Regularization of Ill-conditioned Linear Systems

Sheima Mohammed Eldirdiri Abueldahab (sheima@aims.ac.za)
شيمَاء مُحَمد الدِردِيري أَبو الدَهَب
African Institute for Mathematical Sciences (AIMS)

Supervised by Francis Benyah
University of the Western Cape

June 8, 2007

Abstract

Ill-conditioned linear systems arise in many applications, for example, in the solution of integral equations, and in the solution of non-linear programming problems. In many application of linear algebra, the need arises to find a good approximation $\widehat{\mathbf{x}}$ to a vector $\mathbf{x} \in \mathbb{R}^n$ satisfying an approximating equation $\boldsymbol{A}\mathbf{x} \approx \mathbf{b}$ with ill-conditioned $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, given $\mathbf{b} \in \mathbb{R}^m$. Straightforward the computed solution $\widehat{\mathbf{x}}$ is usually meaningless approximation to $\mathbf{x}$ due to the error in the right-hand side $\mathbf{b}$ and the severe ill-conditioning of the matrix $\boldsymbol{A}$. In order to avoid this difficulty, one typically replaces the linear systems $\boldsymbol{A}\mathbf{x} = \mathbf{b}, \boldsymbol{A} \in \mathbb{R}^{m\times n}, \mathbf{x} \in \mathbb{R}^n, \mathbf{b} \in \mathbb{R}^m$, by a nearby system that is less sensitive to the error in $\mathbf{b}$ and considers the computed solution of the latter system an approximation of $\mathbf{x}$. This replacement is known as regularization. This essay examines various regularization methods for computing stable solution to ill-conditioned linear systems.

Contents

1. Introduction

We will begin our study of ill-conditioned linear systems with some basic concepts of linear algebra.

1.1 Basic concepts of Linear Systems

A linear system is an equation of the form

$$\boldsymbol{A}\mathbf{x} = \mathbf{b}, \tag{1.1}$$

where $\boldsymbol{A}$ is a known $m \times n$ matrix called the *coefficient matrix*, given by

$$\boldsymbol{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & & & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}, \tag{1.2}$$

$\mathbf{b}$ is a known m-vector of constants, given by

$$\mathbf{b} = \begin{bmatrix} b_1 & b_2 & b_3 & \cdots & b_m \end{bmatrix}^T, \tag{1.3}$$

and $\mathbf{x}$ is an unknown n-vector, to be determined, given by

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & x_3 & \cdots & x_n \end{bmatrix}^T. \tag{1.4}$$

Linear systems of equations arise from many situations including the following:

1. Solving systems of non-linear equations by Newton's method.
2. Curve fitting or polynomial interpolation.
3. Numerical solutions of differential equations by the finite difference method.
4. Polynomial approximations to continuous functions on a finite interval $[0, 1]$ results in solving a linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$, where $\boldsymbol{A}$ is a Hilbert matrix.

For any linear systems the **augmented matrix** given by

$$[\boldsymbol{A} \mid \mathbf{b}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} & b_3 \\ \vdots & & & & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} & b_m \end{bmatrix} \tag{1.5}$$

Theorem 1.1. [1] *Consider the system* $\boldsymbol{A}\mathbf{x} = \mathbf{b}$*, with coefficient matrix* $\boldsymbol{A}$ *and augmented matrix* $[\boldsymbol{A} \mid \mathbf{b}]$*, of size* $m \times (n+1)$*. Then the linear system,*

1. $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ *is inconsistent (i.e., no solution exists) if and only if* $\operatorname{rank}(\boldsymbol{A}) < \operatorname{rank}[\boldsymbol{A} \mid \mathbf{b}]$.
2. $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ *has a unique solution if and only if* $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}[\boldsymbol{A} \mid \mathbf{b}] = n$.
3. $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ *has infinitely many solutions if and only if* $\operatorname{rank}(\boldsymbol{A}) = \operatorname{rank}[\boldsymbol{A} \mid \mathbf{b}] < n$.

1.2 Vector and Matrix Norms

Definition 1.2. (Vector Norm) A vector norm is a function $\mathbf{V} \rightarrow \mathbb{R}$ satisfies the following three conditions:

1. $\| \mathbf{x} \| > 0$, for any nonzero vector $\mathbf{x}$.
2. $\| \alpha\mathbf{x} \| = | \alpha | \| \mathbf{x} \|$, for any real scalar α.
3. $\| \mathbf{x} + \mathbf{y} \| \leq \| \mathbf{x} \| + \| \mathbf{y} \|$, for any vectors $\mathbf{x}$ and $\mathbf{y}$.

The three most important vector norms are :

$$\| \mathbf{x} \|_1 = \sum_{i=1}^{n} | \mathbf{x}_i | = | \mathbf{x}_1 | + \cdots + | \mathbf{x}_n |, \tag{1.6}$$

$$\| \mathbf{x} \|_2 = \left[\sum_{i=1}^{n} | \mathbf{x}_i |^2 \right]^{\frac{1}{2}} = (| \mathbf{x}_1 |^2 + \cdots + | \mathbf{x}_n |^2)^{\frac{1}{2}} = (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}}, \tag{1.7}$$

and

$$\| \mathbf{x} \|_\infty = \max_{1 \leq i \leq n} \{ | \mathbf{x}_i | \}. \tag{1.8}$$

Where $\mathbf{x}$ is n-dimensional vector

$$\mathbf{x} = \begin{bmatrix} x_1 & x_2 & \cdots & x_n \end{bmatrix}^T. \tag{1.9}$$

Example 1.3. Find $\| \mathbf{x} \|_\infty$ of the following vector $\mathbf{x}$.

$$\mathbf{x} = \begin{bmatrix} 10 \\ -3 \\ 5 \end{bmatrix}. \tag{1.10}$$

[1] Theorem 1.1 can be found in Ref. [7]

Solution 1.3.

$$\begin{aligned} \| \mathbf{x} \|_\infty &= \max_{1\le i\le 3} | \mathbf{x}_i | \\ &= \max [| 10 |, | -3 |, | 5 |] \\ &= \max [10, 3, 5] \\ &= 10. \end{aligned}$$

Definition 1.4. (Matrix Norm) For any real matrix $\boldsymbol{A}$, a matrix norm $\| \boldsymbol{A} \|$ is a nonnegative number associated with $\boldsymbol{A}$ having the properties

1. $\| \boldsymbol{A} \| \geq 0$, $\| \boldsymbol{A} \| = 0$ iff $\boldsymbol{A} = 0$.
2. $\| \alpha \boldsymbol{A} \| = | \alpha | \| \boldsymbol{A} \|$, for any scalar α.
3. $\| \boldsymbol{A} + \mathbf{B} \| \leq \| \boldsymbol{A} \| + \| \mathbf{B} \|$.
4. $\| \boldsymbol{A}\mathbf{B} \| \leq \| \boldsymbol{A} \| \| \mathbf{B} \|$.

The three most important matrix norms are :

$$\| \boldsymbol{A} \|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} | \mathbf{a}_{ij} | . \tag{1.11}$$

$$\| \boldsymbol{A} \|_2 = \left[\sigma_{\max} \left(\boldsymbol{A}^T \boldsymbol{A}\right)\right]^{\frac{1}{2}}, \tag{1.12}$$

$\sigma_{\max}$ is the largest eigenvalue of the positive-semidefinite matrix $A^T A$.

$$\| \boldsymbol{A} \|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} | \mathbf{a}_{ij} | . \tag{1.13}$$

Example 1.5. [2]Find $\| \boldsymbol{A} \|_\infty$ of the following matrix $\boldsymbol{A}$.

$$\boldsymbol{A} = \begin{bmatrix} 10 & -7 & 0 \\ -3 & 2.099 & 6 \\ 5 & -1 & 5 \end{bmatrix}. \tag{1.14}$$

Solution 1.5. The $\| \boldsymbol{A} \|_\infty$ given by

$$\begin{aligned} \| \boldsymbol{A} \|_\infty &= \max_{1\le i\le 3} \sum_{j=1}^{3} | \mathbf{a}_{ij} | \\ &= \max [(| 10 | + | -7 | + | 0 |), (| -3 | + | 2.099 | + | 6 |), (| 5 | + | -1 | + | 5 |)] \\ &= \max [(10 + 7 + 0), (3 + 2.099 + 6), (5 + 1 + 5)] \\ &= \max [17, 11.099, 11] \\ &= 17. \end{aligned}$$

[2]Example 1.5 can be found in Ref. [5]

2. Ill-Conditioned Linear Systems

This chapter devote to the study of ill-conditioned linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. We will show how small change in the augmented matrix $[\boldsymbol{A} \mid \mathbf{b}]$ affect on the computed solution $\widehat{\mathbf{x}}$ of linear system to be ill-conditioned or well-conditioned. Some classical applications to ill-conditioned linear systems will also discuss (e.g., the Hilbert and the Vandermonde matrices). Since many discussion about ill-conditioned linear systems required knowledge about singular value decomposition (SVD), we will discuss this method and we will show how it can be use to solve the linear systems.

2.1 Definition of Ill-Conditioned Linear Systems

The linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ is said to be **ill-conditioned** if relatively small changes in the entries of augmented matrix $[\boldsymbol{A} \mid \mathbf{b}]$ can cause relatively large changes in the solution. If the system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ is ill-conditioned, then the computed solution to $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ will generally not be very accurate. On the other hand, the system is said to be **well-conditioned** if relatively small changes in the entries of $[\boldsymbol{A} \mid \mathbf{b}]$ results in relatively small changes in the solutions to $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. In that case one should be able to compute solutions more accurately.

Example 2.1. [3] Consider the system

$$\begin{bmatrix} 1 & 3 \\ 2 & 5.999 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 7.999 \end{bmatrix}. \tag{2.1}$$

.

Denoting the system of equations as

$$\boldsymbol{A}\mathbf{x} = \mathbf{b},$$

and making a small change in the right hand side vector of the equation 2.1, gives

$$\begin{bmatrix} 1 & 3 \\ 2 & 5.999 \end{bmatrix} \begin{bmatrix} x_1^{'} \\ x_2^{'} \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix} \tag{2.2}$$

. The exact solution of equation 2.2, is

$$\mathbf{x}^{'} = \begin{bmatrix} x_1^{'} \\ x_2^{'} \end{bmatrix} = \begin{bmatrix} 4 \\ 0 \end{bmatrix}. \tag{2.3}$$

The system 2.1 has exact solution

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \tag{2.4}$$

[3] Example 2.1 can be found in Ref. [5]

The relative error in $\mathbf{b}$ is

$$\frac{\| \mathbf{b} - \mathbf{b}^{'} \|_\infty}{\| \mathbf{b} \|_\infty} = 1.250 \times 10^{-4}.$$

However, the corresponding relative error in the solution vector is

$$\frac{\| \mathbf{x} - \mathbf{x}^{'} \|_\infty}{\| \mathbf{x} \|_\infty} = 3.0.$$

We see the small relative error in the right hand side vector of 1.250×10^{-4} results in a large relative error in the solution vector as 3.0.

This system is **ill-conditioned** since a small change in the right hand side resulted in a large change in the solution vector.

Example 2.2. [4] Consider the system

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix}. \tag{2.5}$$

.

Denoting the system of equations as

$$\boldsymbol{A}\mathbf{x} = \mathbf{b},$$

and making a small change in the right hand side vector of the equation 2.5, gives

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1^{'} \\ x_2^{'} \end{bmatrix} = \begin{bmatrix} 4 \\ 7 \end{bmatrix}. \tag{2.6}$$

The exact solution of equation 2.6, is

$$\mathbf{x} = \begin{bmatrix} x_1^{'} \\ x_2^{'} \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \tag{2.7}$$

The system 2.5 has exact solution

$$\mathbf{x}^{'} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1.999 \\ 1.001 \end{bmatrix}. \tag{2.8}$$

The relative error in $\mathbf{b}$ is

$$\frac{\| \mathbf{b} - \mathbf{b}^{'} \|_\infty}{\| \mathbf{b} \|_\infty} = \frac{0.001}{7.001} = 1.4284 \times 10^{-4}. \tag{2.9}$$

However, the corresponding relative error in the solution vector is

$$\frac{\| \mathbf{x} - \mathbf{x}^{'} \|_\infty}{\| \mathbf{x} \|_\infty} = \frac{0.001}{1.9990} = 5.0025 \times 10^{-4}.$$

[4] Example 2.2 can be found in Ref. [5]

We see the small relative error in the right hand side vector of 1.4284×10^{-4} results in a small relative error in the solution vector as 5.0025×10^{-4}.

This system is **well-conditioned** as a small changes in the right hand side resulted in small change in the solution vector.

2.2 The Condition Number of A Matrix

The condition number is a measure of how close a matrix $\boldsymbol{A}$ is to being singular.

Definition 2.3. (Condition Number) Let $\boldsymbol{A}$ be an $n \times n$ nonsingular matrix. The condition number of $\boldsymbol{A}$, defined by $\mathrm{cond}(\boldsymbol{A})$, denoted by cond$(\boldsymbol{A})$, is defined as

$$\mathrm{cond}(\boldsymbol{A}) = \| \boldsymbol{A} \| \| \boldsymbol{A}^{-1} \| . \tag{2.10}$$

Where $\| . \|$ is a given matrix norm.

Example 2.4. [5] Let

$$A = \begin{bmatrix} 3 & 3 \\ 4 & 5 \end{bmatrix},$$

$$\boldsymbol{A}^{-1} = \frac{1}{3} \begin{bmatrix} 5 & -3 \\ -4 & 3 \end{bmatrix}.$$

And

$$\| \boldsymbol{A} \|_\infty = 9, \ \| \boldsymbol{A}^{-1} \|_\infty = \frac{8}{3},$$

Then

$$\mathrm{cond}(\boldsymbol{A}) = 9 \times \frac{8}{3} = 24.$$

If $\mathrm{cond}(\boldsymbol{A})$ is small (close to 1) then the matrix is said to be **well-conditioned**. On the other hand, if $\mathrm{cond}(\boldsymbol{A})$ is large, that is, if it is significantly larger than one, then the matrix is said to be **ill-conditioned**. The condition number of a matrix $\boldsymbol{A}$ associated with the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ gives a bound on how inaccurate the solution $\mathbf{x}$ will be after approximate solution. In particular, one should think of the condition number as being the rate at which the solution, $\mathbf{x}$, will change with respect to a change in $\mathbf{b}$. Thus, if the condition number is large, even a small error in $\mathbf{b}$ may cause a large error in $\mathbf{x}$. On the other hand, if the condition number is small then the error in $\mathbf{x}$ will not be much bigger than the error in $\mathbf{b}$.

[5] Example 2.4 can be found in Ref. [2]

2.3 Testing The Accuracy of the Solution of a Linear Systems

Once a solution $\widehat{\mathbf{x}}$ of the system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ has been computed , it is natural to test how accurate the computed solution $\widehat{\mathbf{x}}$ is. If the exact solution $\mathbf{x}$ is known, then one could of course compute the relative error $\| \mathbf{x} - \widehat{\mathbf{x}} \| / \| \mathbf{x} \|$, in $\mathbf{x}$. However, in most practical situations, the exact solution is not known. In such cases, the most obvious thing to do is to compute the residual $\mathbf{r} = \mathbf{b} - \boldsymbol{A}\widehat{\mathbf{x}}$ and see how small the relative residual $\| \mathbf{r} \| / \| \mathbf{b} \|$ is. Unfortunately, a small relative residual does not guarantee the accuracy of the solution. The following example illustrates this fact.

Example 2.5. [6] Consider the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ given by

$$\begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix}.$$

Let

$$\widehat{\mathbf{x}} = (3, 0)^T. \tag{2.11}$$

Then

$$\mathbf{r} = \mathbf{b} - \boldsymbol{A}\widehat{\mathbf{x}} = \begin{bmatrix} 3 \\ 3.0001 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1.0001 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0.0002 \end{bmatrix},$$

and relative residual

$$\frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} = 0.000066664, \tag{2.12}$$

which is small, even though the solution $\widehat{\mathbf{x}} = (3, 0)^T$ is nowhere near the exact solution

$$\mathbf{x} = (1, 1)^T. \tag{2.13}$$

The above phenomena can be explained by the following Theorem.

Theorem 2.6. [7] **(The Residual theorem)** *let $\widehat{\mathbf{x}}$ be the computed solution to the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. Then*

$$\frac{\| \mathbf{x} - \widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \leq \| \boldsymbol{A} \| \| \boldsymbol{A}^{-1} \| \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|}, \tag{2.14}$$

or

$$\frac{\| \mathbf{x} - \widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \leq \ \mathrm{cond}(A) \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} \ . \tag{2.15}$$

[6] Example 2.5 can be found in Ref. [2]

[7] Theorem 2.6 can be found in Ref. [2]

Proof 2.6.

$$\begin{aligned} \mathbf{r} &= \mathbf{b} - \boldsymbol{A}\widehat{\mathbf{x}} \\ &= \boldsymbol{A}\mathbf{x} - \boldsymbol{A}\widehat{\mathbf{x}} \\ &= \boldsymbol{A}\left(\mathbf{x} - \widehat{\mathbf{x}}\right), \end{aligned} \tag{2.16}$$

we have

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

taking norms give,

$$\| \mathbf{x} - \widehat{\mathbf{x}} \| \leq \| \boldsymbol{A}^{-1} \| \| \mathbf{r} \| . \tag{2.17}$$

Also from

$$\mathbf{b} = \boldsymbol{A}\mathbf{x}, \tag{2.18}$$

we have

$$\| \mathbf{b} \| \leq \| \boldsymbol{A} \| \| \mathbf{x} \|,$$

that is,

$$\frac{1}{\| \mathbf{x} \|} \leq \frac{\| \boldsymbol{A} \|}{\| \mathbf{b} \|} . \tag{2.19}$$

This combines with 2.17 to give

$$\frac{\| \mathbf{x} - \widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \leq \| \boldsymbol{A} \| \| \boldsymbol{A}^{-1} \| \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|}, \tag{2.20}$$

or

$$\frac{\| \mathbf{x} - \widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \leq \mathrm{cond}(A) \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|}. \tag{2.21}$$

The above Theorem 2.6 tells that the relative error in the computed solution $\widehat{\mathbf{x}}$ depends not only on the relative residual, but also on the quantity $\mathrm{cond}(A)$. A computed solution can be guaranteed to be accurate only when the product $\mathrm{cond}(A)\frac{\| \mathbf{r} \|}{\| \mathbf{b} \|}$ is small.

In previous Example 2.5, we get

$$\mathrm{cond}(\boldsymbol{A}) = 5.0001 \times 10^4 \quad (large), \tag{2.22}$$

and the relative residual

$$\frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} = 0.000066664.$$

So

$$\begin{aligned} \frac{\| \mathbf{x} - \widehat{\mathbf{x}} \|}{\| \mathbf{x} \|} \leq \mathrm{cond}(\boldsymbol{A}) \frac{\| \mathbf{r} \|}{\| \mathbf{b} \|} &= (5.0001 \times 10^4).(0.000066664) \\ &= 3.3333 \quad (large). \end{aligned}$$

2.4 Classical Ill-Conditioned Systems

Ill-conditioned linear systems occur in many applications of mathematics. We give some examples below.

2.4.1 Polynomial Data Fitting: The Vandermonde System

The interpolating n-th degree polynomial $p_n(x)$ at given $(n+1)$ data points (x_i, y_i), is obtained by solving the $(n+1) \times (n+1)$ linear system
$y_i = a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + ... + a_n x_i^n, \quad i = 0, 1, ..., n$ for $a_0, a_1, ..., a_n$.
In matrix form we have

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & \cdots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}. \tag{2.23}$$

The coefficient matrix of the system (2.23) is called the **Vandermonde matrix**. If the $x_i's$ are distinct, then the Vandermonde matrix is nonsingular, and so the system has a unique solution. That is, there is a unique polynomial that interpolates the given points.
For a small n the Vandermonde system can be solved quite easily. However, as n increases the system becomes increasingly ill-conditioned.

Example 2.7. [8] Let the $x_i's$ be the $(n+1)$ equally-spaced points in the interval $[0, 1]$. For example the Vandermonde matrix for $n = 4$ is

$$\mathbf{V}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & (1/4) & (1/4)^2 & (1/4)^3 & (1/4)^4 \\ 1 & (1/2) & (1/2)^2 & (1/2)^3 & (1/2)^4 \\ 1 & (3/4) & (3/4)^2 & (3/4)^3 & (3/4)^4 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \tag{2.24}$$

The condition number, $\mathrm{cond}(\mathbf{V}_4) = 686.4349$. The table below gives the condition numbers of the Vandermonde matrices for $n = 4, 5, 6, 7, 8, 9, 10, 11, 12$.

[8] Example 2.7 can be found in Ref. [2]

n	$\mathrm{cond}(\mathbf{V}_n)$
4	6.8643e+02
5	4.9244e+03
6	3.6061e+04
7	2.6782e+05
8	2.0094e+06
9	1.5193e+07
10	1.1558e+08
11	8.8348e+08
12	6.7806e+09

Table 1: Condition Numbers of Vandermonde Matrices.

2.4.2 Approximation of a Function by a Polynomial: The Hilbert System

Definition 2.8. [9]**(Hilbert Matrix)** The $n \times n$ matrix $\mathbf{H}_n$, with entries $h_{ij} = \dfrac{1}{i+j-1}$, $1 \leq i \leq n$, $1 \leq j \leq n$ 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 standard basis $\{1, x, x^2, \cdots, x^n\}$, for $\mathbb{P}^n$. Suppose a continuous function $f(x)$ defined on the interval $[0,1]$ is to be approximated by a polynomial of degree $n-1$:

$$p_{n-1}(x) = \sum_{i=1}^{n} a_i x^{i-1}, \tag{2.25}$$

such that the error

$$E = \| p_{n-1} - f \|_2^2 = \int_0^1 \left[\sum_{i=1}^{n} a_i x^{i-1} - f(x) \right]^2 dx \tag{2.26}$$

is minimized. The coefficients a_i of the polynomial are easily determined by setting

$$\frac{\partial E}{\partial a_i} = 0, \quad i = 1, 2, \cdots, n$$

$$\frac{\partial E}{\partial a_i} = 2 \int_0^1 \left[\sum_{j=1}^{n} a_j x^{j-1} - f(x) \right] x^{i-1} dx = 0, \quad i = 1, \cdots, n$$

or

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

[9] Definition 2.8 can be found in Ref. [2]

To obtain the latter form, we have interchanged the summation and integration. Letting

$$h_{ij} = \int_0^1 x^{i+j-2}\, dx \tag{2.27}$$

and

$$b_i = \int_0^1 x^{i-1} dx, \qquad i = 1, \cdots, n. \tag{2.28}$$

We have

$$\sum_{j=1}^{n} h_{ij} a_j = b_i, \qquad i = 1, \cdots, n. \tag{2.29}$$

This is exactly the linear system $\mathbf{Ha} = \mathbf{b}$, given by

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

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

For example, the Hilbert matrix of order 5, denoted by $\mathbf{H}_5$ is given by

$$\mathbf{H}_5 = \begin{bmatrix} 1 & 1/2 & 1/3 & 1/4 & 1/5 \\ 1/2 & 1/3 & 1/4 & 1/5 & 1/6 \\ 1/3 & 1/4 & 1/5 & 1/6 & 1/7 \\ 1/4 & 1/5 & 1/6 & 1/7 & 1/8 \\ 1/5 & 1/6 & 1/7 & 1/8 & 1/9 \end{bmatrix}. \tag{2.30}$$

The inverse of $\mathbf{H}_5$ is given by

$$\mathbf{H}_5^{-1} = \begin{bmatrix} 25 & -300 & 1050 & -1400 & 630 \\ -300 & 4800 & -18900 & 26880 & -12600 \\ 10507 & -18900 & 79380 & -117600 & 56700 \\ -1400 & 26880 & -117600 & 179200 & -88200 \\ 630 & -12600 & 56700 & -88200 & 44100 \end{bmatrix}. \tag{2.31}$$

The entries are large compared with the entries of $\mathbf{H}_5$. Therefore some perturbation in the right-hand vector $\mathbf{b}$, gets magnified when multiplied by $\mathbf{H}_5^{-1}$. The condition number of $\mathbf{H}_5$ is $\text{cond}(\mathbf{H}_5) = 4.7661e + 05$ which increases rapidly with n.

n	$\mathrm{cond}(\mathbf{H}_n)$
5	4.766072e+05
6	1.495106e+07
7	4.753674e+08
8	1.525758e+10
9	4.931544e+11
10	1.602529e+13
11	5.223946e+14
12	1.794510e+16

Table 2: Condition Numbers of Hilbert Matrices

This table shows the condition numbers of the Hilbert matrices for $n = 5, 6, 7, 8, 9, 10, 11, 12$. We see the value of the condition number increse when n also increase.

3. Least-squares Solutions to Linear Systems

In several practical situations, we need to solve a linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ where the matrix $\boldsymbol{A}$ is nonsquare and/or singular.In such cases, solutions may not exist at all; in cases where there are solutions, there may be infinitely many solutions. For example, when $\boldsymbol{A}$ is $m \times n$ and $m > n$, we have an overdetermined system (that is, the number of equations is greater than the number of unknowns), and an overdetermined system typically has no solution. In contrast, an underdetermined system ($m < n$) typically has an infinite number of solutions.

In these cases, the best we can hope for is to find a vector $\mathbf{x}$ that will make $\boldsymbol{A}\mathbf{x}$ as close as possible to the vector $\mathbf{b}$. In other words, we seek a vector $\mathbf{x}$ such that $\| \mathbf{r} \| = \| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|$ is minimized. When the 2-norm is used, this solution is referred to as a **least-squares solution** to the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$. The problem of finding least-squares solutions to the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ is known as the **linear least-squares problem** (LSP). The linear least-squares problem is formally defined as follows.

Definition 3.1. [10] Given an $m \times n$ matrix $\boldsymbol{A}$ and m-vector $\mathbf{b}$, the least-squares problem is to find an n-vector $\mathbf{x}$ such that the norm of the residual vector, $\| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2$, is as small as possible.

If the least-squares problem has more than one solution, the one having the 2-norm is called the **minimum norm solution**.

The least-squares solution $\widehat{\mathbf{x}}$, for the the linear system

$$\boldsymbol{A}\mathbf{x} = \mathbf{b}, \tag{3.1}$$

is given by minimizing the 2-norm square of the residual $\boldsymbol{A}\mathbf{x} - \mathbf{b}$, that is, the quantity

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

So

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

The two middle terms are equal, and the minimum is found at the zero of the derivative with respect to $\mathbf{x}$,

$$2\boldsymbol{A}^T \boldsymbol{A}\widehat{\mathbf{x}} - 2\boldsymbol{A}^T \mathbf{b} = 0. \tag{3.4}$$

Then

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

Thus the least-squares solution is

$$\widehat{\mathbf{x}} = \left(\boldsymbol{A}^T \boldsymbol{A}\right)^{-1} \boldsymbol{A}^T \mathbf{b}. \tag{3.6}$$

[10] Definition 3.1 can be found in Ref. [3]

Example 3.2. [11] Find the least-squares solution of $\boldsymbol{A}\mathbf{x} = \mathbf{b}$, where

$$\boldsymbol{A} = \begin{bmatrix} 1 & 274 & 2450 \\ 1 & 180 & 3254 \\ 1 & 375 & 3802 \\ 1 & 205 & 2838 \\ 1 & 86 & 2347 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 162 \\ 120 \\ 223 \\ 131 \\ 67 \end{bmatrix}. \tag{3.7}$$

Solution 3.2. this least-squares solution is given by

$$\text{Step 1: Form} \quad \boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 5 & 1120 & 14,691 \\ 1120 & 297,522 & 3466,420 \\ 14,691 & 3466,402 & 44608,873 \end{bmatrix}. \tag{3.8}$$

$$\text{Step 2: Form} \quad \boldsymbol{A}^T\mathbf{b} = \begin{bmatrix} 703 \\ 182,230 \\ 2164,253 \end{bmatrix}. \tag{3.9}$$

$$\text{Step 3: Solve the equation:} \boldsymbol{A}^T\boldsymbol{A}\mathbf{x} = \boldsymbol{A}^T\mathbf{b}. \tag{3.10}$$

Then

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7.0325 \\ 0.5044 \\ 0.0070 \end{bmatrix}. \tag{3.11}$$

This system is artificially ill-conditioned because the column of $\boldsymbol{A}$ are out of scale.

$$\text{cond}(\boldsymbol{A}) = 1.7576 \times 10^4 \qquad \text{(large)}. \tag{3.12}$$

$$\text{cond}(\boldsymbol{A}^T\boldsymbol{A}) = 3.0891 \times 10^8 \qquad \text{(large)}. \tag{3.13}$$

To see how the computed least-squares solution agree with the data of the vector $\mathbf{b}$, we compute

$$\begin{bmatrix} 1 & 274 & 2450 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 162.4043. \tag{3.14}$$

$$\text{(True value} = 162\text{)}$$

$$\begin{bmatrix} 1 & 180 & 3254 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 120.6153. \tag{3.15}$$

$$\text{(True value} = 120\text{)}$$

[11] Example 3.2 can be found in Ref. [3]

$$\begin{bmatrix} 1 & 375 & 3802 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 222.8193. \tag{3.16}$$

(True value = 223)

3.1 Singular Value Decomposition

The **singular value decomposition** (SVD) of a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ is of great theoretical and practical importance. It is a matrix factorization with applications in many algorithms. Its history goes back more than a century, but its use in numerical computations is much more recent [1].

Definition 3.3. (Singular Value Decomposition) For any real $m \times n$ matrix $\boldsymbol{A}$, there exists orthogonal matrices an $m \times m$ matrix $\mathbf{U}$ and $n \times n$ matrix $\mathbf{V}$, and $\boldsymbol{\Sigma}$ is $m \times n$ "diagonal" matrix, with entries $\sigma_1, \sigma_2, \cdots, \sigma_n$ called the singular values of $\boldsymbol{A}$ such that

$$\boldsymbol{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^T. \tag{3.17}$$

Comment: $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_n)$ has non-negative diagonal elements appearing in non-increasing order such that $\sigma_1 \geq \cdots \geq \sigma_n \geq 0$.

From above equation 3.17 we can get the form

$$\boldsymbol{A}\mathbf{V} = \mathbf{U}\boldsymbol{\Sigma}, \tag{3.18}$$

such that

$$\boldsymbol{A}[\mathbf{v}_1\mathbf{v}_2 \cdots \mathbf{v}_n] = [\mathbf{u}_1\mathbf{u}_2 \cdots \mathbf{u}_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n \\ 0 & \cdots & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & \cdots & \cdots & 0 \end{bmatrix}. \tag{3.19}$$

So

$$\boldsymbol{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i \qquad i = 1, 2, \cdots, n. \tag{3.20}$$

Where the vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are called the right and left singular vectors of $\boldsymbol{A}$, respectively.

$$\boldsymbol{A} = \sum_{i=1}^{n} \sigma_i \mathbf{u}_i \mathbf{v}_i^T \tag{3.21}$$

and

$$\boldsymbol{A}^T = \sum_{i=1}^{n} \sigma_i^T \mathbf{v}_i \mathbf{u}_i^T, \tag{3.22}$$

where $n =$ the numerical $\mathrm{rank}(\boldsymbol{A})$ and $\mathbf{u}_i$ and $\mathbf{v}_i$ are respectively the i-th columns of $\mathbf{U}$ and $\mathbf{V}$.

Comment: The numerical rank of $\boldsymbol{A}$ in this case is equal to the number of non-zero singular values.

Theorem 3.4. [12] *Let $\boldsymbol{A} = \mathbf{U}\Sigma\mathbf{V}^T$ be the singular value decomposition of an $m \times n$ matrix $\boldsymbol{A}(m \geq n)$. Let r be the rank of the matrix $\boldsymbol{A}$. Then*

1. $\mathbf{V}^T(\boldsymbol{A}^T\boldsymbol{A})\mathbf{V} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2, 0, \cdots, 0)_{n\times n}$.

2. $\mathbf{U}^T(\boldsymbol{A}\boldsymbol{A}^T)\mathbf{U} = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \cdots, \sigma_r^2, 0, \cdots, 0)_{m\times m}$.

From this Theorem 3.4 it immediately follows that

1. *The left singular vectors $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are the unit eigenvectors of the matrix $\boldsymbol{A}^T\boldsymbol{A}$.*
2. *The right singular vectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_n$ are the unit eigenvectors of the matrix $\boldsymbol{A}\boldsymbol{A}^T$.*
3. *$\sigma_1^2, \cdots, \sigma_r^2$ are the nonzero eigenvalues of both $\boldsymbol{A}^T\boldsymbol{A}$ and $\boldsymbol{A}\boldsymbol{A}^T$.*

Example 3.5. Let

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}. \tag{3.23}$$

The singular value decomposition gives

$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.24}$$

Then

$$\boldsymbol{A}\boldsymbol{A}^T = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.25}$$

and it's corresponding vector $\mathbf{U}$ is

$$\mathbf{U} = \begin{bmatrix} 0.00000 & 0.70711 & 0.70711 \\ -1.00000 & -0.00000 & 0.00000 \\ 0.00000 & -0.70711 & 0.70711 \end{bmatrix}. \tag{3.26}$$

[12] Theorem 3.4 can be found in Ref. [3]

Also

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \tag{3.27}$$

and it's corresponding vector $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} -0.70711 & 0.70711 \\ -0.70711 & -0.70711 \end{bmatrix}. \tag{3.28}$$

So

$$\boldsymbol{A} = \mathbf{U}\Sigma\mathbf{V}^T = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}. \tag{3.29}$$

Example 3.6. Let

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & -1 \\ 4 & 3 & 1 \\ 6 & 7 & -1 \end{bmatrix}. \tag{3.30}$$

The singular value decomposition gives

$$\Sigma = \begin{bmatrix} 10.67963 & 0.00000 & 0.00000 \\ 0.00000 & 1.98632 & 0.00000 \\ 0.00000 & 0.00000 & 0.00000 \end{bmatrix}. \tag{3.31}$$

Then

$$\boldsymbol{A}\boldsymbol{A}^T = \begin{bmatrix} 6 & 9 & 21 \\ 9 & 26 & 44 \\ 21 & 44 & 86 \end{bmatrix}, \tag{3.32}$$

and it's corresponding vector $\mathbf{U}$ is

$$\mathbf{U} = \begin{bmatrix} -0.20627 & 0.53925 & -0.81650 \\ -0.45417 & -0.79187 & -0.40825 \\ -0.86671 & 0.28662 & 0.40825 \end{bmatrix}. \tag{3.33}$$

Also

$$\boldsymbol{A}^T\boldsymbol{A} = \begin{bmatrix} 53 & 56 & -3 \\ 56 & 62 & -6 \\ -3 & -6 & 3 \end{bmatrix}, \tag{3.34}$$

and it's corresponding vector $\mathbf{V}$ is

$$\mathbf{V} = \begin{bmatrix} -0.676353 & -0.457399 & 0.577350 \\ -0.734296 & 0.357039 & -0.577350 \\ 0.057943 & -0.814438 & -0.577350 \end{bmatrix}. \tag{3.35}$$

So

$$\boldsymbol{A} = \mathbf{U}\Sigma\mathbf{V}^T = \begin{bmatrix} 1 & 2 & -1 \\ 4 & 3 & 1 \\ 6 & 7 & -1 \end{bmatrix}. \tag{3.36}$$

3.2 Solving Linear Systems Using Singular Value Decomposition

The idea of using **SVD** in the solution of the least-squares problem is to determine if a linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ has a solution and, if so, how to compute it.

The least squares solution to the problem **min** $\| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2$ is given by

$$\widehat{\mathbf{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\mathbf{b}, \tag{3.37}$$

substituting $\boldsymbol{A} = \mathbf{U}\Sigma\mathbf{V}^T$, we have

$$\widehat{\mathbf{x}} = \mathbf{V}\Sigma^{+}\mathbf{U}^T\mathbf{b} \tag{3.38}$$

$$\widehat{\mathbf{x}} = \sum_{i=1}^{n} \sigma_i^{+}\mathbf{u}_i^T\mathbf{b}\mathbf{v}_i$$

$$= \sum_{i=1}^{n} \frac{\mathbf{u}_i^T\mathbf{b}}{\sigma_i}\mathbf{v}_i, \tag{3.39}$$

where $n =$ the numerical **rank**$(\boldsymbol{A})$ and $\mathbf{u}_i$ and $\mathbf{v}_i$ are respectively the i-th columns of $\mathbf{U}$ and $\mathbf{V}$, [3].

3.2.1 Singular Value Decomposition and Stability of Linear Systems

We will now analyze the influence of the coefficient matrix of the linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ on the solution using singular value decomposition (SVD) [9].

$$\boldsymbol{A} = \mathbf{U}\Sigma\mathbf{V}^T = \boldsymbol{A} = \sum_{i=1}^{n} \sigma_i\mathbf{u}_i\mathbf{v}_i^T, \tag{3.40}$$

and

$$\widehat{\mathbf{x}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\mathbf{b} = \sum_{i=1}^{n} \sigma_i^{+}\mathbf{u}_i^T\mathbf{b}\mathbf{v}_i. \tag{3.41}$$

Hence if σ_L (The largest singular value of $\boldsymbol{A}$) is small then small changes in $\boldsymbol{A}$ or $\mathbf{b}$ will cause a significant change in $\mathbf{x}$. As we showed before in section 2.2, the conditioned number of a matrix $\boldsymbol{A}$ is defined as:

$$\mathrm{cond}(\boldsymbol{A}) = \| \boldsymbol{A} \| \| \boldsymbol{A}^{-1} \|, \tag{3.42}$$

and for the matrix 2-norm,

$$\mathrm{cond}(\boldsymbol{A}) = \| \boldsymbol{A} \|_2 \| \boldsymbol{A}^{-1} \|_2 = \frac{\sigma_L(\boldsymbol{A})}{\sigma_S(\boldsymbol{A})}. \tag{3.43}$$

Where σ_L and σ_S is the largest and smallest singular values of $\boldsymbol{A}$ respectively.

($\| \boldsymbol{A} \|_2$ is the square root of the largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$ which is the largest singular value of $\boldsymbol{A}$, see equation 1.12).

3.3 Numerical Rank

In practical application that need singular values, we have to know when to accept a computed singular value to be "zero". Accept a computed singular value to be zero if it is less than or equal to $10^{-t} \| \boldsymbol{A} \|_\infty$, where the entries of $\boldsymbol{A}$ are correct to t digits. Having defined a tolerance $\delta = 10^{-t} \| \boldsymbol{A} \|_\infty$ for a zero singular value, we can make the following convention (Golub and Van Loan 1989, 247):

$\boldsymbol{A}$ has "numerical rank" n if the computed singular values $\sigma_1, \sigma_2, \cdots, \sigma_r$ satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \delta \geq \sigma_{n+1} \geq \sigma_r \tag{3.44}$$

Thus, to determine the numerical rank of a matrix $\boldsymbol{A}$, count the "large" singular values only. if this number is n, then $\boldsymbol{A}$ has numerical rank n.

Remark: Note that finding the numerical rank of a matrix will be "tricky" if there is no suitable gap between a set of singular values [3].

4. Regularization Methods

In this chapter, we give a brief introduction to Rank-deficient and discrete ill-posed problems. We introduce the regularization methods to improve the accuracy of the solution of ill-conditioned linear systems. We focussed in two methods, e.g. truncated singular value decomposition **(TSVD)** and Tikhonov methods.

4.1 Rank-deficient and Ill-posed Problems

The numerical treatment of system of equations with an ill-conditioned coefficient matrix depends on the type of ill-conditioning of matrix $\boldsymbol{A}$.
There are two important classes of problems to consider, and many practical problems belong to one of these two classes [4].

1. **Rank-deficient problems** are characterized by the matrix $\boldsymbol{A}$ having a cluster of small singular values, and there is well-determined gap between large and small singular values. This implies that one or more rows and columns of $\boldsymbol{A}$ are nearly linear combinations of some or all of the remaining rows and columns.

2. **Discrete ill-posed problem**. Here, all the the singular values of $\boldsymbol{A}$ gradually decay to zero, also the ratio between the largest and the smallest nonzero singular values is large and there is no notion of a numerical rank for these matrices. The primary difficulty with the discrete ill-posed problems is that they are essentially underdetermined due to the cluster of small singular values of the matrix $\boldsymbol{A}$. Hence it is necessary to incorporate further information about the desired solution in order to stabilise the problem and to single out a useful and stable solution. This is the purpose of regularization.

4.2 Regularization Methods

One way of getting a useful solution to discrete ill-posed or rank-deficient linear problems is by regularization. The main idea of regularization is to find a new problem or method that damps the effect of the noise in the input data and also to make a solution of ill-conditioning linear system $\boldsymbol{A}\mathbf{x} = \mathbf{b}$ more " regular " or smooth. For the solution of ill-conditioned problems the ordinary methods give unacceptable solutions, so that regularization methods are necessary. Since many discussion of ill-conditioned matrices requires knowledge of the **SVD** of the matrix $\boldsymbol{A}$, our essay is mainly focused on **Truncated singular value decomposition** (TSVD) and **Tikhonov** regularization methods.

4.2.1 The Tikhonov Method

Tikhonov regularization method is the most commonly used method of regularization of ill-posed problem. The Tikhonov regularization method involves obtaining the exact or a least-squares solution of linear systems by minimizing the functional

$$\Phi(\mathbf{x}) = \| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2^2, \tag{4.1}$$

subject to $\| L\mathbf{x} \| \leq \beta$, where L is a differential operator.

Standard algorithms often give solutions that decay very rapidly, with large positive and negative values. To stabilize the computation, we add to equation (4.1) a term that penalizes the large components and thereby reduces them. Thus instead of equation (4.1) we minimize the expression

$$\Phi_\alpha(\mathbf{x}) = \| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2^2 + \alpha \| L\mathbf{x} \|_2^2, \tag{4.2}$$

where $\alpha > 0$ is called a regularization parameter. we denote it as

$$\mathbf{x}_\alpha = \textbf{arg min} \left\{ \| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2^2 + \alpha \| L\mathbf{x} \|_2^2 \right\}. \tag{4.3}$$

We consider the case when $L = I$, the identity operator.

$$\Phi_\alpha(\mathbf{x}) = \| \boldsymbol{A}\mathbf{x} - \mathbf{b} \|_2^2 + \alpha \| \mathbf{x} \|_2^2, \tag{4.4}$$

Regularization can be understood as a balance between two requirements [8]:

1. $\mathbf{x}$ should give a small residual $\boldsymbol{A}\mathbf{x}_\alpha - \mathbf{b}$.
2. $\mathbf{x}$ should be small in 2-norm.

This means that the Regularization should balance between these two quantaties $\| \boldsymbol{A}\mathbf{x}_\alpha - \mathbf{b} \|_2$ and $\| \mathbf{x}_\alpha \|_2$. Thus regularization parameter $\alpha > 0$ can be used to "tune" the balance.

We can rewrite equation 4.4 as

$$\begin{aligned} \Phi_\alpha(\mathbf{x}) &= (\boldsymbol{A}\mathbf{x} - \mathbf{b})^T(\boldsymbol{A}\mathbf{x} - \mathbf{b}) + \alpha \mathbf{x}^T\mathbf{x} \\ &= (\boldsymbol{A}\mathbf{x})^T(\boldsymbol{A}\mathbf{x}) - \mathbf{b}^T(\boldsymbol{A}\mathbf{x}) - (\boldsymbol{A}\mathbf{x})^T\mathbf{b} + \mathbf{b}^T\mathbf{b} + \alpha \mathbf{x}^T\mathbf{x}. \end{aligned}$$

The two terms $\mathbf{b}^T(\boldsymbol{A}\mathbf{x})$ and $(\boldsymbol{A}\mathbf{x})^T\mathbf{b}$ are equal, so

$$\Phi_\alpha(\mathbf{x}) = (\boldsymbol{A}\mathbf{x})^T(\boldsymbol{A}\mathbf{x}) - 2(\boldsymbol{A}\mathbf{x})^T\mathbf{b} + \alpha \mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{b} \tag{4.5}$$

$$= (\boldsymbol{A}\mathbf{x})^T(\boldsymbol{A}\mathbf{x}) - 2\mathbf{x}^T\boldsymbol{A}^T\mathbf{b} + \alpha \mathbf{x}^T\mathbf{x} + \mathbf{b}^T\mathbf{b}. \tag{4.6}$$

The minimum is found at $\frac{\partial \Phi_\alpha(\mathbf{x})}{\partial \mathbf{x}} |_{\mathbf{x}=\mathbf{x}_\alpha} = 0$. So

$$\frac{\partial \Phi_\alpha(\mathbf{x})}{\partial \mathbf{x}} |_{\mathbf{x}=\mathbf{x}_\alpha} = 2\boldsymbol{A}^T\boldsymbol{A}\mathbf{x}_\alpha - 2\boldsymbol{A}^T\mathbf{b} + 2\alpha\mathbf{x}_\alpha = 0. \tag{4.7}$$

So

$$(\boldsymbol{A}^T\boldsymbol{A} + \alpha\mathbf{I})\mathbf{x}_\alpha = \boldsymbol{A}^T\mathbf{b}, \tag{4.8}$$

Then

$$\mathbf{x}_\alpha = (\boldsymbol{A}^T\boldsymbol{A} + \alpha I)^{-1}\boldsymbol{A}^T\mathbf{b}. \tag{4.9}$$

Using the singular value decomposition (SVD) of $\boldsymbol{A}$ in equation 4.9 becomes

$$\mathbf{x}_\alpha = \sum_{i=1}^{n} \left[\left(\mathbf{v}_i\sigma_i^T\mathbf{u}_i^T\right)\left(\mathbf{u}_i\sigma_i\mathbf{v}_i^T\right) + \alpha I\right]^{-1}\left(\mathbf{v}_i\sigma_i^T\mathbf{u}_i^T\right)\mathbf{b}, \tag{4.10}$$

so

$$\mathbf{x}_\alpha = \sum_{i=1}^{n} \left(\sigma_i^2 + \alpha\right)^{-1}\left(\mathbf{v}_i\sigma_i\mathbf{u}_i^T\right)\mathbf{b}, \tag{4.11}$$

then

$$\mathbf{x}_\alpha = \sum_{i=1}^{n} \frac{\sigma_i}{\alpha + \sigma_i^2}\left(\mathbf{u}_i^T\right)\mathbf{b}\mathbf{v}_i. \tag{4.12}$$

The effect of the addition in equation (4.4) is to dampen the contribution of the term involving small singular values, so that instead of cutting them off altogether, we modify the method to reduce their impact. A small α has very little effect on the components associated with large σ_i, since for $\alpha \ll \sigma_i^2$

$$\frac{\sigma_i}{\alpha + \sigma_i^2} \cong \frac{1}{\sigma_i}. \tag{4.13}$$

On the other hand, if $\sigma_i^2 \ll \alpha$, then

$$\frac{\sigma_i}{\alpha + \sigma_i^2} \cong \frac{\sigma_i}{\alpha} \ll \frac{1}{\sigma_i}, \tag{4.14}$$

so that the magnification of the components associated with the small singular values is reduced. With a good choice of α, one can then hope to get a relatively smooth solution that is still a reasonably good approximation to the true solution. This is called **SVD** with damping [6].

4.2.2 Truncated Singular Value Decomposition (TSVD)

TSVD is another commonly used method of regularization of Rank-deficient problems and Discrete ill-posed problems. The idea of truncating the **SVD** has been treated as a problem of determining the **"numerical rank"** of the matrix $\boldsymbol{A}$ ($\boldsymbol{A}$ is a noisy representation of a mathematically rank-deficient matrix). All of the computed singular values smaller than some threshold value (α) are treated as zeros which were corrupted by rounding errors into small non-zero quantities. Thus if σ_k is the smallest singular value greater than the truncation threshold $(\sigma_k \geq \alpha)$, then one replace the matrix Σ by a truncated matrix [10].

$$\boldsymbol{\Sigma}_{tr} = \mathrm{diag}(\sigma_1, \cdots, \sigma_k, 0, \cdots, 0) \tag{4.15}$$

whose inverse is given by

$$\boldsymbol{\Sigma}_{tr}^{+} = \mathrm{diag}(\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_k}, 0, \cdots, 0). \tag{4.16}$$

The value k is the **"numerical rank"** of $\boldsymbol{A}$. So we replace the matrix $\boldsymbol{A}$ by its rank k approximation. Then $x_{tr} = x_\alpha$ is the estimate of the solution to the truncated problem, given by

$$\mathbf{x}_{tr} = \mathbf{x}_\alpha = \sum_{i=1}^{k} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \tag{4.17}$$

Example 4.1. Consider the linear system

$$\mathbf{H}_{12}\mathbf{x} = \mathbf{b}. \tag{4.18}$$

where $\mathbf{H}_{12}$ is the 12×12 Hilbert matrix, and

$$\mathbf{b} = \sum_{j=1}^{n} \mathbf{h}_{ij}\mathbf{x}_j. \tag{4.19}$$

This linear system has an exact solution $\mathbf{x} = [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]^T$. The computed solution is given by

$$\widehat{\mathbf{x}} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i. \tag{4.20}$$

Then

$$\begin{aligned}\widehat{\mathbf{x}} = [&1.000000004881659,\ 0.999999523277221,\ 1.000011453630246,\ 0.999883148383246,\\ &1.000618184166931,\ 0.998167348978474,\ 1.003001794695910,\ 0.997860552739088,\\ &0.999135295872923,\ 1.002809344038033,\ 0.998028938426422,\ 1.000484414235235]^T.\end{aligned}$$

The maximum error is

$$\| \mathbf{x} - \widehat{\mathbf{x}} \|_\infty = 0.00300179469591044. \tag{4.21}$$

The error is contaminated with small singular values. The singular values of H_{12} are

$numbers$	Singular values of H_{12}
1	1.79537205956200
2	$3.80275245955037 \times 10^{-1}$
3	$4.47385487521811 \times 10^{-2}$
4	$3.72231223789117 \times 10^{-3}$
5	$2.33089089021769 \times 10^{-4}$
6	$1.11633574832344 \times 10^{-5}$
7	$4.08237611034491 \times 10^{-7}$
8	$1.12286106662802 \times 10^{-8}$
9	$2.25196453231559 \times 10^{-10}$
10	$3.11134836868481 \times 10^{-12}$
11	$2.64910549271184 \times 10^{-14}$
12	$1.09849454516192 \times 10^{-16}$

However the Truncated Singular Value Decomposition (TSVD) method with $\alpha = 10^{-8}$, yields better solution as shown below. The maximum error ($\| \mathbf{x} - \mathbf{x}^{tr} \|_\infty$) for $k = 12$, 11, 10, 9, 8 and 7 are shown in the table below.

k	maximum error
12	$0.269705081406381 \times 10^{-1}$
11	$6.72224601423688 \times 10^{-3}$
10	$1.18729236775472 \times 10^{-5}$
9	$3.84094700600635 \times 10^{-6}$
8	$2.90489981684683 \times 10^{-5}$
7	$1.62703616404802 \times 10^{-4}$

From the table above we notice that the maximum error decreases from $k = 11$ to 9 and after that starts increasing again. Thus we can choose the best regularized solution to be the one corresponding to $k = 9$.

Plugging these singular values into Equation (4.17) it is now easy to see that inclusion of the smallest singular values will cause the sum in Equation (4.17) to blow up yielding a large error in the computed solution. However if these small singular values are cut off we remove these large errors. But as we cut off more and more singular values there comes a point where the singular values is not a faithful representation of the coefficient matrix and hence we need to stop at some minimum value of k. This minimum value of k is equal to 9 for this example linear system.

Figure (4.1) is a plot of computed solutions obtained (by octave) using the TSVD method. The three plots in the figure represent solutions after truncating 1, 2, and 3 smallest singular values one after the other.

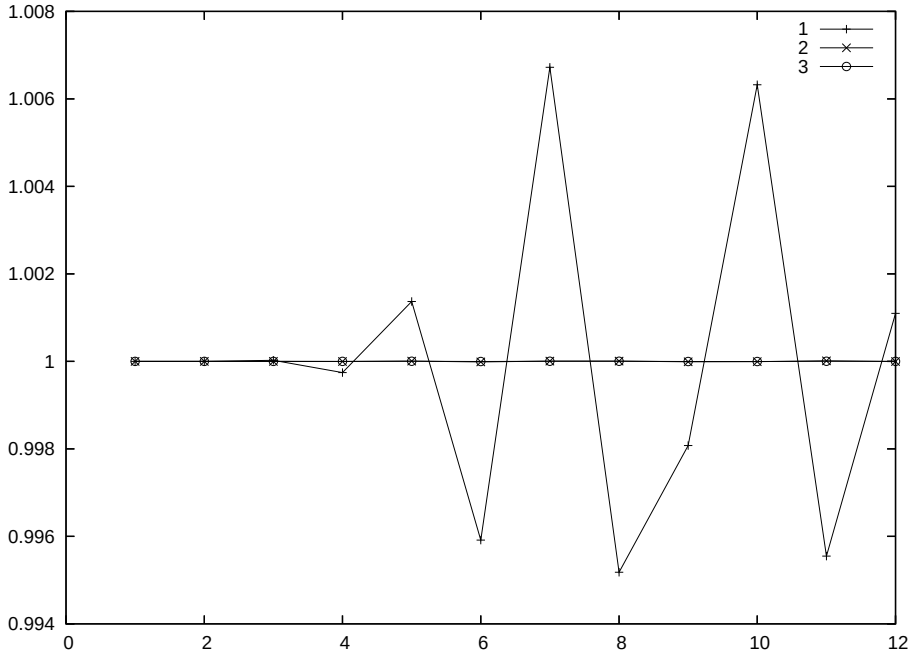

Figure 4.1: Plot of computed regularized solutions using the TSVD method. The smallest 1, 2, and 3 singular values were cut off one after the other.

5. Conclusion

Finally we would like to give a brief summary to this study. In this essay we have briefly introduced the reader to various aspects of ill-conditioned linear systems and their regularization methods. In chapter one, we introduced the linear systems, some basic concepts of linear algebra., and the conditioned number of a matrix. In chapter two, we defined ill-conditioned linear systems, For the purpose of illustration, we have included two (linear) examples for ill-conditioned linear systems, (e.g. the Hilbert and the Vandermonde matrices). Also in chapter three we defined the least-squares solutions to linear systems and singular value decomposition (SVD) as a numerical "tool", which is suited for analysis of the ill-posed problems. Also we identified that the computed solution $\widehat{\mathbf{x}}$ is usually meaningless approximation to $\mathbf{x}$ due to the error in the right-hand side $\mathbf{b}$ and the severe ill-conditioning of the matrix $\boldsymbol{A}$. For this reason, in chapter four we reviewed the regularization methods to improve the accuracy of the solution of ill-posed problems. We focussed in two methods, e.g. truncated singular value decomposition (TSVD) and Tikhonov mehod. One hopes that this work could be extended further to deal with other types of regularization methods.

Appendix A. Truncated Singular Value Decomposition (TSVD)

Here is the octave code for TSVD example and we insert also the results of it as follow.

```
1;

function x = TSVD(A, b, q)
[m, n] = size(A) ;
[U, S, V] = svd(A) ;
s = diag(S) ;


x = zeros(n, 1) ;
for i = 1:q

x = x + (b' * U(:, i)/s(i)) * V(:, i) ;
if x<= 10**(-8)
x=0;
endif

endfor
endfunction

N = 12
H12 = hilb(N)
[U, S, V] = svd(H12)

% The exact solution denoted by x.
x=ones(N,1)
b = H12 * x

%The computed solution denoted by xc.
xc = H12 \ b
xt =[];
for k = N:-1:1

%The TSVD solution denoted by xt.
 xt = [xt TSVD(H12,b,k)];
endfor
xt
```

```
xt =

  Columns 1 through 4:

  1.000000035410423  1.000000010406728  1.000000000128169  0.999999999993725
  0.999995761625934  0.999998969286971  0.999999989054960  0.999999999145378
  1.000126836288086  1.000025022058407  1.000000216120084  1.000000032930959
  0.998345365750319  0.999742665937545  0.999998289259415  0.999999654919425
  1.011673175050563  1.001370185851539  1.000006491267828  1.000001539687882
  0.950432454189308  0.995915975213462  0.999988127076322  0.999996861380750
  1.133947503416035  1.006722246014237  1.000007150407346  1.000002059853554
  0.764130473439041  0.995179805399686  1.000006868068255  1.000002031900902
  1.269705081406381  0.998077220444046  0.999991908851701  0.999997470810440
  0.806915643063877  1.006322162812165  0.999993629378929  0.999997927375082
  1.078626086420750  0.995549233050377  1.000011456491165  1.000003840947006
  0.986101573342976  1.001096511297768  0.999995873687303  0.999998580958273

 Columns 5 through 8:

 0.999999996461529  1.000000151890956  0.999995319603844  1.000106069091871
  1.000000194093540  0.999994106412219  1.000122335866062  0.998253353494705
  0.999997482649582  1.000052197954546  0.999313675155133  1.005457156224012
  1.000012680488899  0.999837296383595  1.001085437561335  0.998057332941534
  0.999972852028113  1.000150789068384  1.000192395006243  0.995467217244331
  1.000016557348865  1.000096569609180  0.999253083336022  0.997356157419837
  1.000018375576711  0.999906488482612  0.999226991564590  1.000712066947196
  0.999987654689213  0.999851527117142  0.999871521112027  1.003392636567184
  0.999977437565833  0.999969007644730  1.000610025132260  1.004301323177792
  1.000003199234471  1.000132143083218  1.000919711186439  1.003036848655090
  1.000029048998168  1.000154263950493  1.000439206328687  0.999587308252251
  0.999984517595727  0.999855303160945  0.998964942748598  0.994134056971875

Columns 9 through 12:

  0.998220027713585  1.021853329482553  0.811827016949171  1.968193407412721
  1.016699883351630  0.901826902329911  1.267158129469560  1.211355960941289
  0.978446218876212  0.999298041036143  1.258596887316880  0.907797344126588
  0.984100114089535  1.053331189093677  1.182716622765847  0.735766433013591
  0.999596918273331  1.070010484453778  1.099090617957318  0.622574166277303
  1.011852504960719  1.064191004336744  1.020782412083546  0.541515632961854
  1.017616517848767  1.045650381835735  0.950479056603555  0.480189690911242
  1.017015122462759  1.020207748562225  0.888094362825359  0.431961081053743
  1.011136208465709  0.991290936630181  0.832808873429604  0.392922531207807
  1.001189570053022  0.960928856644816  0.783684307656640  0.360607676857949
  0.988242349596960  0.930325825448542  0.739850628891383  0.333375279209793
  0.973160612317457  0.900192341869936  0.700553560084826  0.310086830520352
```

Acknowledgements

I would like to express my deepest thanks and gratitude to my supervisor professor Francis Benyah (Department of Math's and Appl. Math's, The University of the Western Cape) who has directed me through this work with an admirable magnitude of skill , care and scientific patience , whose constructive comments , critical remarks and sound advice were simply indispensable for the completion of this work.

Many thanks to Mr. Henri Amuasi, Mr. Eihab basheir, Mr. Ambrose Chongo and Mr. Jan
Groenewald for their help and useful discussions towards getting this work done.

****

I dedicate this work to my family whose love, care, and encouragement helped me a lot.

إِلي أُمي ... مَن أَرضعتني الحب و الحنَان

إِلي أَبِي ... مَن علمني معني العطَاء و الوفَاء

إِلي إِخوتي ... رُفقَاء دَربي في الكفَاح

إِلي كُلِ مَن علمني حرفًا و زَودَني بِنورَ العِلم

إِليكم جميعًا ... أُهدي ثمرةَ جُهدي هذَا المُتوَاضِع ،،،

شيمَاء ،،،

Bibliography

[1] ACM/ESE. http://www.gps.caltech.edu/ tapio/acm118/handouts/svd.pdf. Singular Value Decomposition.

[2] Francis Benyah. *Ill-Conditioning and Regularization in the Solution of Linear Systems*. Seminar, African Institute for Mathematical Science, 2004.

[3] Biswa Nath Data. *Numerical Linear Algebra and Applications*. Brooks/Cole Publishing Company, Calfornia, 1995.

[4] Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. Siam, Society for Industrial and Applied Mathematics, philadelphia, 1998.

[5] Autar K.Kaw. *Introduction to Matrix Algebra*. University of South Florida, USA, 2002.

[6] Peter Linz and Richard L. C. Wang. *Exploring Numerical Methods*. Jones and Bartlett Publishers, California, 2003.

[7] Francis J. Narcowich. http://www.math.tamu.edu/ fnarc/psfiles/rank2005.pdf. The Rank of a Matrix.

[8] Tikhonov. http://matriisi.ee.tut.fi/courses/mat-52500/tikhonov.pdf. Tikhonov REgularization.

[9] Website. http://www.cs.huji.ac.il/ csip/condition.pdf. Numerical Stability - The Condition Number of a Matrix.

[10] Bert W.Rust. http://math.nist.gov/ brust/pubs/truncsvd/ms-truncsvd.ps. Truncating The Singular Value Decompositin For ill-Posed Problems.