm{inv}(D) * b$;

$\mathrm{M} = \mathrm{inv}(D) * A$;

x_J = M b1;

endfunction.

x_J = The Preconditioned System.

NB: D is diagonal matrix and M is a preconditioned matrix.

GAUSS-SEIDEL:

Function x_G = precond(A, b, n).

Format long;

A = Sparse Matrix of a defined order;

[m, n] = size(A);

x = ones$(n, 1)$;

b = A $*$ x

D = diag(diag(A));

L = tril$(A, -1)$;

U = tril$(A, 1)$

b1 = inv$(L + D)^{-1} * A$;

M = inv$(L + D)^{-1} * A$;

x_G = M b1;

endfunction.

x_J = The Preconditioned System.

NB: U is an upper trianglar matrix, L a lower triangular matrix and M is a preconditioned matrix.

APPENDIX B

OCTAVE CODE FOR REGULARIZING ILL-CONDITIONED LINEAR SYSTEMS AND ITS RESULT

Function = Regularize($\mathbf{A}, \mathbf{b}, n$)

format long

$\mathbf{A}$ = hilb(12);

[m, n] = size($\mathbf{A}$);

$\mathbf{x}$ = ones(n, 1);

$\mathbf{b} = \mathbf{A} * \mathbf{x}$;

$\mathbf{A}t\mathbf{A} = \mathbf{A}' * \mathbf{A}$;

$\mathbf{A}t\mathbf{b} = \mathbf{A}' * \mathbf{b}$;

Create the regularization parameters $t = [10^{-16}, ..., 10^{3}]$.

t = zeros(20, 1);

t(1) = 10^{-16};

for $i = 1 : 19$

t(i + 1) = t(i) $*$ 10;

end

if order == 0

D = eye(n);

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg0 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg0(i, :) = ($\mathbf{A}$t + t(i) $*$ D) $\setminus$ $\mathbf{A}$t$\mathbf{b}$;

res0_norm(i) = norm($\mathbf{A} * \mathbf{x}$ − reg(:, i) − $\mathbf{b}$);

sol0_norm(i) = norm(D $*$ $\mathbf{x}$ − reg(:, i));

end.

$\mathbf{x}$_reg0

res0_norm = res0_norm;

sol0_norm = sol0_norm;

[res0_norm′ sol0_norm′]

end.

if order == 1

h = 1;

D1 = DiffOpr(1, n);

L1 = D1′ ∗ D1;

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg1 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg1(i, :) = ($\mathbf{A}$t + t(i) ∗ L1) \ $\mathbf{A}$t$\mathbf{b}$;

res1_norm(i) = norm($\mathbf{A}$ ∗ $\mathbf{x}$_reg(:, i) − $\mathbf{b}$);

sol1_norm(i) = norm(D1 ∗ $\mathbf{x}$_reg(:, i));

end.

$\mathbf{x}$_reg1

res1_norm = res1_norm;

sol1_norm = sol1_norm;

[res1_norm′ sol1_norm′]

end.

if order == 2

h_2 = 0.01;

D2 = DiffOpr(2, n);

B2 = (1/h^2) ∗ D2;

L2 = B2′ ∗ B2;

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg2 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg2(i, :) = ($\mathbf{A}$t + t(i) ∗ L2) \ $\mathbf{A}$t$\mathbf{b}$;

res2_norm(i) = norm($\mathbf{A} * \mathbf{x}\_reg(:, i) - \mathbf{b}$);

sol2_norm(i) = norm(D2 $* \mathbf{x}\_reg(:, i)$);

end.

$\mathbf{x}$_reg2

res2_norm = res2_norm;

sol2_norm = sol2_norm;

[res2_norm′ sol2_norm′]

end.

Reg0=

Columns 1 through 4:

0.99999558987635 0.99994129023543 0.99998956985858 0.99999501375901

0.00011484865408 1.00228840365990 1.00034842281497 1.00014037359154

0.99935815149903 0.97907402327410 0.99731501326943 0.99911916254794

1.00100811961802 1.07220199315852 1.00718065055867 1.00175691516681

1.00019117552870 0.89646910614765 0.99477692842340 0.99954281646107

0.99930647480609 1.07105307745711 0.99772593042956 0.99897772783852

0.99926911621657 0.89481974776813 0.99243789306522 0.99871647970600

0.99986924867098 1.15903339015223 1.01909696856479 1.00195214836498

1.00056880031464 0.99583522180340 1.00214131598349 1.00081607916663

1.00086976910633 0.88274830784933 0.98157967762364 0.99885014402034

1.00042159314940 1.02058294249397 1.00521494238652 1.00051427245869

0.99902212127644 1.02599988063476 1.00219456080894 0.99961477752346

Columns 5 through 8:

0.99999615796750 1.00000318399316 1.00004130287354 1.00006131485466

1.00010848696029 0.99998750395035 0.99932365732858 0.99878578966688

0.99935237030420 0.99976119914157 1.00199815830281 1.00428967063261

1.00110451016100 1.00086685938894 0.99971327688416 0.99814580846151

1.00008702233818 0.99983446257614 0.99809676133742 0.99607061590091

0.99927322529281 0.99911489105196 0.99843604336012 0.99783218666863

0.99915381983932 0.99933468952182 0.99991860111740 1.00085485188883

1.00007839949585 1.00015395147201 1.00146545297977 1.00324893413169

1.00065495922924 1.00088611046582 1.00226540354118 1.00401421008992

1.00075058038528 1.00105671618175 1.00185610317171 1.00276561106508

1.00045878461151 1.00037669386380 1.00003906203083 0.99947245551507

0.99897539434665 0.99860802813603 0.99677653144355 0.99428511343493

Columns 9 through 12:

0.99982759461036 0.99904344396020 0.99981057623282 1.00716677462353

1.00106279761382 1.00940333807493 1.00819859002563 0.97025534101148

1.00119904725954 0.98810959943146 0.98136775960316 0.98859493280536

0.99597555081377 0.98951959648705 0.98954127544851 1.01380648105732

0.99621823053867 0.99865463846272 1.00413559777337 1.02920091397204

0.99967042598279 1.00712409128657 1.01460372120102 1.03346328116079

1.00334123531302 1.01185940941616 1.01858139556544 1.02876949406568

1.00546849361466 1.01223102866891 1.01645835518161 1.01759012100603

1.00530320873138 1.00857502313142 1.00936535936713 1.00196081529613

1.00269365271897 1.00153959816867 0.99848281466878 0.98341117933417

0.99779150595072 0.99181247690929 0.98483040942298 0.96304966181660

0.99087962380398 0.98001700545144 0.96922767091293 0.94166610656110

$Reg1$=

Columns 1 through 4:

0.99984337030413 0.99996624696002 0.99999360011168 0.99999769246978

1.00629382044492 1.00122625262452 1.00020913422909 1.00006457528071

0.94053380032538 0.99002109233257 0.99843882973683 0.99959894674583

1.21192240065814 1.02675047959294 1.00390068391728 1.00076599977069

0.70127773711048 0.98632445271706 0.99814421193350 0.99995872397917

1.09581827413185 0.97665640574577 0.99648826215166 0.99924398486644

1.02336722877166 0.97202638286108 0.99817900385086 0.99954855825103

1.17106319302461 1.08699197553880 1.00865490837008 1.00093583217847

0.84388833351742 1.01876324868609 1.00180799203745 1.00046269820878

1.05873976538266 0.91901141042875 0.99201372691240 0.99947025166967

0.83656328596098 0.99517814267567 0.99982038442358 0.99994904674245

1.11081398434722 1.02711587407747 1.00235015067332 1.00000204938269

Columns 5 through 8:

0.99999961854621 0.99999994094735 0.99999998766239 0.99999999855073

1.00001011207963 1.00000144010224 1.00000024261769 1.00000002686453

0.99994111240620 0.99999219043793 0.99999896219131 0.99999989383996

1.00010224291608 1.00001248184177 1.00000118124187 1.00000010269591

1.00000127796084 1.00000145928561 1.00000046770292 1.00000005717119

0.99992911053781 0.99999047692878 0.99999938387517 0.99999996654903

0.99991178843626 0.99999025422639 0.99999904235602 0.99999990205978

1.00006888183899 1.00000750419835 1.00000043261636 1.00000003347395

1.00007635754385 1.00000795151244 1.00000037573688 1.00000001654656

0.99998930540261 0.99999936242353 0.99999974850404 0.99999996053368

1.00000531729495 1.00000115395127 1.00000022853536 1.00000002945979

0.99996455824570 0.99999575955749 0.99999995386741 1.00000001329310

Columns 9 through 12:

0.99999999968290 0.99999999995858 0.99999999999142 0.99999999999846

1.00000000486121 1.00000000054150 1.00000000009099 1.00000000001083

0.99999998419515 0.99999999849845 0.99999999980536 0.99999999998316

1.00000001082632 1.00000000080921 1.00000000005150 1.00000000000707

1.00000000910318 1.00000000083262 1.00000000008191 1.00000000000643

0.99999999908118 0.99999999991561 1.00000000001540 0.99999999999534

0.99999999156213 0.99999999916125 0.99999999995489 0.99999999998787

1.00000000114667 1.00000000034830 1.00000000006071 1.00000000000151

0.99999999898599 1.00000000021764 1.00000000002200 1.00000000000121

0.99999999296541 0.99999999952729 0.99999999994473 0.99999999999787

1.00000000254603 1.00000000017743 0.99999999999241 1.00000000000648

1.00000000512875 0.99999999999575 0.99999999998501 1.00000000000579

Columns 13 through 16:

0.99999999999967 0.99999999999997 1.00000000000009 1.00000000000005

1.00000000000174 1.00000000000020 1.00000000000017 1.00000000000001

0.99999999999907 0.99999999999885 0.99999999999972 0.99999999999995

1.00000000000006 0.99999999999996 0.99999999999980 0.99999999999995

0.99999999999914 0.99999999999990 0.99999999999974 0.99999999999995

0.99999999999850 0.99999999999984 0.99999999999970 0.99999999999995

0.99999999999859 0.99999999999981 0.99999999999978 0.99999999999997

1.00000000000019 1.00000000000000 1.00000000000003 1.00000000000000

1.00000000000060 1.00000000000006 1.00000000000016 1.00000000000002

1.00000000000040 1.00000000000007 1.00000000000024 1.00000000000003

1.00000000000132 1.00000000000020 1.00000000000043 1.00000000000005

1.00000000000136 1.00000000000021 1.00000000000049 1.00000000000006

Columns 17 through 20:

1.00000000000001	1.00000000000001	0.99999999999999	1.00000000000014
1.00000000000001	1.00000000000001	0.99999999999999	1.00000000000014
1.00000000000000	1.00000000000000	0.99999999999999	1.00000000000014
1.00000000000000	1.00000000000000	0.99999999999998	1.00000000000014
0.99999999999999	1.00000000000000	0.99999999999998	1.00000000000014
0.99999999999999	0.99999999999999	0.99999999999998	1.00000000000014
0.99999999999999	0.99999999999999	0.99999999999998	1.00000000000014
1.00000000000000	0.99999999999999	0.99999999999998	1.00000000000014
1.00000000000000	0.99999999999999	0.99999999999998	1.00000000000014
1.00000000000000	0.99999999999998	0.99999999999998	1.00000000000014
1.00000000000000	0.99999999999998	0.99999999999998	1.00000000000014
1.00000000000000	0.99999999999998	0.99999999999997	1.00000000000014

$Reg2$=

Columns 1 through 4:

0.99999999986089 0.99999999999180 0.99999999999582 0.99999999999896

1.00000000204049 1.00000000015693 1.00000000003530 1.00000000000411

0.99999999402540 0.99999999942425 0.99999999994176 0.99999999999851

1.00000000242300 1.00000000031104 1.00000000000847 1.00000000000393

1.00000000452605 1.00000000058822 1.00000000002027 1.00000000000206

1.00000000083251 1.00000000011132 0.99999999998567 0.99999999999339

0.99999999843441 0.99999999959019 0.99999999997943 0.99999999998801

1.00000000008947 0.99999999982236 1.00000000003921 0.99999999999324

0.99999999773037 0.99999999972420 1.00000000003265 0.99999999999675

0.99999999523447 0.99999999958226 0.99999999998959 0.99999999999996

0.99999999951229 1.00000000011753 0.99999999999194 1.00000000000844

1.00000000535442 1.00000000060201 0.99999999997985 1.00000000001594

Columns 5 through 8:

0.99999999999964 0.99999999999997 1.00000000000002 1.00000000000001

1.00000000000128 1.00000000000017 1.00000000000000 0.99999999999999

1.00000000000057 1.00000000000007 0.99999999999997 0.99999999999998

1.00000000000036 0.99999999999999 0.99999999999995 0.99999999999998

0.99999999999923 0.99999999999982 0.99999999999993 0.99999999999998

0.99999999999783 0.99999999999966 0.99999999999992 0.99999999999998

0.99999999999729 0.99999999999960 0.99999999999993 0.99999999999999

0.99999999999811 0.99999999999972 0.99999999999996 0.99999999999999

0.99999999999918 0.99999999999989 1.00000000000000 1.00000000000000

1.00000000000048 1.00000000000012 1.00000000000005 1.00000000000001

1.00000000000255 1.00000000000044 1.00000000000012 1.00000000000002

1.00000000000459 1.00000000000075 1.00000000000019 1.00000000000003

Columns 9 through 12:

0.99999999999999 0.99999999999999 0.99999999999985 0.99999999999934

0.99999999999999 0.99999999999999 0.99999999999989 0.99999999999950

1.00000000000000 0.99999999999999 0.99999999999993 0.99999999999966

1.00000000000000 0.99999999999999 0.99999999999997 0.99999999999983

1.00000000000000 0.99999999999999 1.00000000000001 0.99999999999999

1.00000000000000 0.99999999999999 1.00000000000005 1.00000000000016

1.00000000000000 1.00000000000000 1.00000000000010 1.00000000000033

1.00000000000000 1.00000000000000 1.00000000000014 1.00000000000050

1.00000000000000 1.00000000000001 1.00000000000019 1.00000000000067

0.99999999999999 1.00000000000002 1.00000000000023 1.00000000000085

0.99999999999999 1.00000000000003 1.00000000000028 1.00000000001032

0.99999999999998 1.00000000000004 1.00000000000032 1.00000000000120

Columns 13 through 16:

1.00000000002429 1.00000000030304 0.99999999893369 0.99999999361121

1.00000000001694 1.00000000022115 0.99999999928457 0.99999999522508

1.00000000000960 1.00000000013926 0.99999999963545 0.99999999683896

1.00000000000225 1.00000000005737 0.99999999998633 0.99999999845284

0.99999999999491 0.99999999997547 1.00000000033720 1.00000000006672

0.99999999998757 0.99999999989357 1.00000000068808 1.00000000168060

0.99999999998023 0.99999999981167 1.00000000103894 1.00000000329448

0.99999999997290 0.99999999972976 1.00000000138980 1.00000000490836

0.99999999996556 0.99999999964785 1.00000000174066 1.00000000652224

0.99999999995823 0.99999999956594 1.00000000209152 1.00000000813612

0.99999999995090 0.99999999948402 1.00000000244237 1.00000000975000

0.99999999994356 0.99999999940210 1.00000000279323 1.00000001136388

APPENDIX C

OCTAVE CODE FOR REGULARIZING FREDHOLM INTEGRAL EQUATION OF THE FIRST KIND AND ITS RESULT

Function = Regularize($\mathbf{A}, \mathbf{b}, n$).

$\mathbf{A}$ =Generated matrix from an integral equation of size seventeen.

[m, n] = size($\mathbf{A}$);

$\mathbf{x}$ = ones(n, 1);

$\mathbf{b} = \mathbf{A} * \mathbf{x}$;

$\mathbf{A}$t$\mathbf{A} = \mathbf{A}' * \mathbf{A}$;

$\mathbf{A}$t$\mathbf{b} = \mathbf{A}' * \mathbf{b}$;

Create the regularization parameters $t = [10^{-16}, ..., 10^{3}]$.

t = zeros(20, 1);

t(1) = 10^{-16};

for $i = 1 : 19$

t(i + 1) = t(i) $*$ 10;

end.

if order == 0

D = eye(n);

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg0 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg0(i, :) = ($\mathbf{A}$t + t(i) $*$ D) $\setminus$ $\mathbf{A}$t$\mathbf{b}$;

res0_norm(i) = norm($\mathbf{A} * \mathbf{x}$ − reg(:, i) − $\mathbf{b}$);

sol0_norm(i) = norm(D $*$ $\mathbf{x}$ − reg(:, i));

end.

$\mathbf{x}$_reg0

plot(−log10(res0_norm), −log10(sol0_norm));

end.

if order == 1

h = 1;

D1 = DiffOpr(1, n);

L1 = D1′ ∗ D1;

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg1 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg1(i, :) = ($\mathbf{A}$t + t(i) ∗ L1) \ $\mathbf{A}$t$\mathbf{b}$;

res1_norm(i) = norm($\mathbf{A}$ ∗ $\mathbf{x}$_reg(:, i) − $\mathbf{b}$);

sol1_norm(i) = norm(D1 ∗ $\mathbf{x}$_reg(:, i));

end.

$\mathbf{x}$_reg1

plot(−log10(res1_norm), −log10(sol1_norm))

end.

if order == 2

h_2 = 0.01;

D2 = DiffOpr(2, n);

B2 = (1/h^2) ∗ D2;

L2 = B2′ ∗ B2;

Compute the regularized solutions corresponding to each regularization parameter. $\mathbf{x}$_reg2 = zeros(n, length(t));

for i = 1 : length(t)

$\mathbf{x}$_reg2(i, :) = ($\mathbf{A}$t + t(i) ∗ L2) \ $\mathbf{A}$t$\mathbf{b}$;

res2_norm(i) = norm($\mathbf{A}$ ∗ $\mathbf{x}$_reg(:, i) − $\mathbf{b}$);

sol2_norm(i) = norm(D2 ∗ $\mathbf{x}$_reg(:, i));

$\mathbf{x}$_reg2

plot(−log10(res2_norm), −log10(sol2_norm))

end.

$\mathbf{x}$_Reg1 =

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
-8.9437e+0	-2.8742e+0	5.2110e-1	-4.4836e-1	4.3307e-1	1.5411e+0
3.4315e+1	1.6576e+1	3.1678e+0	1.8536e+0	1.6785e+0	1.7170e+0
-2.5763e+1	-1.1872e+1	-4.3959e-1	1.6881e+0	1.3564e+0	1.1540e+0
-8.1029e-3	-5.4719e-3	-2.8565e-3	-1.7156e-3	-1.1120e-3	-1.6808e-4
1.5684e+0	3.1900e-1	-1.0315e-1	8.5582e-1	1.3725e+0	8.2317e-1
-2.0151e+1	-6.2096e+0	3.5506e+0	2.9824e+0	1.8544e+0	8.8795e-1
-3.0313e-3	-1.6476e-3	-6.6537e-4	-7.3133e-4	-8.9552e-4	-1.0003e-3
4.9852e+0	1.5110e+0	-6.5348e-1	4.9395e-2	6.7894e-1	8.6305e-1
4.4937e+1	2.3933e+1	5.7945e+0	1.0486e+0	4.7272e-1	1.1316e+0
-2.4875e+1	-1.4818e+1	-4.9543e+0	-9.2321e-1	2.1452e-1	1.3426e+0
1.2109e+0	2.0287e+0	2.0587e+0	1.2644e+0	1.4579e+0	1.8075e+0
8.8293e+0	5.7723e+0	3.4963e+0	3.2145e+0	2.6884e+0	2.0446e+0
-7.1561e+0	-1.7755e+0	2.3208e+0	2.7563e+0	2.4968e+0	1.6931e+0
-2.6143e-1	-7.1066e-2	6.9475e-2	7.0345e-2	8.6132e-2	2.1866e-1
-4.9130e-2	-2.5449e-2	-6.3015e-3	-2.9768e-3	-3.0347e-3	-3.5252e-3
-1.5204e-2	-6.8933e-3	-7.2102e-4	-5.0457e-4	-6.3954e-4	-1.9306e-4
5.9559e+0	4.2383e+0	2.9013e+0	2.7469e+0	2.8764e+0	3.1254e+0

Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
1.7848e+0	1.7698e+0	2.0501e+0	2.4069e+0	1.9569e+0	1.3621e+0
1.6946e+0	1.6723e+0	1.8768e+0	2.1336e+0	1.7876e+0	1.3214e+0
1.1362e+0	1.1406e+0	1.2279e+0	1.3479e+0	1.2998e+0	1.1897e+0
7.4886e-5	1.2341e-3	1.2826e-2	9.3873e-2	5.1579e-1	9.5870e-1
6.6482e-1	7.3600e-1	5.8606e-1	3.7633e-1	5.8895e-1	8.3219e-1
6.7660e-1	7.3082e-1	5.6407e-1	3.1818e-1	4.2156e-1	5.9281e-1
-8.7150e-4	-1.7358e-3	-8.6222e-3	-9.840e-3	4.2898e-2	2.3592e-1
9.5228e-1	9.6995e-1	7.8760e-1	5.3930e-1	6.0629e-1	6.9293e-1
1.4348e+0	1.4429e+0	1.2363e+0	9.5903e-1	1.0164e+0	1.0298e+0
1.6501e+0	1.5907e+0	1.4629e+0	1.2997e+0	1.2970e+0	1.2492e+0
1.7285e+0	1.5734e+0	1.5738e+0	1.5791e+0	1.4633e+0	1.3568e+0
1.6663e+0	1.4954e+0	1.6226e+0	1.7676e+0	1.5186e+0	1.3619e+0
1.3811e+0	1.3594e+0	1.5684e+0	1.7711e+0	1.4511e+0	1.2782e+0
5.9848e-1	9.6500e-1	1.1961e+0	1.3929e+0	1.2235e+0	1.1239e+0
-2.6799e-3	2.5916e-3	4.3946e-2	2.9142e-1	7.6219e-1	9.2083e-1
9.9413e-4	3.3414e-3	1.9827e-2	1.7465e-1	5.9036e-1	7.7890e-1
3.0781e+0	2.8980e+0	2.4931e+0	1.6765e+0	9.6202e-1	7.5094e-1

Column 13	Column 14	Column 15	Column 16	Column 17	Column 18
0.99952	0.96646	0.99395	0.99807	0.99864	0.99882
0.99780	0.96816	0.99420	0.99810	0.99864	0.99882
0.98473	0.97125	0.99471	0.99816	0.99865	0.99882
0.95240	0.97367	0.99523	0.99823	0.99867	0.99883
0.91635	0.97657	0.99586	0.99832	0.99869	0.99883
0.84607	0.97473	0.99600	0.99836	0.99871	0.99883
0.73491	0.96626	9.9542	0.99834	0.99872	0.99883
0.92990	1.00290	0.99965	0.99881	0.99879	0.99884
1.0742	1.02980	1.0028	0.99917	0.99884	0.99885
1.1654	1.04600	1.0046	0.99940	0.99889	0.99886
1.2035	1.05110	1.0052	0.99950	0.99892	0.99886
1.1914	1.04530	1.0045	0.99947	0.99894	0.99887
1.1362	1.03010	1.0027	0.99933	0.99894	0.99887
1.0491	1.00800	1.0000	0.99910	0.99894	0.99887
0.94725	0.98278	0.99700	0.99882	0.99893	0.99887
0.86209	0.96075	0.99433	0.99857	0.99891	0.99887
0.81990	0.94973	0.99300	0.99844	0.99890	0.99887

$\mathbf{x}$_Reg2 =

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6
3.2434353	3.8698193	3.5831519	2.6307741	1.7127983	0.9228078
2.1195973	2.4507759	2.3495901	1.9444993	1.4169926	0.9143018
0.9746398	1.1016834	1.1790306	1.2779945	1.1251945	0.9060205
0.0129868	0.0464812	0.2488778	0.6925389	0.8527269	0.8995588
-0.1272902	-0.3343805	-0.1619312	0.2926023	0.6294548	0.8987991
-0.1219766	-0.3072576	-0.2154045	0.0736196	0.4839333	0.9086499
-0.0131434	-0.0227473	-0.0041868	0.0758580	0.4651249	0.9380455
0.7928620	0.4515691	0.4326341	0.3846340	0.6448897	1.0007187
1.5567492	1.0133085	0.9472347	0.8161117	0.9073638	1.0709604
1.9766777	1.5633650	1.4181065	1.2302218	1.1629245	1.1290077
2.0270492	1.9782635	1.7433220	1.5293400	1.3487780	1.1613351
1.7929152	2.1085105	1.8430321	1.6567490	1.4285634	1.1606735
1.3439432	1.8190411	1.6743448	1.5949241	1.3907668	1.1256928
0.7051324	1.0912175	1.2623432	1.3636640	1.2456226	1.0602511
0.0201782	0.2235134	0.7558206	1.0188199	1.0202938	0.9721249
0.0226790	0.1787181	0.5182579	0.6525046	0.7521472	0.8711335
2.7138733	1.6365075	0.6651896	0.3094552	0.4704665	0.7656677

Column 7	Column 8	Column 9	Column 10	Column 11	Column 12
0.8995849	0.9825752	0.9967444	0.9982540	0.9984059	0.9984211
0.9192529	0.9859340	0.9971402	0.9983325	0.9984524	0.9984644
0.9387790	0.9892679	0.9975334	0.9984107	0.9984989	0.9985078
0.9579232	0.9925257	0.9979184	0.9984881	0.9985454	0.9985511
0.9765372	0.9956553	0.9982893	0.9985641	0.9985917	0.9985944
0.9944527	0.9985929	0.9986392	0.9986379	0.9986377	0.9986377
1.0119667	1.0013228	0.9989656	0.9987094	0.9986836	0.9986810
1.0300149	1.0039006	0.9992735	0.9987790	0.9987292	0.9987242
1.0448963	1.0058981	0.9995193	0.9988423	0.9987742	0.9987674
1.0538310	1.0069971	0.9996705	0.9988962	0.9988183	0.9988105
1.0550497	1.0070018	0.9997072	0.9989385	0.9988612	0.9988535
1.0478315	1.0058451	0.9996227	0.9989687	0.9989029	0.9988964
1.0324728	1.0035859	0.9994236	0.9989873	0.9989435	0.9989391
1.0101701	1.0003948	0.9991275	0.9989962	0.9989831	0.9989817
0.9827919	0.9965257	0.9987609	0.9989981	0.9990219	0.9990243
0.9525174	0.9922694	0.9983542	0.9989959	0.9990604	0.9990668
0.9212665	0.9878839	0.9979340	0.9989923	0.9990987	0.9991094

Column 13	Column 14	Column 15	Column 16	Column 17	Column 18
0.9984226	0.9984228	0.9984228	0.9984228	0.9984219	0.9984245
0.9984656	0.9984658	0.9984658	0.9984658	0.9984650	0.9984673
0.9985087	0.9985087	0.9985088	0.9985087	0.9985081	0.9985101
0.9985517	0.9985517	0.9985517	0.9985517	0.9985511	0.9985529
0.9985947	0.9985947	0.9985947	0.9985947	0.9985942	0.9985957
0.9986377	0.9986377	0.9986377	0.9986377	0.9986373	0.9986386
0.9986807	0.9986807	0.9986807	0.9986807	0.9986803	0.9986814
0.9987237	0.9987237	0.9987237	0.9987237	0.9987234	0.9987242
0.9987667	0.9987667	0.9987667	0.9987667	0.9987665	0.9987670
0.9988097	0.9988097	0.9988097	0.9988096	0.9988096	0.9988098
0.9988527	0.9988526	0.9988526	0.9988526	0.9988526	0.9988527
0.9988957	0.9988956	0.9988956	0.9988956	0.9988957	0.9988955
0.9989387	0.9989386	0.9989386	0.9989386	0.9989388	0.9989383
0.9989816	0.9989816	0.9989816	0.9989816	0.9989818	0.9989811
0.9990245	0.9990246	0.9990246	0.9990246	0.9990249	0.9990239
0.9990675	0.9990676	0.9990676	0.9990676	0.9990680	0.9990668
0.9991104	0.9991105	0.9991105	0.9991105	0.9991111	0.9991096