ON DISCRETE LINEAR SYSTEMS

METHODS OF REACHABILITY, OBSERVABILITY AND STABILITY. SOME THEORETICAL AND PRACTICAL ASPECTS.

HALLGRÍMUR ÓSKARSSON

Bachelor's thesis
2019:K26

LUND UNIVERSITY

Faculty of Science
Centre for Mathematical Sciences
Numerical Analysis

Abstract

Linear time-invariant systems are ordinary differential equation systems that arise in control engineering where they are used to model e.g. signal processing, chemical processing and economics. A study is conducted on linear time-invariant systems, their solution and three key properties they have: Observability, reachability and stability. Three algorithms, Gaussian elimination, singular value decomposition and QR decomposition, are studied for their effectiveness to determine whether a system is reachable and/or observable, and examples are given to show why the singular value decomposition is the preferred method.

Populärvetenskaplig Sammanfattning

Linjära tids-invarianta system är ordinära differential ekvation system som förekommer inom Reglerteknik där de används för att modellera blandt annat signal-processering, kemiska processer och ekonomi. Ett studie utförs på linjära tids-invarianta system, deras lösningar och tre nyckel egenskaper som de besitter: Observabilitet, åtkomlighet, och stabilitet. Tre algoritmer, Gaussian elimination, singular värde dekomposition och QR dekomposition studeras för att bedömma om ett system är åtkomligt och/eller observerbart, och exempel ges för att visa varför singularvärde dekomposition är den föredragna metoden.

Contents

Introduction

Linear time-invariant systems are mathematical models based on first order linear differential equations that arise in many fields such as physics, biology and economics among others. Such systems can e.g. describe the voltage in an RLC electrical circuit, or how the balance of a bank account with compounded interest evolves over time. The system has three key components, an input, a state and an output.

A central theme of importance regarding linear time-invariant systems and their solutions is whether the system is:

- Reachable, that is whether any state in the state space of the system can be reached from the origin, using a suitable input.
- Observable, i.e. can the system's initial state be determined simply by observing the outputs of the system.
- Stable, i.e. is the solution of the system stable.

In this thesis each of these properties is defined and methods are provided to determine whether a system has them. These properties will be derived for discrete linear time-invariant systems only, but they are analog for continuous systems and thus the results provided will work for those too. Finally some algorithms used in the computation of these properties will be studied along with their effectiveness

Chapter 1 will introduce the linear time-invariant system, what its components are and what a general solution to the system is. Chapter 2 will define reachability and observability and properties that are used to determine them. Chapter 3 introduces the Z-transform, a method to find a solution to a linear time-invariant system that converts the system to a rational polynomial equation. Chapter 4 will introduce stability, two of its many definitions, how stability is determined and how the two definitions discussed relate to each other. Finally chapter 5 will discuss common algorithms used in numerical methods that are used in the determination of reachability and observability. Examples will be provided in each chapter to demonstrate their meaning.

Chapter 1

The Linear Time-Invariant System

1.1 The Linear Time-Invariant System Structure

The linear time-invariant system is a system of first order linear differential or difference equation with constant coefficients. It is often called the *State vector equation*. It is of the form [9, p. 2]:

$$\begin{aligned} x(k+1) &= Ax(k) + bu(k),\ x(0) = x_0 \\ y(k) &= cx(k) + du(k) \end{aligned} \tag{1.1}$$

where:

- A is an $n \times n$ matrix.
- b and c are n column and row vectors respectively.
- d is a scalar.
- $x(k)$ is an n column vector called *the state vector*.
- $u(k)$ is a scalar called *the input*.
- $y(k)$ is a scalar called *the output*.
- $u(k)$ and $y(k)$ are often referred to as input and output *signals*.
- x_0 is the *initial state* of $x(k)$.
- k is the independent variable.

The continuous case is noted the same, except it uses t as the independent variable. Such a system is often denoted (A, b, c, d). The two most common inputs are the *unit impulse signal* and the *unit step signal* [7, p. 30]:

$$\delta(k) = \begin{cases} 1, & \text{if } k = 0 \\ 0, & \text{otherwise} \end{cases} \qquad \text{The unit impulse signal}$$

$$u(k) = \begin{cases} 0, & \text{if } k < 0 \\ 1, & \text{if } k \geqslant 0 \end{cases} \qquad \text{The unit step signal}$$

Both signals appear often in physical systems and the unit impulse signal has the special property that any other input can be written as a sum of unit impulses.

The system above is called a *single-input, single-output system* as opposed to a multi-input, multi-output system where b, c and d can be matrices (denoted by B, C and D) instead of vectors. Such a system arises as well in control theory but the focus here will be on the single input, single output system since the theory discussed here applies to both systems.

1.2 Solutions to a Discrete Linear Time-Invariant System

A general solution to the state of (1.1) is found with a simple iterative process:

$$\begin{aligned} x(0) &= x_0 \\ x(1) &= Ax_0 + bu(0) \\ x(2) &= Ax(1) + bu(1) = A^2x_0 + Abu(0) + bu(1) \\ &\vdots \\ x(k) &= A^kx_0 + \sum_{j=0}^{k-1} A^{k-j-1}bu(j), k \geqslant 1 \end{aligned} \tag{1.2}$$

And in turn, the output is given by:

$$y(k) = cA^kx_0 + \sum_{j=0}^{k-1} cA^{k-j-1}bu(j) + du(k), k \geqslant 1 \tag{1.3}$$

1.3 Examples

The cattle Ranch problem

Consider a simplified model for the population of a cattle farm [3, p. 24]. The life stages of the cattle are split into three age categories: calve, mature cows and old cows. In one year, calves grow into mature cows, and mature cows into old cows (given they're not removed from the population). The mature and old cows reproduce, making a calve over the year. Let:

- $x_1(k)$, $x_2(k)$ and $x_3(k)$ denote the number of calves, mature cows and old cows at year k respectively.
- a_1 and a_2 be the reproduction rate of the mature cows and the old cows respectively.
- a_3, a_4, a_5 be the survival rate of the calves, mature cows and old cows respectively, from one year to the next, which can be affected by e.g. deceases or accidents.
- a_6 the proportional number of mature cows removed from the population at year k, which in turn gives the output of the ranch.
- $u(k)$ is the number of outside calves acquired to be introduced to the population to ensure genetic diversity in the cattle.

Based on these assumptions the number of cows at each live stage in year $k+1$ is given by:

- $x_1(k+1) = a_1x_2(k) + a_2x_3(k) + u(k)$
- $x_2(k+1) = a_3x_1(k) - a_6x_2(k)$
- $x_3(k+1) = a_4x_2(k) + a_5x_3(k)$.

So by letting $x(k+1) = \begin{bmatrix} x_1(k+1) \\ x_2(k+1) \\ x_3(k+1) \end{bmatrix}$, $c = \begin{bmatrix} 0 & a_6 & 0 \end{bmatrix}$ and $b = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, the model can be expressed as:

$$\begin{aligned} x(k+1) &= \begin{bmatrix} 0 & a_1 & a_2 \\ a_3 & -a_6 & 0 \\ 0 & a_4 & a_5 \end{bmatrix} x(k) + \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} u(k) \\ y(k) &= \begin{bmatrix} 0 & a_6 & 0 \end{bmatrix} x(k) \end{aligned}$$

Assuming no calves are added into the population, and using (1.2), with an initial value $x(0) = x_0$, the state solution is given by:

$$x(k) = \begin{bmatrix} 0 & a_1 & a_2 \\ a_3 & -a_6 & 0 \\ 0 & a_4 & a_5 \end{bmatrix}^k x_0$$

And in turn the output of the ranch is given by:

$$y(k) = \begin{bmatrix} 0 & a_6 & 0 \end{bmatrix} \begin{bmatrix} 0 & a_1 & a_2 \\ a_3 & -a_6 & 0 \\ 0 & a_4 & a_5 \end{bmatrix}^k x_0$$

Chapter 2

Observability and Reachability

Two important properties of linear control systems are reachability and observability.

2.1 Reachability

Reachability is the idea that a system, with a zero initial state, can be manipulated to yield any desired state with a suitable input. For example, an economist might want to be able to control the rate of inflation by increasing or decreasing taxes.

Definition: Consider the system (1.1). It is called *completely reachable* if, for the initial state $x(0) = 0$ and any desired state x_f, there exists a finite positive integer k_1 and a discrete scalar input u, such that $x(k_1) = x_f$ [9, p. 46].

For a linear time invariant system there is a simple condition that can be used to determine whether the system is completely reachable or not.

Theorem: Let the $n \times n$ matrix $\mathcal{R} = [b|Ab|A^2b|...|A^{n-1}b]$ be the *reachability matrix* of the system (1.1). Then, the system is completely reachable if and only if Rank($\mathcal{R}$) $= n$ [9, p. 46].

Proof: "$\Leftarrow$". Assume rank($\mathcal{R}$) $= n$. The goal is to find scalar inputs $u(0)$, $u(1)$, ..., $u(n-1)$, so that any x_f can be attained. Let u be the vector of the scalar inputs in reverse order, $u = \begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix}$. By eq. (1.2) the state vector $x(n)$ is given by:

$$\begin{aligned} x(n) &= \underbrace{A^n x_0}_{=0} + \sum_{j=0}^{n-1} A^{n-j-1} b u(j) \\ &= A^{n-1} b u(0) + A^{n-2} b u(1) + ... + Abu(n-2) + bu(n-1) \\ &= [b|Ab|A^2 b|...|A^{n-1} b] \begin{bmatrix} u(n-1) \\ u(n-2) \\ \vdots \\ u(0) \end{bmatrix} = \mathcal{R} u \\ \iff x(n) &= \mathcal{R} u \end{aligned}$$

Since $\mathcal{R}$ is of full rank n, it is invertible and the input vector u can be uniquely determined with matrix inversion. By picking $k_1 = n$ in the definition above any desired state x_f can be attained, and thus the system is completely reachable.

"⇒". Now, consider the case where the system is completely reachable but assume rank($\mathcal{R}$) $< n$. Then the columns of $\mathcal{R}$ are not linearly independent, and there exists a state x_f that is not in the column space of $\mathcal{R}$, which is a contradiction to the assumption that the system is completely reachable. Thus the rank of $\mathcal{R}$ must equal n. □

Note: The desired state might be reached before the time k_1, but it is guaranteed to be reached when $k_1 = n$.

Example:

Consider the cattle ranch problem from chapter 1. The reachability matrix is given by:

$$\mathcal{R} = [b|Ab|A^2 b] = \begin{bmatrix} 1 & 0 & a_1 a_3 \\ 0 & a_3 & -a_3 a_6 \\ 0 & 0 & a_3 a_4 \end{bmatrix}$$

If the survivability of the calves drops to zero, then rank($\mathcal{R}$) = 1 and if the survivability of the mature cows drops to zero, then rank($\mathcal{R}$) = 2 and in either case the system will not be reachable.

On a final note, reachability must not be confused with the concept of *controllability*, which is the idea of a system being able to reach the zero state from any given initial state.

Definition:

The system (1.1) is called *completely controllable* if for any initial state $x(0) = x_0$, there exists a finite, positive integer k_1 and input $u(k), k = 0, 1, ..., k_1 - 1$ such that $x(k_1) = 0$ [9, p. 56].

In the continuous case reachability and controllability are equivalent, however in the discrete case there exist systems that can be controllable but not reachable. But as Rugh notes [10, p. 463], there exist systems that are controllable but highly trivial such as:

$$\begin{aligned} x(k+1) &= 0x(k) + 0u(k) \\ x(0) &= x_0 \end{aligned}$$

and the fact that a system might fail to be reachable because the interval for k might be simply too short or that the matrix A might not be invertible. Because of this reachability is usually considered rather than controllability when discrete systems are analysed.

2.2 Observability

Observability is the idea that one can observe any given state of a system only from its output.

Definition: Consider again the system (1.1), and let $u(k) = 0$ for all $k \geqslant 0$. The system is called *completely observable* if there exists a finite positive integer k_1 such that knowledge of $y(0)$, $y(1)$, ..., $y(k_1 - 1)$ is sufficient to uniquely determine the initial state x_0 [9, p. 48].

In the linear time invariant case, there is a simple condition that can be used to determine if a system is completely observable.

Theorem: Let the $n \times n$ matrix $\mathcal{O} = \left[\begin{array}{|c|} \hline c \\ \hline cA \\ \hline \vdots \\ \hline cA^{n-1} \\ \hline \end{array}\right]$ be the *observability matrix* of the system (1.1). Then, the system is completely observable if and only if $\text{rank}(\mathcal{O}) = n$ [9, p. 48].

Proof:

”$\Leftarrow$”. Assume that $\text{rank}(\mathcal{O}) = n$, let $u(k) = 0$, $k \geqslant 0$, and $y = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix}$ be a vector of the outputs. Using eq. (1.3), the output is given by $y(k) = cA^k x_0$ for $k = 0, 1, ..., n-1$. Then:

$$y = \begin{bmatrix} y(0) \\ y(1) \\ \vdots \\ y(n-1) \end{bmatrix} = \begin{bmatrix} cx_0 \\ cAx_0 \\ \vdots \\ cA^{n-1}x_0 \end{bmatrix} = \left[\begin{array}{c} c \\ \hline cA \\ \hline \vdots \\ \hline cA^{n-1} \end{array}\right] x_0 = \mathcal{O}x_0$$

Since $\mathcal{O}$ is of full rank n and thus invertible, and assuming all the outputs in y are known, any initial state x_0 can be uniquely determined by picking $k_1 = n$ in the definition and performing matrix inversion. Thus the system is completely observable.

”$\Rightarrow$”. Proof by contradiction. The goal here is to show that if $\mathcal{O}$ is not of full rank n, then two different initial states x_0 can be found from the outputs of the system. Assume the system is completely observable but the $\text{rank}(\mathcal{O}) < n$. Then, since $\mathcal{O}$ is not invertible, there exists a vector $x_0 \neq 0$ such that $\mathcal{O}x_0 = y = 0$.

However, the initial vector $x_0 = 0$ also gives the output vector $y = 0$, i.e. two different vectors x_0 exist that can be derived from $y = 0$. This is a contradiction to the assumption that the system is completely observable. Thus the rank of $\mathcal{O}$ has to equal n. □

Example:

Consider again the cattle ranch problem from chapter 1. Its observability matrix is given by:

$$\mathcal{O} = \begin{bmatrix} 0 & a_6 & 0 \\ a_3 a_6 & -a_6^2 & 0 \\ -a_3 a_6^2 & a_6(a_1 a_3 + a_6^2) & a_2 a_3 a_6 \end{bmatrix}$$

Thus, if a_2, a_3 or a_6 equal 0, i.e. if:

- No mature cows are taken from the population.
- The survival rates of the calves drops to zero.
- the reproduction rate of the old cows drops to zero

the system will not be observable.

While the definitions of reachability and observability, along with the conditions for a system to be completely reachable/observable, were only derived for discrete linear time invariant systems, the analogous conditions for a continuous linear time invariant system are the same [9, p. 40, p. 44].

Chapter 3

The Z-Transform

A fundamental part of discrete linear systems is the determination of a meaningful solution to a system. While (1.2) gives a general solution it is not necessarily meaningful. One way to solve such a problem is to use the Z-transform, which converts a system of functions such as (1.1) into an infinite power series.

Step by step process of the method is:

1. Taking the Z-transform of a system.
2. Finding a closed form expression of the transform, which yields a rational polynomial.
3. Simplify the closed form if needed.

The resulting function can then be compared to a table of inverse Z-transforms to determine its exact form, be it the state or the output. The Z-transform is the discrete analog of the Laplace transform for continuous functions, and is often called the *generating function.*

Definition: Let $f : \mathbb{Z} \to \mathbb{R}$, be a given discrete function. The *unilateral Z-transform*, $\mathscr{Z}[f](z)$ of f, is a function $F : \mathbb{C} \to \mathbb{C}$, defined as [2, p. 247]:

$$\mathscr{Z}[f](z) = F(z) = \sum_{k=0}^{\infty} f(k) z^{-k}$$

In a similar manner, the *bilateral* transform has the same definition, except the summation index starts at $-\infty$ in that case.

To determine the closed form expression of a Z-transform one can e.g. inspect the power series generated by the transform.

3.1 Power Series Inspection

This method is best described by an example. It revolves around manipulation of the sequence of $F(z)$ to yield a closed form expression.

Example:

Let $f(k) = (-1)^k, k \geqslant 0$. Then, its Z-transform is given by:

$$\mathcal{Z}[f](z) = F(z) = \sum_{k=0}^{\infty} \frac{(-1)^k}{z^k} = 1 - \frac{1}{z} + \frac{1}{z^2} - \frac{1}{z^3} + \cdots$$

Now consider the sequence:

$$\begin{aligned}
&F(z) + \frac{1}{z}F(z) = F(z)\left(\frac{z+1}{z}\right) \\
&= 1 - \frac{1}{z} + \frac{1}{z^2} - \frac{1}{z^3} + \cdots + \left(\frac{1}{z} - \frac{1}{z^2} + \frac{1}{z^3} - \frac{1}{z^4} + \cdots\right) \\
&= 1 + \frac{(-1)^k}{z^{k+1}}
\end{aligned}$$

The last term in that sequence approaches 0 when $|z| > 1$. Thus the sequence becomes:

$$\begin{aligned}
F(z)\left(\frac{z+1}{z}\right) &= 1 \\
\Longleftrightarrow F(z) &= \frac{z}{z+1}
\end{aligned}$$

Which is the generating function for f with region of convergence $|z| > 1$.

3.2 Region of Convergence

The importance of the region of convergence of Z-transforms is that without it the transform can not be uniquely determined. Consider the two functions:

$$u(k) = \begin{cases} 1, & \text{if } k \geqslant 0 \\ 0, & \text{otherwise} \end{cases}$$

and

$$v(k) = -u(-k-1) = \begin{cases} -1 & \text{, if } k \leqslant -1 \\ 0 & \text{, otherwise} \end{cases}$$

The Z-transform of both functions is given by $U(z) = V(z) = \frac{z}{z-1}$ but the regions of convergence is $|z| > 1$ for $U(z)$ while the region for $V(z)$ is given by $|z| < 1$. This demonstrates the importance of regions of convergence. It's however worth noting that e.g. [9, p. 61] denotes the Z-transform as the *formal power series* of the indeterminate z and does not question its region of convergence or if it converges at all.

3.3 Properties of the Z-transform

The Z-transform has a number of properties that will be listed below without proof [13].

- **Linearity:** $\mathcal{Z}[\alpha f + \beta g](z) = \alpha F(z) + \beta G(z)$, $\alpha, \beta \in \mathbb{R}$
- **Translation:**
  - $\mathcal{Z}[f(\cdot - 1)](z) = z^{-1}F(z)$
  - $\mathcal{Z}[f(\cdot - n)](z) = z^{-n}F(z)$
  - $\mathcal{Z}[f(\cdot + 1)](z) = zF(z) - zf(0)$
  - $\mathcal{Z}[f(\cdot + 2)](z) = z^2F(z) - z^2f(0) - zf(1)$
  - $\mathcal{Z}[f(\cdot + n)](z) = z^nF(z) - \sum_{p=0}^{n-1} z^{n-p}f(p)$
- **Scaling:** $\mathcal{Z}[\alpha^n f(\cdot)](z) = F(z/\alpha)$, $\alpha \in \mathbb{R}\backslash 0$
- **powers multiplication:** $\mathcal{Z}[k^n f(\cdot)](z) = (-1)^k \left(z\frac{d}{dz}\right)^k F(z)$
- **Convolution:** Let $h(k) = \sum_{j=0}^{\infty} f(k-j)g(j)$, i.e. the convolution of functions f and g. Then, $\mathcal{Z}[h](z) = F(z)G(z)$

3.4 Z-transform Pairs

Below is a table of well known Z-transform pairs [2, p. 252]. The functions $\delta(k)$ and $u(k)$ denote the unit impulse and unit step function respectively, and α and γ are scalars in $\mathbb{R}$.

Signal	Z-transform	ROC
$\delta(k-n)$	z^{-n}	$\mathbb{C}$
$u(k)$	$\frac{z}{z-1}$	$\|z\| > 1$
$-u(-k-1)$	$\frac{z}{z-1}$	$\|z\| < 1$
$ku(k)$	$\frac{z}{(z-1)^2}$	$\|z\| > 1$
$k^2u(k)$	$\frac{z(z+1)}{(z-1)^3}$	$\|z\| > 1$
$(-(\alpha)^k)u(k-1)$	$\frac{z}{z-\alpha}$	$\|z\| < \|\alpha\|$
$\alpha^k u(k)$	$\frac{z}{z-\alpha}$	$\|z\| > \|\alpha\|$
$\gamma^k \cos(\alpha k)u(k)$	$\frac{z(z-\gamma cos(\alpha))}{z^2-2\gamma\cos(\alpha k)z+\gamma^2}$	$\|z\| > \|\alpha\|$
$\gamma^k \sin(\alpha k)u(k)$	$\frac{z\gamma\sin(\alpha)}{z^2-2\cos(\alpha)z+\gamma^2}$	$\|z\| > \|\alpha\|$

3.5 The inverse Z-Transform

As mentioned in the beginning of this chapter, for the application of a Z-transform of a system to be of any use, a process of inverting the transform has to be available. There are three methods to do so that will be mentioned here [2, p. 260]:

- Inspection of the power series expansion of the generating function.
- Partial fraction decomposition of the generating function (PFD).
- Contour integration.

The inverse Z-transform is denoted by $f(k) = \mathcal{Z}^{-1}[F](k)$.

3.5.1 Contour Integral

The contour integration method defines the inverse Z-transform of function F as [7, p. 758]:

$$\mathcal{Z}^{-1}[F](k) = \frac{1}{2\pi i}\oint_{r} F(z)z^{k-1}dz$$

where i is the imaginary unit and r is a counter clockwise contour around the origin of the complex plane. This method requires advanced knowledge in complex analysis and will not be further discussed here.

3.5.2 Power Series Expansion

This method revolves around analyzing the power series expansion of the function $F(z)$, e.g. if $F(z)$ is a proper, reduced rational function, then by performing the long division of the fraction one can acquire the infinite power series which can be used along with the Z-transform pair table to determine the inverse function of F.

Example:

Consider the function $F(z) = \frac{z}{z-\alpha}$, $\alpha \in \mathbb{R}$ with $|z| > |\alpha|$. Carrying out the long division of the fraction gives:

$$\frac{z}{z-a} = 1 + \frac{a}{z} + \frac{a^2}{z^2} + ... = \sum_{k=0}^{\infty} a^k z^{-k}$$

By comparing this sequence with the definition of a Z-transform it is readily seen that the inverse Z-transform is $f(k) = a^k,\ k = 1, 2, 3, ...$

3.5.3 Partial Fraction Decomposition

Consider the rational function $F(z) = \frac{P(z)}{Q(z)}$, where P and Q are polynomials in z. The method revolves around decomposing the function into a sum of terms where the denominator in each term is a factor of $Q(z)$, then the terms are inspected in the same manner as above, and their inverse transform determined. By the linearity of Z-transforms, the inverse transform is thus the sum of those terms, and is found by comparing them to a list of Z-transform pairs.

Example:

Consider the function $F(z) = \frac{z^2+3z}{z^2-8z+15}$ with $|z| > 5$. To simplify the calculation, the partial fraction decomposition of $z^{-1}F(z)$ will be derived, and then $F(z)$ will be determined by the proper multiplication of the variable z. The goal is to find coefficients A and B such that:

$$z^{-1}F(z) = \frac{z+3}{z^2-8z+15} = \frac{z+3}{(z-5)(z-3)} = \frac{A}{z-5} + \frac{B}{z-3}$$

Taking the last two equalities and multiplying the denominator over to the right hand side gives:

$$\begin{aligned}
&\frac{z+3}{(z-5)(z-3)} = \frac{A}{z-5} + \frac{B}{z-3} \\
\iff & A(z-3) + B(z-5) = z+3 \\
\iff & z(A+B) - (3A+5B) = z+3
\end{aligned}$$

Comparison of the coefficients on both sides of the last equation gives the linear system:

$$\begin{cases} A+B=1 \\ 3A+5B=-3 \end{cases}$$

which yields $A = 4, B = -3$. Thus the partial fraction decomposition of $F(z)$ is given by $F(z) = \frac{4z}{z-5} - \frac{3z}{z-3}$. Now by comparing this generating function to the table above thus gives an inverse transform $f(k) = 4 \cdot 5^k - 3^{k+1}$.

3.6 Solutions to the State Vector Equation

These tools can now be used to represent the equations in (1.1) as Z-transforms in the following way, by taking the Z-transform of (1.1):

$$\begin{aligned}
&\mathscr{Z}[x(k+1)](z) = A\mathscr{Z}[x(k)](z) + b\mathscr{Z}[u(k)](z) \\
\iff & zX(z) - zx_0 = AX(z) + bU(z) \\
\iff & (zI-A)X(z) = zx_0 + bU(z) \\
\iff & X(z) = (zI-A)^{-1}[zx_0 + bU(z)]
\end{aligned}$$

Using this expression, the Z-transform of the output is given by:

$$\begin{aligned}
Y(z) &= cX(z) + dU(z) \\
&= c\left[(zI-A)^{-1}zx_0 + (zI-A)^{-1}bU(z)\right] + dU(z) \qquad (3.1) \\
&= cz(zI-A)^{-1}x_0 + \left[c(zI-A)^{-1}b + d\right]U(z)
\end{aligned}$$

and finally, now by performing partial fraction decomposition as needed, the inverse transform of the state or the output can be found.

Example:

Consider the system:

$$\begin{aligned} x(k+1) &= \begin{bmatrix} 3 & 1 \\ 0 & 5 \end{bmatrix} x(k) + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u(k) \\ y(k) &= \begin{bmatrix} 1 & 0 \end{bmatrix} x(k) \\ x(0) &= 0 \end{aligned}$$

with $u(k) = 1, k \geqslant 0$, and Z-transform $U(z) = \frac{z}{z-1}$. The matrix $(zI - A)^{-1}$ is given by:

$$\frac{1}{(z-3)(z-5)} \begin{bmatrix} z-5 & 1 \\ 0 & z-3 \end{bmatrix}$$

and thus the expression for the Z-transform of the state is:

$$\begin{aligned} X(z) &= \frac{1}{(z-3)(z-5)} \begin{bmatrix} z-5 & 1 \\ 0 & z-3 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \frac{z}{z-1} \\ &= \frac{z}{(z-1)(z-3)(z-5)} \begin{bmatrix} z-5 \\ 0 \end{bmatrix} \\ &= \frac{z}{(z-1)(z-3)} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} -\frac{1}{2}\frac{z}{z-1} + \frac{1}{2}\frac{z}{z-3} \\ 0 \end{bmatrix} \end{aligned}$$

where the last equality is derived with partial fraction decomposition. Using this state, and the equation derived in (3.1), yields the Z-transformed output:

$$Y(z) = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} -\frac{1}{2}\frac{z}{z-1} + \frac{3}{2}\frac{z}{z-3} \\ 0 \end{bmatrix} = -\frac{1}{2}\frac{z}{z-1} + \frac{3}{2}\frac{z}{z-3}$$

and finally, using the table of transforms, the output of the system is given by:

$$y(k) = -\frac{1}{2} + \frac{1}{2}3^k = \frac{1}{2}(3^k - 1)$$

Chapter 4

Stability

Two different types of stability will be discussed here, *asymptotic stability* and *bounded-input-bounded-output stability* (BIBO). Results will also be provided to show how they relate to each other. Stability is usually the first property that is considered when the behaviour of a system is analyzed [9, p. 113]. Other types of stability, such as exponential stability, exist but will not be discussed here.

4.1 Asymptotic Stability

The first type of stability discussed is the asymptotic stability, which analyzes the stability of a system as time passes.

Definition: Consider System (1.1) and let $u(k) = 0, \forall k \geqslant 0$. The system is called *asymptotically stable* if for any inital state x_0 $\lim_{k\to\infty} x(k) = 0$ [9, p.121].

Informally, if it is assumed that the input to the system is always zero, the solution to the system will, as time passes, converge to zero.

The following theorem gives a method to determine the asymptotic stability of a linear time-invariant system.

Theorem: The system (1.1) is asymptotically stable if and only if the eigenvalues of A all have magnitude less than unity [9, p. 122].

Proof: Using (1.2) and the definition of stability, (1.1) is asymptotically stable if:

$$\lim_{k\to\infty} x(k) = \lim_{k\to\infty} A^k x_0 = 0 \tag{4.1}$$

By the *Jordan normal form* [4, p. 354] there exists a nonsingular matrix T such that A can be written as $A = TJT^{-1}$, where J is a block diagonal matrix:

$$J = \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix}$$

and each block J_i is of the form:

$$J_i = \begin{bmatrix} \lambda_i & 1 & 0 & & \\ 0 & \lambda_i & 1 & \ddots & \\ & \ddots & \ddots & \ddots & 0 \\ & & \ddots & \ddots & 1 \\ & & & 0 & \lambda_i \end{bmatrix}$$

with λ_i being an eigenvalue of A. Furthermore, each block J_i can be written as a sum of a scaled identity matrix $D_i = \lambda_i I$ and a nilpotent matrix N_i, $J_i = D_i + N_i$, where:

$$N_i = \begin{bmatrix} 0 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 0 \\ & & & \ddots & 1 \\ & & & & 0 \end{bmatrix}$$

Let j denote the dimension of J_i. Then, $N_i^j = 0$, and furthermore, each successive power of N_i "shifts" the superdiagonal to the right, i.e.:

$$N_i^2 = \begin{bmatrix} 0 & 0 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & \ddots & 0 \\ & & & & 0 \end{bmatrix}$$

and so on. Thus (4.1) can be written as:

$$\lim_{k\to\infty} x(k) = \lim_{k\to\infty} (TJT^{-1})^k x_0 = 0 \tag{4.2}$$

By expanding (4.2), it becomes:

$$\begin{aligned} &\lim_{k\to\infty} (TJT^{-1})(TJT^{-1})...(TJT^{-1})x_0 = 0 \\ \iff &\lim_{k\to\infty} TJ^kT^{-1}x_0 = 0 \end{aligned}$$

Now, inspecting the matrix J^k:

$$\begin{aligned} J^k &= \underbrace{\begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix} \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix} \cdots \begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix}}_{k \text{ times}} \\ &= \begin{bmatrix} J_1^k & & & \\ & J_2^k & & \\ & & \ddots & \\ & & & J_m^k \end{bmatrix} \\ &= \begin{bmatrix} (D_1+N_1)^k & & & \\ & (D_2+N_2)^k & & \\ & & \ddots & \\ & & & (D_m+N_m)^k \end{bmatrix} \end{aligned}$$

By the *binomial theorem*, each block equals:

$$\begin{aligned} (D_i+N_i)^k = &\binom{k}{0}D_i^k + \binom{k}{1}D_i^{k-1}N_i + ... + \binom{k}{j-1}D_i^{k-j+1}N_i^{j-1} \\ &+ \underbrace{\binom{k}{j}D_i^{k-j}N_i^j + ... + \binom{k}{k}N_i^k}_{=0} \end{aligned}$$

$$= \binom{k}{0}\begin{bmatrix} \lambda_i^k & & \\ & \ddots & \\ & & \lambda_i^k \end{bmatrix} + \binom{k}{1}\begin{bmatrix} 0 & \lambda_i^{k-1} & & \\ & \ddots & \ddots & \\ & & \ddots & \lambda_i^{k-1} \\ & & & 0 \end{bmatrix} + ...$$

$$+ \binom{k}{j-1}\begin{bmatrix} 0 & & \lambda_i^{k-j+1} \\ & \ddots & \\ & & 0 \end{bmatrix}$$

which is an upper triangular matrix with each element being the eigenvalue to some power $k - l$, times some constant $c_l = \binom{k}{l}$, for $l = 0, 1, ..., j - 1$. Thus, as k goes to infinity, each element of each block of J will converge to zero if and only if $|\lambda_i| < 1$, which in turn means that (4.2) will hold. This concludes the proof. □

A similar argument exists for continuous systems, where they are asymptotically stable if and only if the real parts of the eigenvalues are negative.

Although the Jordan normal form is used in this proof, it is often avoided to use it in numerical algorithms because of the roundoff difficulties that result from the lack of orthogonality in T [1, p. 503].

4.2 Bounded-Input-Bounded-Output Stability

Bounded-input-bounded-output stability relates the boundedness of an input to a system to the boundedness of the resulting output.

Definition: Consider System (1.1) and let $x_0 = 0$. The system is called *Bounded-input-bounded-output stable* if the following holds. For any input function u having bound v, that is $|u(k)| \leqslant v$ for all $k \geqslant 0$, there exists a q such that $|y(k)| \leqslant q$ for all $k \geqslant 0$ [9, p. 122].

As with asymptotic stability, a theorem is provided that determines the bounded-input-bounded-output stability of a system.

Theorem: Consider System (1.1) and let $h(k) = cA^{k-1}b, k \geqslant 1$. then the system is bounded-input-bounded-output stable if and only if [9, p. 122]:

$$\sum_{k=1}^{\infty} |h(k)| < \infty$$

Proof:

"$\Leftarrow$": Let $x_0 = 0$, $|u(k)| \leqslant v, \forall k$, $p = \sum_{k=1}^{\infty} |h(k)|$, and consider the output given by (1.3). Taking the absolute value of both sides of it gives:

$$\begin{aligned}
|y(k)| &= \left| \sum_{j=0}^{k-1} cA^{k-j-1}bu(j) + du(k) \right| \\
&\leqslant \sum_{j=0}^{k-1} |cA^{k-j-1}bu(j)| + |du(k)| && \text{Triangle inequality} \\
&\leqslant \sum_{j=0}^{k-1} |cA^{k-j-1}b|v + |d|v && |u(k)| \leqslant v \\
&= (|cA^{k-1}b| + |cA^{k-2}b| + ... + |cAb| + |cA^0b|)v + |d|v \\
&= (|h(k)| + |h(k-1)| + ... + |h(2)| + |h(1)|)v + |d|v \\
&= \sum_{l=1}^{k} |h(l)|v + |d|v \\
&\leqslant pv + |d|v
\end{aligned}$$

where the last inequality follows from the assumption of the theorem. By picking $q = pv + |d|v$ in the definition above the output will be bounded and the system is bounded-input bounded-output stable.

"$\Rightarrow$": Proof by contradiction. The goal here is to show that if $\sum_{k=1}^{\infty} |h(k)| = \infty$ then there exists a bounded input that results in an unbounded output. Thus assume $x_0 = 0$, $\sum_{k=1}^{\infty} |h(k)| = \infty$, the system is bounded-input-bounded-output stable, i.e. for each input $u(k)$ and output $y(k)$ there exist constants v and q respectively such that $|u(k)| \leqslant v, \forall k$ and $|y(k)| \leqslant q, \forall k$ and let n be an integer such that $\sum_{k=1}^{n-1} |h(k)| > q + |d|$ which is always possible since the sum diverges.

Now, let the signal $u(j)$ be such that:

$$u(j) = \begin{cases} 1, & \text{if } cA^{k-j-1}b > 0 \\ 0, & \text{if } cA^{k-j-1}b = 0 \\ -1, & \text{if } cA^{k-j-1}b < 0 \end{cases}$$

where k is fixed but arbitrary and $0 \leqslant j \leqslant k-1$. This signal has magnitude 1 or less and is thus bounded. It also has the property that each term in the output given by (1.3) is non-negative, i.e.:

$$\begin{aligned}
y(k) &= \sum_{j=0}^{k-1} cA^{k-j-1}bu(j) + du(k) \\
&= \sum_{j=0}^{k-1} |cA^{k-j-1}b| + du(k) \\
&= |cA^{k-1}b| + |cA^{k-2}b| + \cdots + |cA^{0}b| + du(k) \\
&= \sum_{j=1}^{k-1} |h(j)| + du(k)
\end{aligned}$$

But now consider $y(n)$. This gives:

$$\begin{aligned}
y(n) &= \sum_{j=1}^{n-1} |h(j)| + du(n) \\
&> q + |d| + du(n)
\end{aligned}$$

which contradicts the assumption that $|y(k)| \leqslant q$, and thus that the system is bounded-input-bounded-output stable. Thus it is necessary for the sum to be finite if the system is to be bounded-input-bounded-output stable. □

4.3 Relation of Asymptotic stability and Bounded-Input Bounded-Ouput stability

To conclude this chapter, a relation between asymptotic stability and bounded-input bounded-output stability is provided.

Theorem: If (1.1) is asymptotically stable, then it is bounded-input bounded-output stable [9, p. 122].

Proof:

Since (1.1) is asymptotically stable we know that all the eigenvalues of A have magnitude less than unity. It follows from the *spectral radius theorem* [6, p. 617] and the continuity of norms that:

$$\lim_{k\to\infty} A^k = 0 \Rightarrow \lim_{k\to\infty} ||A^k|| = 0$$

Then there exists a number d such that $0 < d < 1$ and $||A^k|| < d^k$. Now consider:

$$\sum_{k=1}^{\infty} |h(k)| = \sum_{k=1}^{\infty} |cA^{k-1}b|$$

By the *Cauchy-Schwarz inequality* [12, p. 21] it follows that:

$$\sum_{k=1}^{\infty} |cA^{k-1}b| \leqslant ||c|| \cdot ||b|| \sum_{k=1}^{\infty} ||A^{k-1}|| < ||c|| \cdot ||b|| \cdot \sum_{k=1}^{\infty} d^{k-1}$$

and since $d < 1$, the right hand side of this equation converges to some constant p. Thus we have:

$$\sum_{k=1}^{\infty} |h(k)| < p$$

and thus it follows that (1.1) is bounded-input-bounded-output stable. □

Theorem: If (1.1) is completely observable, completely reachable and bounded-input bounded-output stable, then it is also asymptotically stable [9, p. 122].

Proof:

Since (1.1) is bounded-input bounded-output stable we know that $\sum_{k=1}^{\infty} |cA^{k-1}b|$ is finite, which implies that $\lim_{k\to\infty} cA^{k-1}b = 0$. Now since k approaches infinity we can pluck out an A on the left side and on the right side to get:

$$\begin{aligned}
&\lim_{k\to\infty} (c)(A^{k-1})(b) = 0\\
&\lim_{k\to\infty} (cA)(A^{k-3})(Ab) = 0\\
&\lim_{k\to\infty} (cA^2)(A^{k-5})(A^2b) = 0\\
&\vdots\\
&\lim_{k\to\infty} (cA^i)(A^{k-i-j})(A^jb) = 0, \text{ for } i, j = 0, 1, ..., n-1
\end{aligned}$$

The relation above can be rewritten as:

$$\begin{aligned}
&\lim_{k\to\infty} \left[\begin{array}{|c|} \hline c \\ \hline cA \\ \hline \vdots \\ \hline cA^{n-1} \\ \hline \end{array}\right] A^{k-i-j}[b|Ab|A^2b|...|A^{n-1}b] = 0\\
\Longleftrightarrow &\lim_{k\to\infty} \mathcal{O}A^{k-i-j}\mathcal{R} = 0
\end{aligned}$$

where $\mathcal{O}$ and $\mathcal{R}$ are the observability and reachability matrices respectively. Now since both of them are of full rank n by assumption it follows that:

$$\lim_{k\to\infty} A^{k-i-j} = 0$$

which in turn implies that (1.1) is asymptotically stable. This concludes the proof. □

Example:

Consider the system [9, p. 125]:

$$\begin{aligned} x(k+1) &= \begin{bmatrix} \frac{1}{2} & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix} x(k) + \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} u(k) \\ y(k) &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} x(k) \end{aligned}$$

with A, b, c taking on the usual notation. Since A is triangular its eigenvalues are on the diagonal and the largest one in absolute value is 1, thus the system is not asymptotically stable. By using eigenvalue decomposition of $A = P\Lambda P^{-1}$, the sum $\sum_{k=1}^{\infty} |h(k)|$ can be determined and is given by:

$$\begin{aligned} \sum_{k=1}^{\infty} |h(k)| &= \sum_{k=1}^{\infty} |c\left(P\Lambda P^{-1}\right)^k b| \\ &= \sum_{k=1}^{\infty} |cP\Lambda^k P^{-1} b| \\ &= \sum_{k=1}^{\infty} \left| \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & -\frac{2}{\sqrt{5}} & 0 \\ 0 & \frac{1}{\sqrt{5}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \left(\frac{1}{2}\right)^k & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & (-1)^k \end{bmatrix} \begin{bmatrix} 1 & 2 & 0 \\ 0 & \sqrt{5} & 0 \\ 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \right| \\ &= \sum_{k=1}^{\infty} \left(\frac{1}{2}\right)^{k-1} \end{aligned}$$

which is a convergent series and thus the system is bounded-input bounded-output stable.

Furthermore, the ranks for the observability and reachability matrices is 2 which is not surprising given that the system is not asymptotically stable, meaning that either its bounded-input bounded-output stability, or rank conditions for the observability and reachability matrices would have to fail.

Chapter 5

Numerical Methods to Determine Rank

How do we determine the rank of large matrices? In cases that appear in applications in engineering and physics the matrices in question can have very large dimensions, say 1000×1000. For this reason the use of numerical methods is paramount, but care must be taken because of the finite arithmetic of computers. Because of the finite arithmetic, we need to be sure that the algorithm used to determine rank is *stable*, i.e. that small perturbations in input do not give large perturbations in the output. For more on stability see [12, Ch.14, 15].

5.1 The singular Value Decomposition

The best method known is to use the *singular value decomposition* [5, p. 167]. It is given by:

$$A = U\Sigma V^T$$

The matrix Σ is diagonal with the singular values of A on its diagonal, arranged from largest to the smallest, and U and V are unitary. It not only finds the rank of A by looking at the singular values, but it also tells us how close in the 2-norm we are from a matrix of a lower rank, by looking at the smallest singular value. It is also *backwards stable*, a particularly nice form of stability, i.e. the singular values attained by the factorization are exact for a matrix close by the original matrix in the 2-norm [8, p. 131].

The singular value decomposition is however computationally expensive. For example the commonly used Golub-Kahan method has two phases:

1. Decompose A into $Q^T BP$ where B is bidiagonal and Q, P are orthogonal. If only the singular values are required then Q and P can be discarded.

2. Compute the singular value decomposition of B, e.g. with a variant of the QR algorithm or by the divide and conquer method [12, p. 239].

The cost of the algorithm for a $m \times n$ matrix is dominated by the first phase which has a cost of $\sim 4mn^2 - \frac{4}{3}n^3$ flops, compared to $\sim 2mn^2 - \frac{2}{3}n^3$ flops for a Householder QR decomposition and $\sim \frac{2}{3}m^3$ flops for Gaussian Elimination. [12, p. 237. p. 75. p. 160]. So why not use QR decomposition or Gaussian elimination to determine the rank instead? It turns out these two methods are not stable methods of rank determination.

5.2 The QR Decomposition

Consider the $n \times n$ matrix $A = QR$. Since Q is of full rank we know that $\text{rank}(A) = \text{rank}(R)$. Thus if A is rank deficient there should be an element on the diagonal of R that is small when compared to the machine epsilon. However consider the matrix:

$$\begin{aligned}\text{Kah}_n(c) &= \begin{bmatrix} 1 & -c & -c & \dots & -c \\ 0 & s & -cs & \dots & -cs \\ \vdots & \ddots & s^2 & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -cs^{n-2} \\ 0 & \dots & \dots & 0 & s^{n-1} \end{bmatrix} \\ &= \begin{bmatrix} 1 & & & & \\ & s & & & \\ & & s^2 & & \\ & & & \ddots & \\ & & & & s^{n-1} \end{bmatrix} \begin{bmatrix} 1 & -c & -c & \dots & -c \\ 0 & 1 & -c & \dots & -c \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -c \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix}\end{aligned}$$

where $c > 0, s > 0$ and $c^2 + s^2 = 1$ [4, p. 279]. This is a famous example discovered by William Kahan. Performing a QR decomposition on $\text{Kah}_{100}(0.2)$ reveals the smallest diagonal element of R to be $r_{100,100} = 0.13256413$, a number far away from zero when compared to the standard 64 bit machine epsilon $\epsilon = 2.220446 \cdot 10^{-16}$, but at the same time the smallest singular value of $\text{Kah}_{100}(0.2)$ is $\sigma_n = 3.678056 \cdot 10^{-09}$. Thus the Kahan matrix gets closer and closer to a singular matrix as its dimensions grow.

n	R_{nn}	σ_n	σ_1/σ_n	σ_{n-1}/σ_n
5	$9.215999 \cdot 10^{-1}$	$6.126013 \cdot 10^{-1}$	1.838307	1.681975
10	$8.321862 \cdot 10^{-1}$	$2.788470 \cdot 10^{-1}$	4.622766	3.336641
50	$3.678284 \cdot 10^{-1}$	$9.287521 \cdot 10^{-5}$	$4.990962 \cdot 10^{5}$	$4.427926 \cdot 10^{3}$
100	$1.325641 \cdot 10^{-1}$	$3.678056 \cdot 10^{-9}$	$2.177658 \cdot 10^{9}$	$4.029607 \cdot 10^{7}$
150	$4.777569 \cdot 10^{-2}$	$1.456591 \cdot 10^{-13}$	$7.257142 \cdot 10^{13}$	$3.667112 \cdot 10^{11}$
200	$1.721820 \cdot 10^{-2}$	$5.786992 \cdot 10^{-18}$	$2.190496 \cdot 10^{18}$	$3.326517 \cdot 10^{15}$
250	$6.205382 \cdot 10^{-3}$	$1.131002 \cdot 10^{-20}$	$1.281683 \cdot 10^{21}$	$6.134230 \cdot 10^{17}$
300	$2.236399 \cdot 10^{-3}$	$9.008557 \cdot 10^{-21}$	$1.789173 \cdot 10^{21}$	$2.775550 \cdot 10^{17}$

Table 5.1: A comparison table of the trailing element of R and some of the singular values of $\text{Kah}_n(0.2)$ as its dimension grows. The number σ_1/σ_n is known as the *condition number* of the matrix, and σ_{n-1}/σ_n denotes the ratio of the second smallest and smallest singular value.

5.3 Gaussian Elimination

In the case of Gaussian Elimination consider the $n \times n$ matrix [5, p.168]:

$$A_n = \begin{bmatrix} 1 & -1 & \dots & -1 \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -1 \\ 0 & \dots & 0 & 1 \end{bmatrix}$$

This matrix is in reduced row echelon form and thus is an example of a matrix that could result from Gaussian elimination. It has a nice, upper triangular structure with all the diagonal elements equal to 1, and hence $\det(A_n) = 1$ for all n. So it clearly has full rank. However, for A_{100} its smallest singular value is $\sigma_n = 2.637323 \cdot 10^{-18}$, so in the same manner as the Kahan matrix it gets arbitrarily close to a singular matrix as its dimension grows.

n	σ_n	σ_1/σ_n	σ_{n-1}/σ_n
5	$9.298533 \cdot 10^{-2}$	$2.942754 \cdot 10^{1}$	$1.623324 \cdot 10^{1}$
10	$2.929643 \cdot 10^{-3}$	$1.918487 \cdot 10^{3}$	$5.127458 \cdot 10^{2}$
50	$2.659590 \cdot 10^{-15}$	$1.162226 \cdot 10^{16}$	$5.640279 \cdot 10^{14}$
100	$4.772302 \cdot 10^{-18}$	$1.314330 \cdot 10^{19}$	$3.143180 \cdot 10^{17}$
150	$1.162393 \cdot 10^{-17}$	$8.134010 \cdot 10^{18}$	$1.290450 \cdot 10^{17}$
200	$3.369086 \cdot 10^{-18}$	$3.751085 \cdot 10^{19}$	$4.452261 \cdot 10^{17}$
250	$1.572475 \cdot 10^{-17}$	$1.006099 \cdot 10^{19}$	$9.539123 \cdot 10^{16}$
300	$7.556553 \cdot 10^{-18}$	$2.514857 \cdot 10^{19}$	$1.985035 \cdot 10^{17}$

Table 5.2: A comparison of the smallest singular values of A_n as n grows. Even though it has a structure that shows full rank, the singular values show that as its dimension grows it becomes closer and closer to a singular matrix.

By performing singular value decomposition of $A_{100} = U\Sigma V^T$, we can define $\tilde{\Sigma} = \Sigma + \delta\Sigma$ such that $\tilde{\sigma}_{100,100} = 0$. Then $\tilde{A} = U\tilde{\Sigma}V^T$ will be the singular matrix that is closest to A in the 2-norm.

These cases shown here are rare to happen in mathematical modelling but are nonetheless one of the reason why the singular value decomposition is considered the most reliable method for rank determination even though computationally QR decomposition and Gaussian elimination are cheaper to perform.

Chapter 6

Conclusions

The aim of this thesis was the study of linear time-invariant systems, and the concepts of reachability, observability and stability of those systems and how these concepts can be used to get desirable results when using them to model systems in real life.

While the conditions for these concepts are easily derived, a difficult problem that we face when applying these methods is the finite arithmetic of the computers performing the tasks at hand. This shows the importance of the study of perturbation and how it affects numerical computations.

Historically finite arithmetic has been a large obstacle to overcome for numerical analysts. Early hardware came with its own standards for floating point precision and operations that worked on one system could lead to different results when done on a different system [11]. This led to the standardization of floating point precision with the IEEE-754 standard, which most common computers use today.

It is interesting to note that the singular value calculations performed in this thesis were tested on two different computers with two different methods in the Python programming language, one which computes matrices U, Σ and V and another which discards U and V at each step in the calculation. On each occasion the results given had small differences when compared to each other.

Bibliography

[1] H. Anton and R.C. Busby. *Contemporary Linear Algebra.* Wiley, 2002.

[2] Richard Baraniuk. *Signals and Systems.* Rice University, Houston, 2003.

[3] S. Barnett and R.G. Cameron. *Introduction to Mathematical Control Theory, 2nd. edition.* Oxford University Press, Houston, 1985.

[4] G.H. Golub and C.H. Van Loan. *Matrix Computations, 4th. edition.* John Hopkins University Press, Maryland, 2013.

[5] V. Klema and A. Laub. The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176, April 1980.

[6] C.D. Meyer. *Matrix Analysis and Applied Linear Algebra.* Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2000.

[7] Allan V. Oppenheim. *Signals and Systems, 2nd. edition.* Prentice Hall, New Jersey, 1997.

[8] C. Paige. Properties of numerical algorithms related to computing controllability. *IEEE Transactions on Automatic Control*, 26(1):130–138, February 1981.

[9] Wilson J. Rugh. *Mathematical Description of Linear Systems.* Marcel Decker Inc., New York, 1975.

[10] Wilson J. Rugh. *Linear Systems Theory.* Prentice Hall, New Jersey, 1996.

[11] C. Severance. Ieee 754: An interview with william kahan. *Computer*, 31(3):114–115, March 1998.

[12] L. Trefethen and D. Bau. *Numerical Linear Algebra.* SIAM, Philadelphia, 1997.

[13] Eric W. Weisstein. Z-transform. `http://mathworld.wolfram.com/Z-Transform.html`. Accessed: 2018.10.01.

Bachelor’s Theses in Mathematical Sciences 2019:K26
ISSN 1654-6229

LUNFNA-4029-2019

Numerical Analysis
Centre for Mathematical Sciences
Lund University
Box 118, SE-221 00 Lund, Sweden

http://www.maths.lth.se/